SUOMEN TILASTOSEURAN VUOSIKIRJA 2013–2014

ÅRSBOK FÖR STATISTISKA SAMFUNDET I FINLAND 2013–2014

THE YEARBOOK OF THE FINNISH STATISTICAL SOCIETY 2013–2014

2014

SISÄLLYSLUETTELO

ESIMIEHEN PALSTA

Kimmo Vehkalahti
Suomen Tilastoseuran esimies 2012 ja 2013

Tilanne näyttää hyvältä! Huudahdus viittaa edellisessä vuosikirjassa aloittamaani, perinteisiä esipuheita täydentävään katsaukseen ”*Uuden hallituksen tervehdys*”, jota nyt käsillä olevassa kirjassa jatkaa seuran tämänhetkinen esimies **Jyrki Möttönen**. Kaksoiskatsaus tuli tarpeelliseksi, kun siirryimme noudattamaan biennaalimallia, siis toimimaan perinteisen vuosittaisen kalenterikierron sijaan kahden vuoden jaksoissa. Mallin implikoima oma esimieskauteni on jo takanapäin, joten tarkastelen tilannetta nyt ajallisesti etäämmältä.

Ajatus biennaalimallista oli noussut agendalle jo toimiessani seuran varaesimiehenä. Lopullinen sysäys tuli tammikuussa 2012, kun olin lupautunut siirtymään esimieheksi. Jyväskylän yliopisto oli luvannut järjestää Tilastopäivät, mutta pyydetyt luennoitsijat eivät suostuneetkaan, mikä johti vaikeuksiin. Olimme yhtäkkiä auttamatta myöhässä aikatauluista. Silloinen esimies **Timo Alanko** ehdotti, että Tilastopäivät siirrettäisiin seuraavaan syksyyn tai kevääseen 2013, jolloin voitaisiin kenties samalla siirtyä biennaalimalliin. Havaitsimme pian, että syksy on hankala ajankohta. Emme myöskään halunneet järjestää Tilastopäiviä kolmatta kertaa peräkkäin Helsingissä, joten otimme vuoden aikalisän. Tästä sai alkunsa koko seuran toimintamallin muutos kaksivuotiseen rytmitykseen. Se on sittemmin osoittautunut toimivammaksi nykyisen työelämän hektisyydessä, jossa kaikki seuran aktiviteetit hoidetaan *oto* eli oman toimen (mikäli sellainen on) ohella.

Samoihin aikoihin, vuoden 2012 alkupuolella, käynnistyi toinen esimieskauttani värittänyt, mielenkiintoinen, harvinaislaatuinen ja todella monisäikeinen prosessi: **Hannu Niemen** muotokuvan maalauttaminen ja lahjoittaminen Helsingin yliopiston (Suomen laajimpaan, toim.huom.) muotokuvakokoelmaan. Yliopistolla oli syksyllä 2011 herätty huomaamaan, ettei moiseen tarkoitukseen ole luvallista kerätä rahaa *(sic)*, vaan keräys tulee organisoida jonkin tieteellisen seuran kautta. **Risto Lehtonen** oli neuvonut ottamaan yhteyttä Suomen Tilastoseuraan, jonka varaesimies löytyi samalta laitokselta. Monien vaiheiden jälkeen muotokuva julkistettiin juhlallisesti ja yliopiston protokollaohjeita noudattaen helmikuussa 2013, ja saman vuoden kesänä prosessi sai lopullisen päätöksensä, kun se kaikkine yksityiskohtineen raportoitiin rahankeruulupia hallinnoineelle viranomaiselle eli Helsingin poliisilaitokselle.

Heti ensimmäiselle hallitukselleni totesin vuosikirjan kokoamisen ja Tilastopäivien järjestelyiden olevan niin pahasti kesken, että oli syytä alkaa harkita biennaalimallia.

Päätin katsaukseni linjaukseen: ”*Haluan, että seuran hallituksesta saadaan toimiva kokonaisuus, jossa jokainen pääsee osallistumaan erilaisiin tehtäviin ja esittämään näkemyksiään siitä, millaiseksi tätä pian 100 vuotta täyttävää seuraa kannattaisi muovata, jotta se toimisi järkevästi tässä nykyisessä, hektisessä maailmassa. Osa perinteistä joutaa romukoppaan, osaa on hyvä kunnioittaa ja kaikkea voi muuttaa.*”

Biennaalimalli kiteytyi sittemmin paperiin, jossa viitoitin tiekarttaa kohti seuramme 100-vuotisjuhlia. Paperin päivitetty versio sisältyy tähän vuosikirjaan. Perinteinen **Leo Törnqvistin** nimeä kantava gradupalkinto muuttui jaettavaksi kahden vuoden välein ja sai rinnalleen **Samuli Ripatin** aloitteesta synnytetyn, joka neljäs vuosi jaettavan, väitöskirjapalkinnon. Molemmat palkinnot jaettiin uudessa muodossaan kevään 2013 Tilastopäivillä Jyväskylässä. Ehdokkaita oli enemmän kuin koskaan, ja nekin julkistettiin ensi kertaa. Palkinnon saaneet esittelevät työnsä tässä vuosikirjassa.

Käytännön toiminnan sujuvoittamiseksi oli kaksi asiaa, jotka oli ehdottomasti saatava siihenastista paremmalle tolalle: ***1) verkkosivut*** ja ***2) taloushallinto***. Jälkimmäisen modernisointi oli usean hallituskauden mittainen rupeama, josta päävastuun kantoivat **Maria Valaste**, **Leena Kalliovirta** ja **Kaisa Mäntysaari**. Verkkosivujen sisällön päivitys tehtiin hallituksen sisäisenä ryhmätyönä, mutta sivuston rakenteen, ulkoasun ja hallinnointivälineiden ajantasaistamiseen päätettiin hakea ulkopuolinen tekijä.

Sonja Lumme selvitti mahdollisia toimijoita, mutta hintataso tuntui olevan seuran ulottumattomissa. Keksin lähestyä entistä graduohjattavaani **Kati Tiirikaista**, jonka tiesin toimivan alalla, ja lähetin hänelle viestin: ”*Haluaisimme uudistaa homehtuneet webbisivumme, muttemme ole innokkaita maksamaan siitä järjettömiä konsulttipalkkioita, emme etenkään mitään kk-maksuja. Ajatus olisi saada selkeä sivusto, jota voimme ylläpitää omin neuvoin”.* Kati otti haasteen vastaan ja uudisti sivumme *www.tilastoseura.fi* nimellisellä palkkiolla ja seuran ainaisjäsenyydellä.

Tästä vuosikirjasta löytyy esimiehen palstan kirjoittamisen ohessa survomani kuva Tieteellisten seurain valtuuskunnan (TSV) jäsenseuroista. Tilastoseuran jäsenmäärä on kohtuullisen suuri verrattuna moniin vastaaviin seuroihin. Etenkin, jos vertaillaan Euroopan laajuisesti ja suhteutetaan vielä maan väestömäärään, olemme yllättävänkin suuri seura. Tämän tajusin itse vasta **Maailman tilastovuonna 2013**, edustaessani seuraa **ISI**:n 59. kokouksessa Hongkongissa, jossa perustimme uuden eurooppalaisen kattojärjestön **FENStatS**. Uusia jäseniä kannattaa edelleen rekrytoida aktiivisesti, etenkin tilastotieteen monilta sovellusaloilta. Myös tilastotieteen pääaineopiskelijoita on hyvä saada mukaan, jotta he pysyvät hyvässä seurassa myös valmistuttuaan.

Kiitän kaikkia yhteistyöstä ja toivotan menestystä nykyiselle ja tuleville hallituksille!

SIHTEERIN KOMMENTTI

Kaisa Mäntysaari (M.Sc.)

Kahteen kuluneeseen vuoteen Suomen Tilastoseuran sihteerinä on mahtunut monenlaista mukavaa. Tilastopäivien, iltapäiväseminaarien ja vuosikirjan lisäksi työn alla on ollut mm. nettisivujen uudistus. Seuran tiedottaminen onkin enenevässä määrin siirtynyt verkkoon ja nettisivujen päivittäminen on ollut yksi sihteerin päätehtävistä jäsenrekisterin ylläpidon ohella.

Seuran sivuilla tiedotetaan tätä nykyä ajankohtaisista tilastoalan tapahtumista ja avoimista työpaikoista. Nettisivuilta löytyvät myös mm. hallituksen kokousten pöytäkirjat ja jäsenkirjeet sekä Tilastopäivien ja iltapäiväseminaarien esitysmateriaalit. Myös seuraan liittyminen ja yhteystietojen päivittäminen tapahtuu nettisivuilta löytyvien lomakkeiden avulla. Osoite www.tilastoseura.fi on siis se osoite, jonka jokaisen jäsenen on syytä muistaa.

Haluan kiittää yhteistyöstä seuran entistä ja nykyistä esimiestä Kimmo Vehkalahtea ja Jyrki Möttöstä. Lämmin kiitos kaikesta avusta myös rahastonhoitaja Leena Kalliovirralle sekä tehtävään perehdyttäneelle edeltäjälleni Marjo Pyy-Martikaiselle. Kiitos myös kaikille molempien kausien hallitusten jäsenille. Erityiskiitokset Kati Tiirikaiselle uusien nettisivujen suunnittelusta, toteutuksesta ja käyttötuesta sekä Hilkka Lehtoselle tämän vuosikirjan taitosta.

UUDEN HALLITUKSEN TERVEHDYS

Jyrki Möttönen
Suomen Tilastoseuran esimies 2014

Ensimmäiseksi haluan kiittää Kimmo Vehkalahtea hänen monivuotisesta työstään Suomen Tilastoseuran hallituksessa. Kimmon esimiesaikana Tilastoseuran toiminnassa siirryttiin biennaalirytmitykseen eli Tilastopäivät järjestetään ja vuosikirja julkaistaan kahden vuoden välein. Esimiehen palstalla Kimmo selvittää tähän kaksivuotismalliin siirtymisen syitä ja taustoja.

Tilastopäivät

Tilastopäivät järjestettiin Jyväskylän yliopistossa Mattilanniemen kampusalueella 30.-31.5.2013 otsikolla *Statistics for environment, ecology and forestry.* Keynote-puhujana oli Alan E. Gelfand, Duke University ja kutsuttuina puhujina olivat Juha Heikkinen Metlasta, Otso Ovaskainen Helsingin yliopistosta ja Lauri Mehtätalo Itä-Suomen yliopistosta. Järjestelyistä vastasivat Suomen Tilastoseura ja Jyväskylän yliopisto (Matematiikan ja tilastotieteen laitos).

Tilastopäivien konferenssipäivälliset pidettiin ravintola Piatossa lähellä konferenssipaikkaa.

SAS Instituutin taloudellinen tuki päivien järjestämiselle oli merkittävä ja Tilastoseura on saadusta tuesta erittäin kiitollinen.

Ensi vuonna Tilastopäivät järjestetään Helsingissä ja aihepiiriksi on suunniteltu isoihin havaintoaineistoihin liittyviä tilastollisia menetelmiä. Tilastopäivien tarkempi ohjelma selviää lähikuukausien aikana ja siitä tiedotetaan Tilastoseuran jäsenkirjeessä.

Palkinnot

Seura jakaa joka toinen vuosi Leo Törnqvist -palkinnon parhaalle suomalaisessa yliopistossa tai korkeakoulussa hyväksytylle tilastotieteen pro gradu -tutkielmalle. Tilastoseura valitsi parhaaksi vuosina 2011–2012 tehdyksi pro gradu -tutkielmaksi Tytti Pasasen (Tampereen yliopisto) työn.

“Two-Level Structural Equation Modeling with Non-Normal Observed Variables for Assessing Poverty in Laos”. Ensimmäistä kertaa jaettavan väitöskirjapalkinnon voittaja oli Jukka Sirénin (Helsingin yliopisto) väitöskirja “Statistical models for inferring the structure and history of populations from genetic data”. Väitöskirjapalkinto jaetaan parhaalle neljän edellisen vuoden aikana yliopistossa tai korkeakoulussa hyväksytylle tilastotieteen väitöskirjalle.

Iltapäiväseminaarit

Iltapäiväseminaareja järjestetään edellisten vuosien tapaan muutaman kerran vuodessa kiinnostavien aiheiden esilletullessa ja halukkaiden järjestäjien löytyessä. Edellisen vuosikirjan ilmestymisen jälkeen on järjestetty neljä iltapäiväseminaaria.

Iltapäiväseminaari *Statistical and computational challenges in Large-Scale genomic Epidemiology* järjestettiin Biomedicumissa 8.10.2012. Luennoijat olivat Johannes Ket-

tunen (FIMM), Michael Inouye (University of Melbourne), Matti Pirinen (FIMM), Jukka Corander (Helsingin yliopisto). Kommenttipuheenvuoron esitti Juni Palmgren (FIMM). Järjestelyistä vastasivat FIMM yhdessä Tilastoseuran kanssa.

Iltapäiväseminaari *Mitä tilastot ja rekisterit kertovat tämän päivän nuorista?* järjestettiin 23.4.2013 Helsingin taloushallintopalvelun Saldo-auditoriossa. Luennoijat olivat

Reija Paananen (THL), Elise Haapamäki (Helsingin kaupungin tietokeskus), Liisa Larja (Tilastokeskus), Seija Saari (Helsingin kaupungin tietokeskus) ja Vesa Keskinen (Helsingin kaupungin tietokeskus). Järjestelyistä vastasivat Helsingin kaupungin Tietokeskus yhdessä Suomen Tilastoseuran kanssa.

Iltapäiväseminaari *Tilastojen luku- ja käyttötaito* järjestettiin Helsingin taloushallintopalvelun Saldo-auditoriossa 5.11.2013. Luennoijat olivat Jussi Melkas (YTL), Minna Torppa **(**Forum Virium Helsinki), Reija Helenius **(**Tilastokeskus), Riikka Muje (Lyseonpuiston lukio, Rovaniemi). Järjestelyistä vastasivat Helsingin kaupungin Tietokeskus yhdessä Suomen Tilastoseuran kanssa.

Iltapäiväseminaari *Kokemuksia sähköisistä tiedonkeruumenetelmistä* järjestettiin Helsingin kaupunkisuunnitteluviraston info- ja näyttelytilassa Laiturissa 10.6.2014. Luennoitsijat olivat Kirsti Pohjanpää (Tilastokeskus), Tara Junes (Tilastokeskus), Sini Askelo (Helsingin kaupungin tietokeskus), Vesa Keskinen (Helsingin kaupungin tietokeskus) ja Maija Mattila (Helsingin kaupunkisuunnitteluvirasto). Kommenttipuheenvuoron esitti Risto Lehtonen (Helsingin yliopisto). Järjestelyistä vastasivat Helsingin kaupungin Tietokeskus yhdessä Suomen Tilastoseuran kanssa.

Marraskuussa 2014 järjestetään seuraava iltapäiväseminaari, jonka järjestelyistä vastaa Suomen Syöpärekisteri yhdessä Tilastoseuran kanssa.

Suomen Tilastoseura: KOHTI 100-VUOTISJUHLIA

Kimmo Vehkalahti
3.11.2012 (30.10.2014)

Aikataulua vuosiksi 2014–2020

Vuosi	Tilastopäivät[1]	Palkinnot[2]		Vuosikirja[3]
		Pro gradu[4]	Väitöskirja[5]	
2009	Kuopio	2008		2008
2010	Helsinki	2009		2009
2011	Helsinki	2010		2010
2012[6]				2011–2012
2013	Jyväskylä	2011–2012	2009–2012	
2014				2013–2014
2015	Helsinki	2013–2014		
2016				2015–2016
2017		2015–2016	2013–2016	
2018				2017–2018
2019		2017–2018		
2020[7]				2019–2020
2021		2019–2020	2017–2020	
. . .				

[1] Järjestetään keväisin (toukokuussa).
[2] Jaetaan Tilastopäivien yhteydessä.
[3] Julkaistaan loppusyksystä (marraskuussa).
[4] Leo Törnqvist -palkinto (vuodesta 1978).
[5] Väitöskirjapalkinto (vuodesta 2013).
[6] Seuran toimintarytmin muutos (”biennaalimalli”).
[7] Seura 100 vuotta (juhlaseminaari Helsingissä?).

TILASTOPÄIVÄT 2013
Statistical Days 2013

Topic: Statistics for environment, ecology and forestry
Dates: May 30–31, 2013
Venue: University of Jyväskylä
Organizers: Finnish Statistical Society and University of Jyväskylä (Department of Mathematics and Statistics)

DAY 1 – Thursday, May 30, 2013

11:00-11:55 Registration
12:00-12:15 Opening
Kimmo Vehkalahti, President of the Finnish Statistical Society

Keynote Session
Chair: *Antti Penttinen*, University of Jyväskylä

12:15-13:00 Wildfires in South Africa; Cherry trees in Japan
Professor Alan E. Gelfand, Duke University
13:00-13:15 Break
13:15-14:00 Point pattern modeling for degraded presence-only data over large regions
Professor Alan E. Gelfand, Duke University
14:00-14:30 Coffee break

Invited Session I
Chair: *Juha Karvanen*, University of Jyväskylä

14:30-15:15 Assessing uncertainties in national greenhouse gas inventory
Professor Juha Heikkinen, Metla
15:15 -15:25 Break

Invited Session II
Chair: *Jyrki Möttönen*, Vice President of the Finnish Statistical Society

15:25-16:10 Applications of Bayesian statistics in ecology and evolutionary biology
Professor Otso Ovaskainen, University of Helsinki
16:10-16:20 Break

Contributed Session I
Chair: *Jyrki Möttönen*, Vice President of the Finnish Statistical Society

16:20-16:40 Bayesian scale space analysis of temporal changes in satellite images
Leena Pasanen, University of Oulu
16:40-17:00 Testing spatial hypotheses for marked spatial point patterns
Mari Myllymäki, Aalto University

Sponsor's address and Conference dinner
17:00-17:30 Forecasting with SAS Forecast Server
Nina Survo, SAS Institute
19:00-22:00 Conference dinner: Restaurant Piato, Agora

DAY 2 – Friday, May 31, 2013

Contributed Session II
Chair: *Leena Kalliovirta*, Treasurer of the Finnish Statistical Society

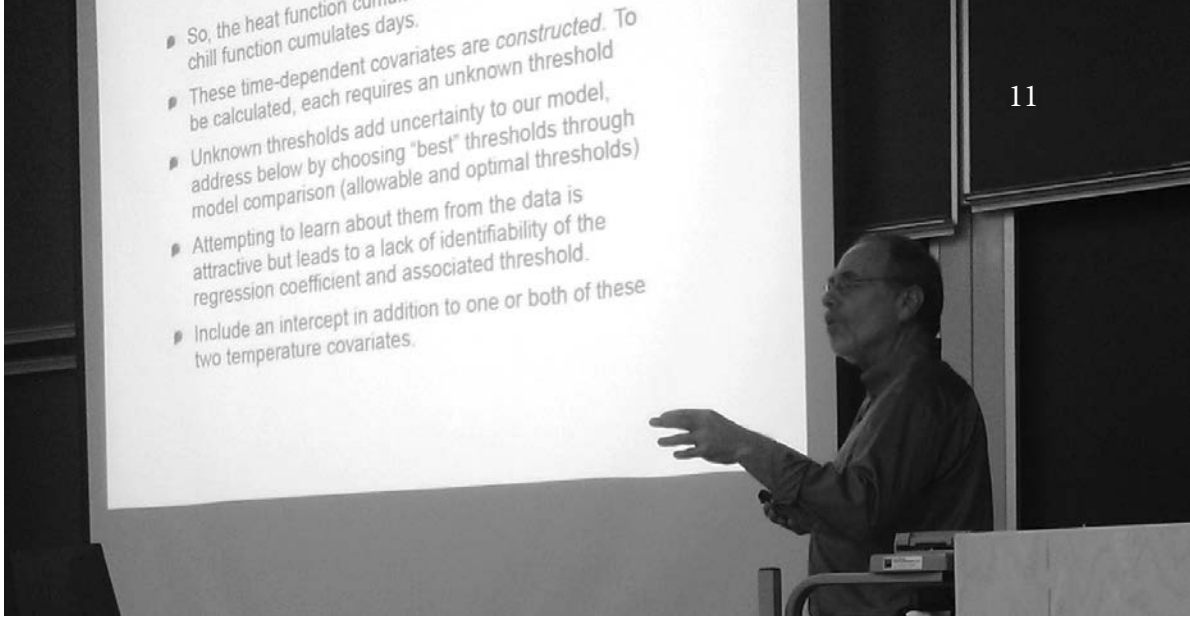
09:00-09:20 Towards more cost-effective identification of freshwater macroinvertebrates
Johanna Ärje, University of Jyväskylä
09:20-09:40 Estimating aggregated nutrient fluxes in four Finnish rivers
via Gaussian state space models
Jouni Helske, University of Jyväskylä
09:40-09:45 Break

Invited Session III
Chair: *Jukka Nyblom*, University of Jyväskylä

09:45-10:30 Applications of mixed effects models in forestry
Dr Lauri Mehtätalo, University of Eastern Finland

Poster Session
10:30-11:10 Poster presentations with coffee

Final Session
11:15-11:30 Leo Törnqvist Prize: The Best Master's Thesis in Statistics (2011–2012)
11:30-11:45 Dissertation Prize: The Best Doctoral Thesis in Statistics (2009–2012)
11:45-12:00 Final Discussion and Closing



Wildfires in South Africa; Cherry Trees in Japan

Alan E. Gelfand,
Department of Statistical Science, Duke University

Abstract

We consider two challenging ecological problems and show how they are naturally investigated using spatial survival models with time varying covariates.

The first problem involves explanation of the occurrence of wild fires. This process is of interest because over half of the world's terrestrial ecosystems depend on fire to maintain ecological structure and function and the ecological role of fire regimes can be strongly influenced by weather and climate. To undertake this analysis, we developed an extensive database of observed fires with high-resolution meteorological data to explore fire regimes in the Mediterranean ecosystem in the Cape Floristic Region (CFR) of South Africa during the period 1980–2000. We need to consider the influence of seasonally (quarterly) anomalous weather on fire probability. In addition to these local-scale influences, the Antarctic Ocean Oscillation (AAO) is a potentially important large scale influence with regard to global circulation patterns.

The second involves explaining first flowering times. The objective here is to learn about changes in the length and onset of the growing season. This process has to be examined at individual tree/plant level and in response to weather, in particular daily temperature, rather than aggregating to climate. We are broadly interested in comparison of first flowering time (or bud burst) across species but here we focus on explaining spatial variation in first flowering time. We consider first flowering dates for trees of a single species in Japan at 45 locations over 52 years, collected through 2009. The

challenge with this process is to provide suitable functions of the weather – heating and chilling functions – to employ in the explanation. The difficulty is that these functions are not explicitly defined since they require measurement beginning from unknown starting dates as well as unknown thresholds. We have uncertainty in the specification of the functional covariates.

We present both analyses and our findings along with some future challenges.

Point pattern modeling for degraded presence-only data over large regions

Alan E. Gelfand
Dep't of Statistical Science, Duke University

Explaining species distribution using local environmental features is a long standing ecological problem. Often, available data is collected as a set of presence locations only thus precluding the possibility of a presence-absence analysis. We propose that it is natural to view presence-only data for a region as a point pattern over that region and to use local environmental features to explain the intensity driving this point pattern. This suggests hierarchical modeling, treating the presence data as a realization of a spatial point process whose intensity is governed by environmental covariates. Spatial dependence in the intensity surface is modeled with random effects involving a zero mean Gaussian process. Highly variable and typically sparse sampling effort as well as land transformation degrades the point pattern so we augment the model to capture these effects. The Cape Floristic Region (CFR) in South Africa provides a rich class with such species data. The potential, i.e., nondegraded presence surfaces over the entire area are of interest from a conservation and policy perspective.

Our model assumes grid cell homogeneity of the intensity process where the region is divided into ~ 37, 000 grid cells. To work with a Gaussian process over a very large number of cells we use predictive process approximation. Bias correction by adding a heteroscedastic error component is implemented. The model was run for a number of different species. Model selection was investigated with regard to choice of environmental covariates. Also, comparison is made with the now popular Maxent approach, though the latter is much more limited with regard to inference. In fact, inference such as investigation of species richness immediately follows from our modeling framework.

Assessing uncertainties in national greenhouse gas inventory[1]

Juha Heikkinen
Finnish Forest Research Institute (Metla)
juha.heikkinen@metla.fi

Introduction

United Nations Framework Convention on Climate Change and its Kyoto Protocol require from participating countries annual reports of human-induced emissions and removals of greenhouse gases (GHG) to and from the atmosphere, because ”Accurate, consistent and internationally comparable data on GHG emissions is essential for the international community to take the most appropriate action to mitigate climate change…” (unfccc.int, National Reports). While Statistics Finland is the national responsible unit for Finland’s GHG-inventory (www.stat.fi/tup/khkinv), the role of Finnish Forest Research Institute (Metla) is to assess sources and sinks of GHG in Land use, Land-use change and Forestry (LULUCF) sector (www.metla.fi/ghg). The other sectors are Energy, Industrial processes, Solvents, Agriculture, and Waste.

Quantification of uncertainties is an essential part of the inventory, with a view of directing the resources available for research toward reducing the uncertainties over time (IPCC 2000, p. 1.4). Specifically, the 2006 guidance for GHG-inventories introduces the concept of *key category* to identify, within sectors, the categories (or within categories, the pools) which have a significant influence on a country’s total inventory in terms of the absolute level, trend, or uncertainty in emissions and removals. These categories (or pools) should be the priority for countries during inventory resource allocation for data collection, compilation, quality assurance/quality control and reporting (IPCC 2006, Vol 1. p. 1.6, Table 4.1).

This paper reports some experiences from ongoing research aimed at improving the assessment of uncertainty in GHG-inventory for Forestry sector. The inventory is based on a combination of sample- and model-based inferences. The focus here is on the kind of issues that can also be of interest in other contexts, where changes in

[1] Based on joint research with Aleksi Lehtonen and Jaakko Repola from Finnish Forest Research Institute (Metla), and Göran Ståhl, Hans Petersson, and Sören Holm from Swedish University of Agricultural Sciences (SLU)

population totals or means are estimated on such basis. In particular, two carbon pools are considered, living tree biomass and the aggregate pool DOM+SOM containing dead wood, litter, and soil organic matter. The changes in these pools contribute vast majority of CO_2 emissions and removals within category Forest Land Remaining Forest Land.

Background

Guidelines for GHG-inventories identify two approaches for estimating changes in carbons stocks (IPCC 2006, Vol 4. 2.2.1). Finnish inventory applies the *Gain-Loss Method* for living tree biomass in Forest Land Remaining Forest land by subtracting the estimated annual drain due to loggings and natural mortality from the estimated increment due to tree growth. Swedish inventory, on the other hand, applies the *Stock-Difference Method* based on differences between subsequent estimates of tree biomass stock.

National forest inventories (NFI) provide reliable estimates of tree stem volume and its increment (e.g., Tomppo et al. 2011, Korhonen et al. 2013). In Finnish NFI, breast height diameter *d* is measured from more than 400 000 trees (*tally trees*) during the five years' rotation, and height *h*, upper (6m) diameter and increment (growth) from every 7th (*sample trees*). Uncertainty in volume models based on these measurements can usually be neglected, and the conversion from stem volume to stem biomass is relatively reliable.

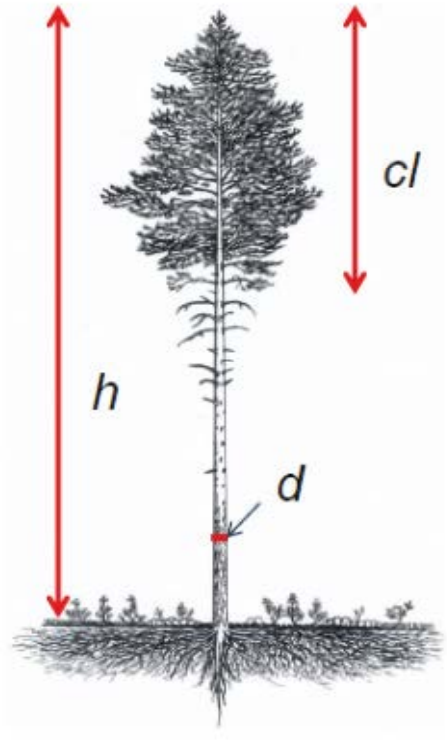


Figure 1. Some tree measurements in NFI.

Other tree biomass compartments, including branches, foliage (needles or leaves), and roots are more problematic. Measurement of biomass (dry weight) is laborious and destructive. Species- and compartment-specific models have been developed based on a set of moderate number of trees, which cannot exactly be considered as a probability sample from Finnish forests (Repola 2008, 2009). The predictors in these regression models are tree-level variables measured in NFI such as *d*, *h*, and crown length *cl* (Figure 1). Different versions of the models are available, using different sets of predictors, with simplest ones based on *d* only.

Carbon inputs to DOM+SOM pool consist of litterfall from trees, residues from logging, and natural mortality (Figure 2). This input, together with weather

information, drives soil model Yasso07 (Tuomi et al. 2009), which is used in the Finnish inventory to assess the changes in DOM+SOM carbon.

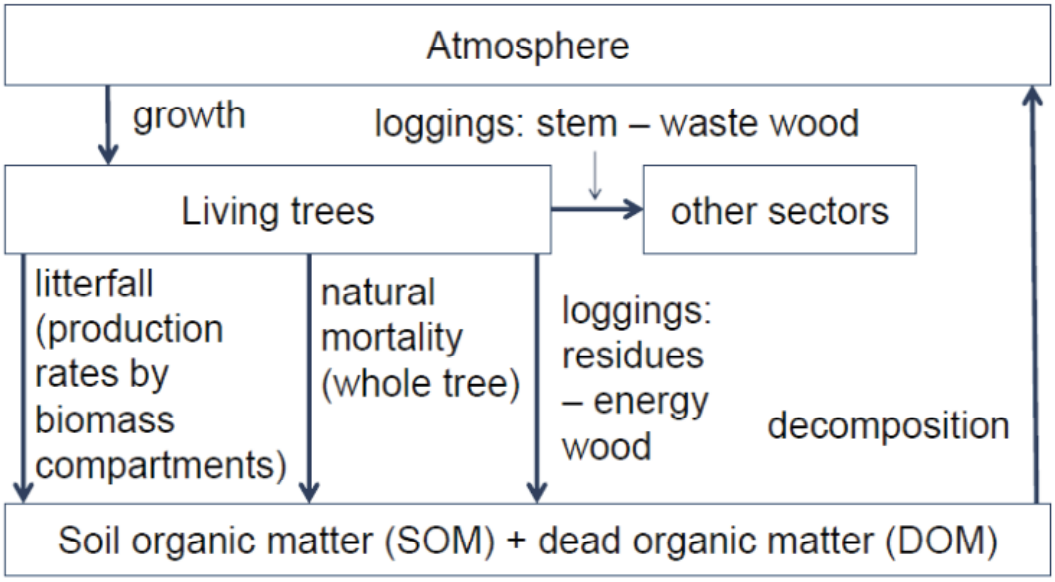


Figure 2. Some CO_2 (or carbon) fluxes to and from category Forest Land Remaining Forest Land, and between its two main pools.

Litter input from living trees is estimated by multiplying current stocks (Figure 3a) with compartment-specific annual litter production rates based on experiments, monitoring and longenvity studies (e.g., how many needle cohorts can be found at one time). For example, litter production rate of pine stem is 0.0052, while that of pine needles is 0.245 and of birch leaves 0.79 (not 1, since a part of the leave biomass is 'captured' by the branches before the leaf falls). From the viewpoint of estimating the litterfall, foliage is thus much more important than stemwood, but its estimates are also much less precise.

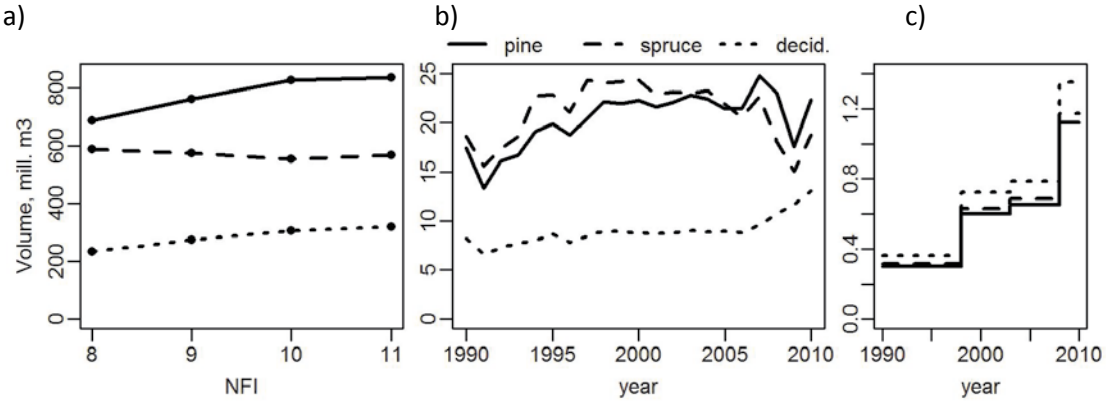


Figure 3. Stem volume of living trees from NFI8 (1986-94) to NFI11 (2009-13) (a), annual logging volumes (b), and volume of natural mortality (c) on mineral soil forest land.

Logging volumes are available from annual statistics on commercial wood removals (Figure 3b). Residues from loggings contain that part of the logged stems, which cannot be utilized (*waste wood*), and the whole biomass of other biomass compartments, except for that portion which is used as energy wood. Again, other less well known compartments are more important than the stem.

Only occasional studies are available of the volume of natural mortality, the most recent based on the volume of trees that died between two measurements of permanent NFI plots. They indicate an increasing trend (Figure 3c), but the precision of the estimates is known to be rather poor. On the other hand, the proportional contribution of natural mortality to the total DOM+SOM input is quite small (Figure 4).

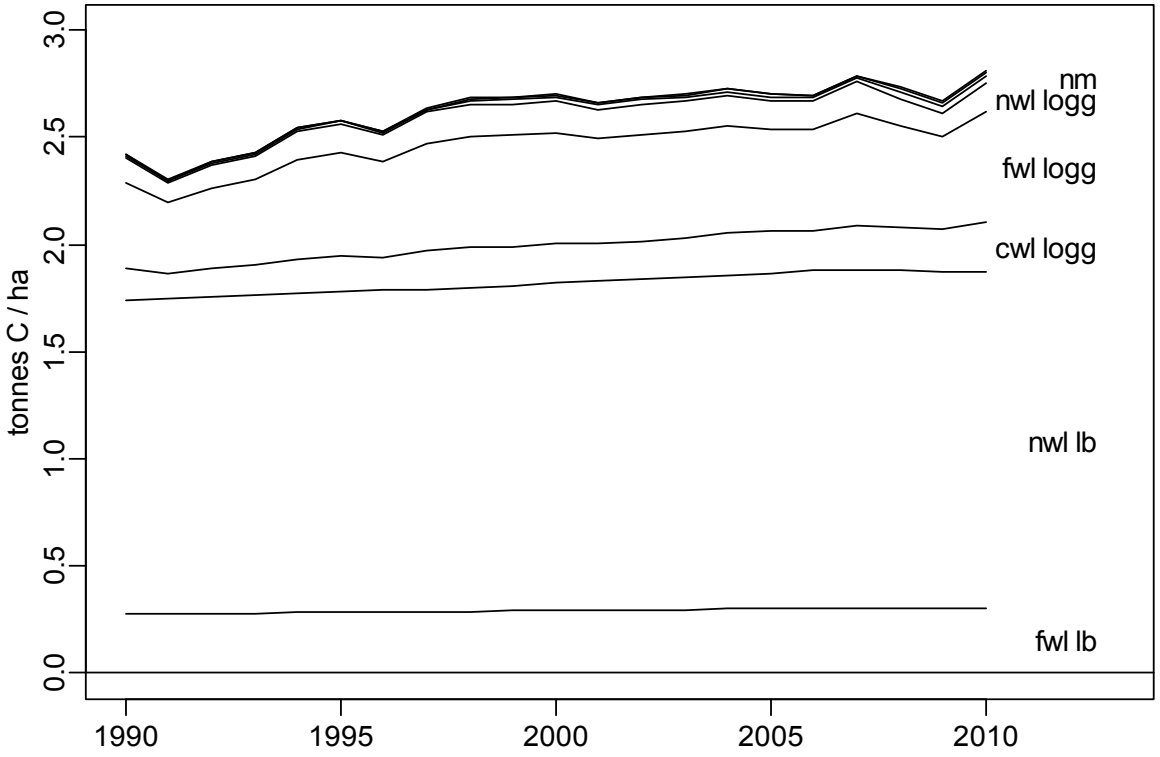


Figure 4. Input of coarse woodly litter (cwl), fine woody litter (fwl), and non-woody litter (nwl) from living trees (lb), loggings (logg), and natural mortality (nm) to DOM+SOM.

Living tree biomass

In Ståhl et al. (2014), we developed a method for quantifying the combined uncertainty stemming from sampling and model errors in the NFI-based estimates of the biomass stock of living trees, and its changes. For simplicity, we focused on Stock-Difference Method. Essentially, the mean stock per hectare over a given region, $\overline{Y}$, is estimated using a suitably scaled weighted mean, $\hat{y}$, of model-predicted biomasses of sample trees in the NFI sample from that region. Letting $\overline{y}$ denote the (unknown) true mean biomass over the sample we can decompose estimation error, $\hat{y} - \overline{Y}$, into model error $\hat{y} - \bar{y}$ and sampling error $\overline{y} - \overline{Y}$, the variances of which were estimated, respectively, based on the modelling data and variability of biomass predictions in the NFI data.

One problem encountered, which is quite typical, was that sufficient information for estimating the model error was not provided in the model publications. For Swedish models (Marklund 1988) only the residual variation around the fitted model was reported, and the covariance between parameter estimates was also missing in Repola (2008, 2009). In the current context, residual variation is not very influential, because NFI samples are large. On the other hand, uncertainty in model parameter estimates is much more consequential, because the same model with the same estimation error is applied to all trees of the same species. In other words, random errors in parameter estimates, stemming from random selection of modelling data, feature as systematic errors in model-based inventory.

To appreciate the importance of covariance between parameter estimates, consider the toy examle illustrated in Figure 5. Modelling data sets of ten (x, y) pairs in each were simulated with $x \sim \mathrm{N}(4,1)$ and $y \sim \mathrm{N}(x, 0.5^2)$, and linear model $y = \alpha + \beta x + \varepsilon$ was fitted to each set. In 100 replications, standard deviations of $\hat{\alpha}$ and $\hat{\beta}$ were 0.84 and 0.21, but the standard deviation in predictions for $x = 4$ was only 0.16. This is explained by the strong negative correlation of $\hat{\alpha}$ and $\hat{\beta}$, which is a rule when the x-values in the modelling data are not centred to 0, and by the fact that the prediction point is well within the range of modelling data.

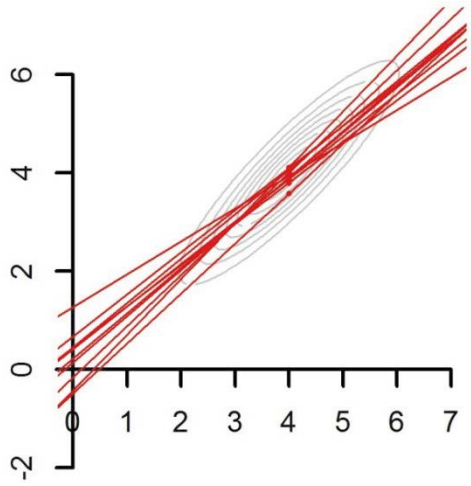


Figure 5. Linear models fitted to data simulated from bivariate normal distribution, whose density is contoured in gray.

The results of Ståhl et al. (2014) indicate that, for the entire aboveground biomass, uncertainty due to sampling is more influential than that due to estimation of model

parameters. The reason for this is that the stem constitutes a large portion of total aboveground biomass, and is well predicted by the models. Furthermore, the role of parameter uncertainty becomes practically negligible when estimating the change in biomass stock. This, in turn, is explained by the fact that the systematic errors resulting from parameter estimation are similar for both of the two stock estimates and therefore largely cancel out in the change estimate. However, model uncertainty is much more crucial for the carbon balance in DOM+SOM, where litter input estimation is based on current stock and largely depends on the biomass of those tree compartments, which are less well predicted by the models than the stem.

DOM+SOM

Yasso07 supports a Monte Carlo facility combining the uncertainty in its own parameters with that provided for the inputs in order to produce a posterior distribution of soil carbon budgets (Tuomi et al. 2009). To that end, we need to be able to generate simulated litter input series, whose distribution reflects the associated uncertainty.

The general form of litter inputs for year t is

$$L_{t,g} = \sum_{k} \sum_{c,s \in g} p_{s,k,c}\, b_{s,t,k,c} V_{s,t,k},$$

where

$g \in$ {cwl,fwl,nwl } indicates the type of litter (e.g., foliage and fineroots are non-woody litter),
s indicates the source of litter: living trees, loggings, or natural mortality,
k indicates the tree species,
c indicates the tree compartment (stem, branches, etc.),
p is the litter production rate
b is the conversion and expansion factor of stem volume to biomass, and
V is the total stem volume.

The total volumes of living trees are obtained directly from NFI and assumed to be free from model errors. The compartment-specific expansion factors are estimated from the NFI sample trees. For living trees, they are estimated separately for each NFI rotation, whereas for loggings and natural mortality, one set of factors, estimated from trees removed between successive measurements of permanent sample plots, is applied for the whole series of volumes.

Sources of uncertainty in $L_{t,g}$ include

NFI sampling affecting V's of living trees and natural mortality, and all b's,
possibly systematic error in logging volumes,
parameter uncertainty in biomass models affecting all b's, and
uncertainty in litter production rates p.

Strong and complex correlations are induced to the estimates of $L_{t,g}$ for different t's and g's. For example

for given s, t, and k, the same sample trees contribute to factors b of all compartments c

same volume estimates are applied for all c,

for nm and logg, same b's are applied for each t

volume estimates for different species are negatively correlated, etc.

We could largely avoid dealing with these correlations analytically by simply simulating p's, b's, and V's and aggregating simulated $L_{t,g}$'s from them. However, the number of values to be simulated per each source and time point would then be 16-fold in comparison to simulating the $L_{t,g}$'s directly.

Our work in progress hence involves analytic derivation of the covariance matrix of the vector of estimates of $L_{t,g}$, taking into account all known sources of uncertainty, and its implementation to the operational GHG-inventory system. Preliminary results do exist, and again they confirm the importance of taking correlations properly into account (Figure 6).

Conclusion

In GHG-inventory the exact values for uncertainty are not as important as the identification of the key categories, the components that are both influential and uncertain. To that end, proper handling of correlations is often more essential than accurate estimates of standard errors.

References

IPCC 2000. Penman, J & al. (eds). Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories. Intergovernmental Panel on Climate Change (IPCC), IPCC/OECD/IEA/IGES, Hayama, Japan.

IPCC 2006. Eggleston, H.S. & al. (eds). 2006 IPCC Guidelines for National Greenhouse Gas Inventories. IGES, Japan.

Korhonen, K.T., Ihalainen, A., Viiri, H., Heikkinen, J., Henttonen, H., Hotanen, J.-P., Mäkelä, H., Nevalainen & S., Pitkänen, J. 2013. Suomen metsät 2004-2008 ja niiden kehitys 1921-2008. Metsätieteen aikakauskirja 2013: 269-608.

Marklund, L.G. 1988. Biomass functions for pine, spruce and birch in Sweden. Swedish University of Agricultural Sciences, Department of Forest Survey, Report 45.

Repola, J. 2008. Biomass equations for birch in Finland. Silva Fennica 42: 605-624.

Repola, J. 2009. Biomass equations for Scots pine and Norway spruce in Finland. Silva Fennica 43: 625-647.

Ståhl, G., Heikkinen, J., Petersson, H., Repola, J. & Holm, S. 2014. Sample based estimation of greenhouse gas emissions from forests - a new approach to account for both sampling and model errors. Forest Science 60: 3-13.

Tomppo, E., Heikkinen, J., Henttonen, H.M., Ihalainen, A., Katila, M., Mäkelä, H., Tuomainen, T. & Vainikainen, N. 2011. Designing and conducting a forest inventory - case: 9th National Forest Inventory of Finland. Springer, Managing Forest Ecosystems 21.

Tuomi, M., Thum, T., Järvinen, H., Fronzek, S., Berg, B., Harmon, M., Trofymow, J.A., Sevanto, S. & Liski, J. 2009. Leaf litter decomposition - Estimates of global variability based on Yasso07 model. Ecological Modelling 220: 3362-3371.

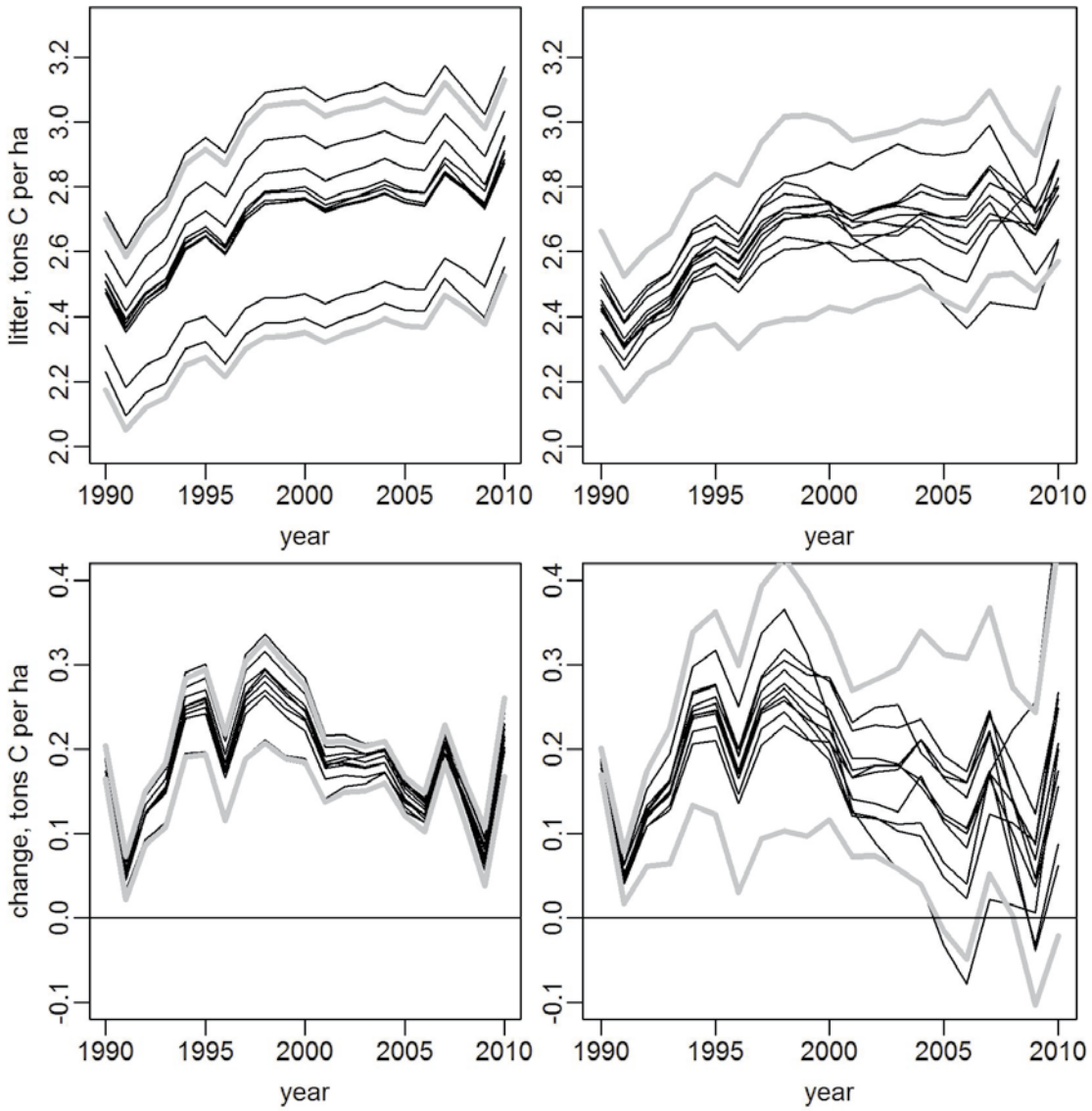


Figure 6. Top: 10 simulated series of total litter input for South Finland (black) and confidence intervals (gray). In the right, temporal correlations have been neglected. Bottom: The same for soil carbon change simulated by Yasso07 using the inputs of top row.

Applications of Bayesian statistics in ecology and evolutionary biology

Otso Ovaskainen

In observational studies of ecology and evolutionary biology, the process of interest is seldom observed directly, and thus inference relies on data that are indirect and influenced by an observation process. Often observational data also have a spatial or temporal component, bringing an additional level of complexity for the analysis phase. Bayesian state-space modeling has become an increasingly popular approach in ecology and evolutionary biology because it allows for the separation of the biological process from the observation process, and because it can incorporate hierarchical structures needed to account for different levels of biological organization. I illustrate the use of Bayesian approaches in ecology and evolutionary biology with the help of case studies that relate to animal movement, spatial population biology, community ecology and quantitative genetics. The underlying mathematical approaches involve diffusion-advection-reaction models (animal movement), individual-based models tailored to specific systems (spatial population biology), linear models with residual correlations structured by space, time or relatedness (spatial population biology and quantitative genetics), and hierarchical models that allow one to draw inferences on large but sparse data sets (community ecology).

Bayesian Scale Space Analysis of Temporal Changes in Satellite Images

Leena Pasanen and Lasse Holmström
Department of Mathematical Sciences, University of Oulu, Finland

1 Introduction

We consider the detection of land cover changes using pairs of Landsat ETM+ satellite images that consist of eight spectral bands. In order to simplify this multidimensional change detection task, the image pair is first transformed to a one dimensional image. However, when the transformation is non-linear, the true change in the images may be masked by complex noise. For example, when changes in the normalized difference vegetation index (NDVI) is considered, the variance of the noise may not be constant over the image and methods based on image thresholding could be ineffective.

We use Bayesian statistical modeling for the change detection. However, as the posterior distribution of the transformed image might be difficult to obtain, we estimate it by first obtaining the joint posterior distribution of the original two noiseless satellite images, drawing a large sample from this posterior and then applying the transformation to the sampled images.

All changes may not be detected when the posterior of the transformed image is used directly. For example, large areas where the intensity has changed only slightly, may not be discernible from noise. Therefore, in order to detect both the high-intensity small scale changes and low-intensity large scale changes, we will apply a scale space approach that employs multi-level smoothing. This means that the changes are detected from the posterior distributions of the smooths of the transformed image.

The method used is called iBSiZer (Bayesian SiZer for images) that was first used for digital images [6]. IBSiZer is an instance of SiZer (SIgnificant ZERo crossings of derivatives) methodology that was first proposed in [1, 2]. For the detection of change in NDVI-difference images iBSiZer was used in [8].

2 The satellite images

Landsat ETM+ satellite images consist of eight image bands, each representing a different wavelength of light. Each band can be considered as an $M \times N$ array of real numbers x_{ij}, but in mathematical derivations we treat them as a vector $\mathbf{x} = [x_1, \ldots, x_n] \in \mathbb{R}^n,\ n = NM$. We combine the bands corresponding to the two instants of time considered into one $16n \times 1$ vector $\mathbf{x} = [\mathbf{x}_{11}^T, \ldots, \mathbf{x}_{18}^T, \mathbf{x}_{21}^T, \ldots, \mathbf{x}_{28}^T]^T$,

where $\mathbf{x}_{ij}$ is the band j at time i. An observed satellite image $\mathbf{y}$ is modeled as

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\varepsilon}, \tag{1}$$

where, $\mathbf{x}$ is the true image and $\boldsymbol{\varepsilon}$ is the corrupting noise.

For a pair of satellite images represented by $\mathbf{v} = [\mathbf{v}_{11}^T, \ldots, \mathbf{v}_{28}^T]^T$ the NDVI difference is computed as

$$\mathbf{N}_\mathbf{v} \equiv \frac{\mathbf{v}_{24} - \mathbf{v}_{23}}{\mathbf{v}_{24} + \mathbf{v}_{23}} - \frac{\mathbf{v}_{14} - \mathbf{v}_{13}}{\mathbf{v}_{14} + \mathbf{v}_{13}}. \tag{2}$$

The true and the noisy NDVI difference images are denoted by $\mathbf{N}_\mathbf{x}$ and $\mathbf{N}_\mathbf{y}$, respectively.

3 The method

In order to detect temporal changes in two NDVI images in multiple scales four steps are taken:

1. The posterior distribution of $\mathbf{x}$ is obtained.
   We assume that the noise has a Gaussian distribution, hence the likelihood function $p(\mathbf{y}|\mathbf{x})$ is a Gaussian density. As the prior distribution of $\mathbf{x}$, we use a Gaussian smoothing prior that penalizes for image roughness as measured by the second differences of neighboring pixel intensities [6]. For Landsat ETM+ images, the prior models the prior temporal dependence in the images corresponding to the same band as well as the smoothness of each band image $\mathbf{x}_{ij}$. This prior model for Landsat ETM+ images is discussed in more detail in [8].

2. The posterior distribution of $\mathbf{N}_\mathbf{x}$ is obtained.
   The posterior $p(\mathbf{N}_\mathbf{x}|\mathbf{y})$ is estimated by drawing a large sample from the posterior of $\mathbf{x}$ and then applying the transformation (2) to each sampled $\mathbf{x}$.

3. The posterior distribution of smooths of $\mathbf{N}_\mathbf{x}$ is obtained with several smoothing levels.
   Lets denote by $\mathbf{S}_\lambda$ a smoothing operator with smoothing level $\lambda \geq 0$. The posteriors $p(\mathbf{S}_\lambda \mathbf{N}_\mathbf{x}|\mathbf{y})$ are estimated by smoothing each sampled $\mathbf{N}_\mathbf{x}$. The smoother we use here is a discrete spine smoother (see [5, 3, 4]).

4. On each smoothing level detect the areas where $\mathbf{N}_\mathbf{x}$ is credibly positive or negative.
   We detect the pixels that have high enough posterior probability of being greater or less than zero. The inference is simultaneous over all pixels of the image and we use the "simultaneous credible intervals" (CI) method that was first proposed for one dimensional data in [3] and then extended for digital images in [7]. The result of each smoothing level is presented as an iBSiZer map where white, black and gray represents the credibly positive, negative and neither pixels.

To illustrate the idea of iBSiZer, we will detect changes in a pair of images based on a real Landsat ETM+ satellite image and manually constructed changes, using the difference of their NDVI images. The test image pair is constructed from a 176×165 subimage of a full Landsat ETM+ satellite image [6].

The observed difference in NDVI images $\mathbf{N_y}$, resulting posterior mean of $\mathbf{N_x}$, and the posterior means and credibility maps with three different smoothing levels are displayed in Figure 1. Note, the strong heteroscedasticity of the noise in $\mathbf{N_y}$. If simple thresholding would be applied to the noisy NDVI-difference image, the true changes would be masked by noise [8]. However, the iBSiZer maps seem to cope with this heteroscedasticity well. The two smallest scale maps detect the small changed areas with high absolute intensity. The large positive area in the lower right corner is detected in the larger scale. Note, that none of the maps alone would reveal all the interesting features.

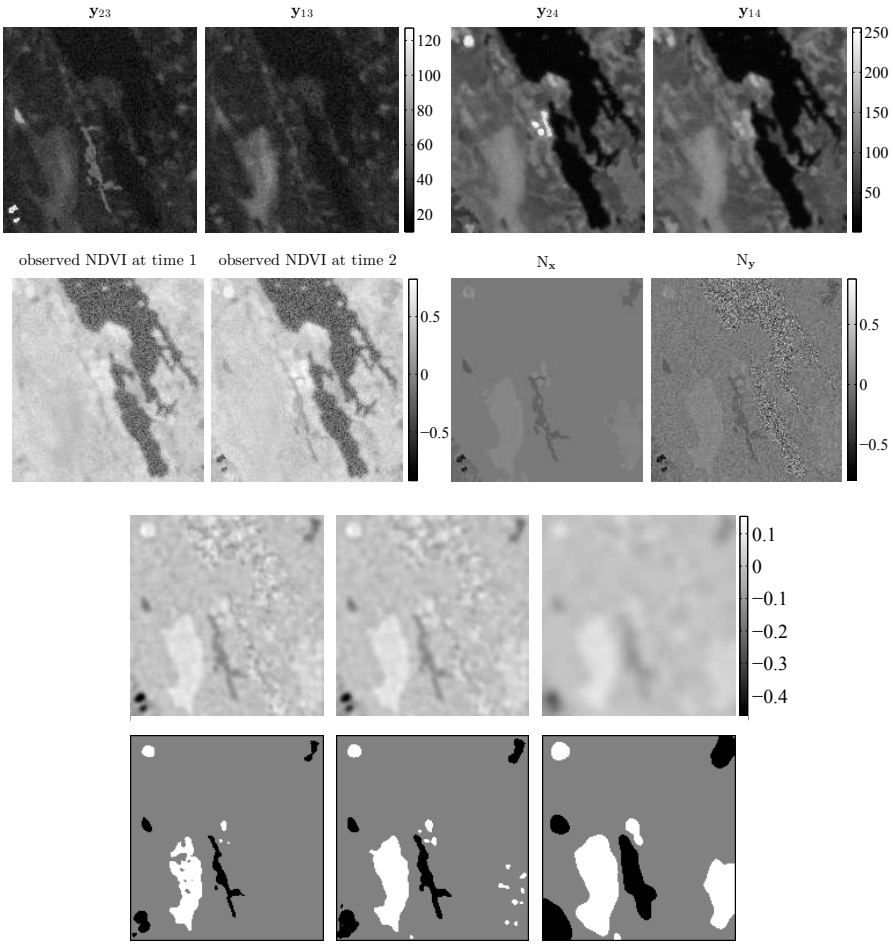


Figure 1: A partly artificially constructed pair of satellite images. 1st row: Noisy band images. 2nd row: Noisy NDVI-images, the true and the observed difference of NDVI-images. 3th row: Posterior means of $\mathbf{S}_\lambda\mathbf{N_x}$, $\lambda = 0, 1, 100$. 4th row: Corresponding iBSiZer-maps. Black, white and gray represents the pixels that are credibly negative, positive or neither.

References

[1] P. Chaudhuri and J. S. Marron. SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447):807–823, 1999.

[2] P. Chaudhuri and J. S. Marron. Scale space view of curve estimation. *The Annals of Statistics*, 28(2):408–428, 2000.

[3] P. Erästö and L. Holmström. Bayesian multiscale smoothing for making inferences about features in scatter plots. *Journal of Computational and Graphical Statistics*, 14(3):569–589, 2005.

[4] P. Erästö and L. Holmström. Bayesian analysis of features in a scatter plot with dependent observations and errors in predictors. *Journal of Statistical Computation and Simulation*, 77(5):421–431, 2007.

[5] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models. A roughness penalty approach*. Chapman & Hall, 1994.

[6] L. Holmström and L. Pasanen. Bayesian scale space analysis of differences in images. *Technometrics*, 54(1):16 – 29, 2012.

[7] L. Holmström, L. Pasanen, R. Furrer, and S. R. Sain. Scale space multiresolution analysis of random signals. *Computational Statistics & Data Analysis*, 55(10):2840 – 2855, 2011. Available on-line at http://dx.doi.org/10.1016/j.csda.2011.04.011.

[8] L. Pasanen and L. Holmström. Bayesian scale space analysis of temporal changes in Landsat ETM+ satellite images. Submitted for publication. Available on-line at http://cc.oulu.fi/~lpasanen/iBSiZer/ChangeDetection/ChangeDetection.pdf, 2012.

Towards more cost-effective identification of freshwater macroinvertebrates

Johanna Ärje[1], Salme Kärkkäinen[1], Tuomas Turpeinen[2] and Kristian Meissner[3]
[1]University of Jyväskylä, Department of Mathematics and Statistics
[2]University of Jyväskylä,
Department of Mathematical Information Technology
[3]Finnish Environment Institute (SYKE),
Monitoring and Assessment Unit, Jyväskylä, Finland

A growing number of anthropogenically caused drivers of ecosystem service change such as overexploitation, impacts of land use, pollution and alien species pose potential threats to aquatic ecosystems. Therefore we urgently need to assess, preserve and assure good future water quality and promote sustainable water resource management on a global scale. Targeted environmental legislation such as the EU Water Framework Directive require the inclusion of several biological indicator groups for freshwater ecological status assessment. Illustrating the complexity of this requirement is the fact that boreal freshwaters of Europe alone harbor hundreds of macroinvertebrate and thousands of microscopic periphyton and phytoplankton species.

The current funding mismatch between growing demands and the actual assets spend on actual biomonitoring call for new more cost effective ways to reach legislative goals. Here, we focus only on macroinvertebrate biomonitoring where a large portion of the total cost and time spend is due to manual identification of taxa by highly trained experts. However many of the processes and procedures used in automated taxa recognition of macroinvertebrates are extendable to other biological groups as well. We explore a novel approach of taxa identification based on a combination computer vision and a novel Bayes classifier, Random Bayes Array (RBA), to automatically identify macroinvertebrates from single posture images without human input. The RBA is an ensemble of Bayes classifiers, each using randomly selected geometric and color scale features extracted from the images. An ensemble approach is less prone to outliers and random feature selection reduces the impact of highly noisy features therefore eliminating the need for user based feature selection. In our preliminary tests on smaller data sets identification rates attained over 95%. Here, we test our novel classifier on a realistic set consisting of 6814 individuals belonging to 35 taxa. With the single posture data used here we achieved identification rates of 82% on average. Already these results are a first step towards resolving the mismatch between allocated resources to biomonitoring and legislative demands. However we are confident that our future work using multiple posture data and more refined features will significantly improve results further making cost effective automated biomonitoring an attainable goal.

Estimating aggregated nutriet fluxes in four Finnish rivers via Gaussian state space models

Jouni Helske*[1], Jukka Nyblom[1], Petri Ekholm[2], and Kristian Meissner[2]

[1]University of Jyväskylä, Department of Mathematics and Statistics

[2]Finnish Environment Institute

Abstract

Reliable estimates of the nutrient fluxes carried by rivers from land-based sources to the sea are needed for efficient abatement of marine eutrophication. The infrequent data calls for ways to reliably estimate the nutrient concentrations of the missing days. Here we use the Gaussian state space models with daily water flow as a predictor variable to predict missing nutrient concentrations and yearly fluxes for four agriculturally impacted Finnish rivers. Our results show no trends in yearly nutrient fluxes over the last 25 years for all of the four rivers examined.

*Corresponding author: Jouni Helske, jouni.helske@jyu.fi, University of Jyväskylä, Department of Mathematics and Statistics, P.O.Box 35 (MaD) Jyväskylä, FI 40014

1 Introduction

Abatement of marine eutrophication calls for reliable estimates of the nutrient fluxes carried by rivers from land-based sources to the sea. Monitoring programs of rivers typically involve daily measurements of water flow, but due to the costs, much more infrequent sampling of phosphorus and nitrogen concentrations. Our primary interest is a total nutrient flux over time span of length s (e.g. a calendar year) yearly nutrient flux, denoted by

$$m_{t,s} = \sum_{i=1}^{s} q_{t+i} c_{t+i},$$

where q_t is the water flow on the day t and c_t t is the daily nutrient concentration. Neglecting the measurement errors, if we had the values q_t and c_t measured on each day, then we would have correct nutrient fluxes.

2 Data

Our data consist of the concentrations of total phosphorus and total nitrogen, and water flow measurements from four Finnish rivers, Aurajoki, Paimionjoki, Porvoonjoki and Vantaanjoki, during 1985–2010. Daily measurements on nutrient concentrations are available for only 5–10% of the time period, while water flow measurements are usually available for each day.

3 Methods

For the prediction of the missing phosphorus and nitrogen concentration measurements we use a Gaussian state space model which consists of first order autoregressive component, and time varying regression components with daily water flow measurements as a predictor variables. It can be argued

(Wartiovaara, 1975; Rankinen et al., 2010) that the high water flow due to the precipitation has two opposite effects on the nutrient loadings. On other hand it increases the loading from the agriculture, but at the same time it decreases the relative loading from wastewaters.therefore we will use both log-flow and reciprocal of log-flow as predictor variables. The graphical representation of this model is shown in Figure 1. As the phosphorus and nitrogen concentration measurements are correlated, we model them together but separately for each river, allowing the observational disturbances e_t and the state disturbances η_t to correlate between nutrient-specific models.

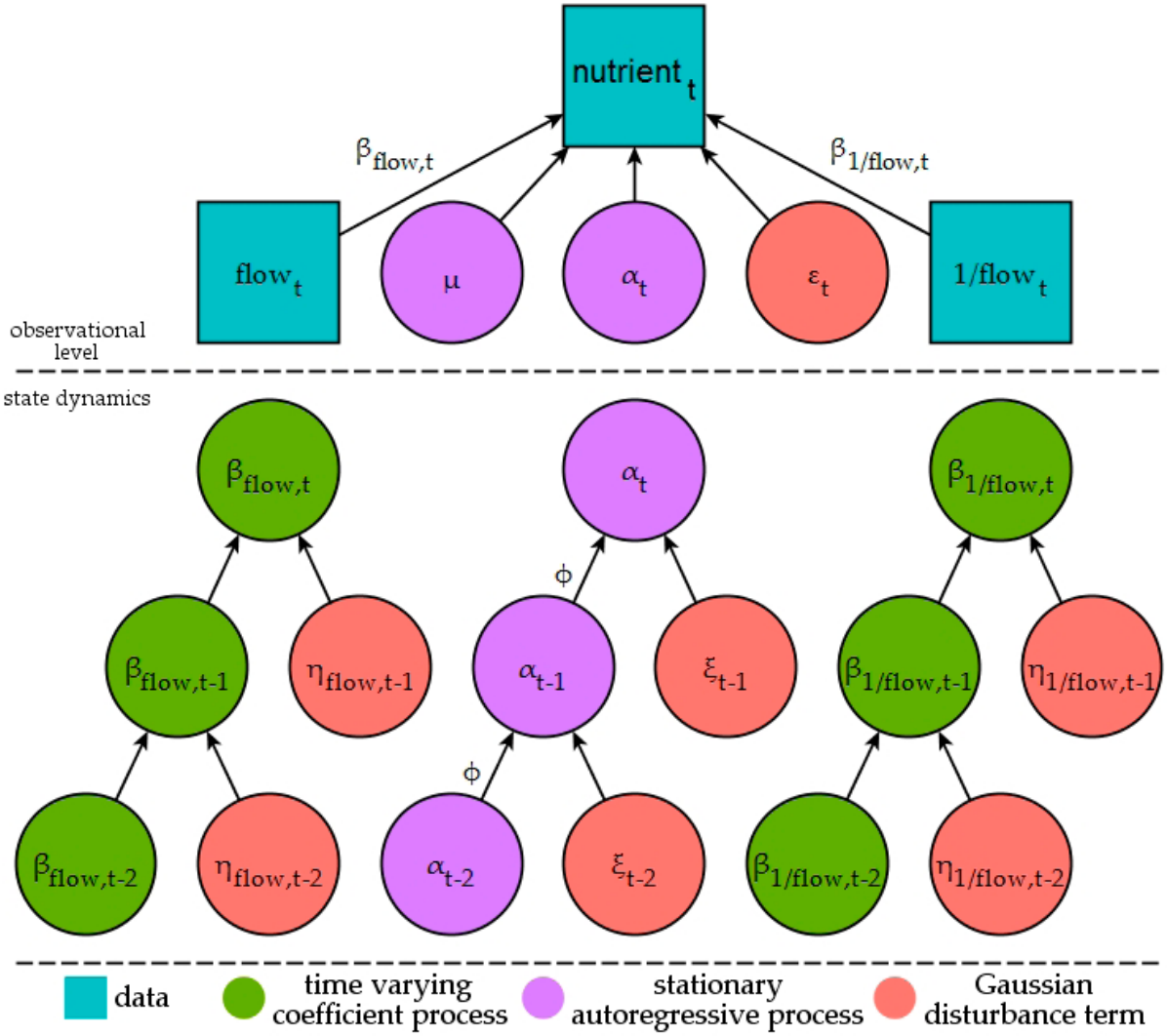


Figure 1: Graphical representation of Gaussian state space model for the single nutrient concentration series.

The unknown parameters of the model are estimated by maximum likelihood method. As a special case, when autoregressive coefficient $\phi = 0$ and $\mathrm{Var}(\eta_t) = 0$, an ordinary regression model is obtained. Given the estimated model, we estimate the yearly fluxes via conditional simulation of missing nutrient concentrations.

4 Results

The yearly estimates of nutrient fluxes with their simulated 95% prediction intervals are shown in the Figure 2. Each river exhibits a similar fluctuating

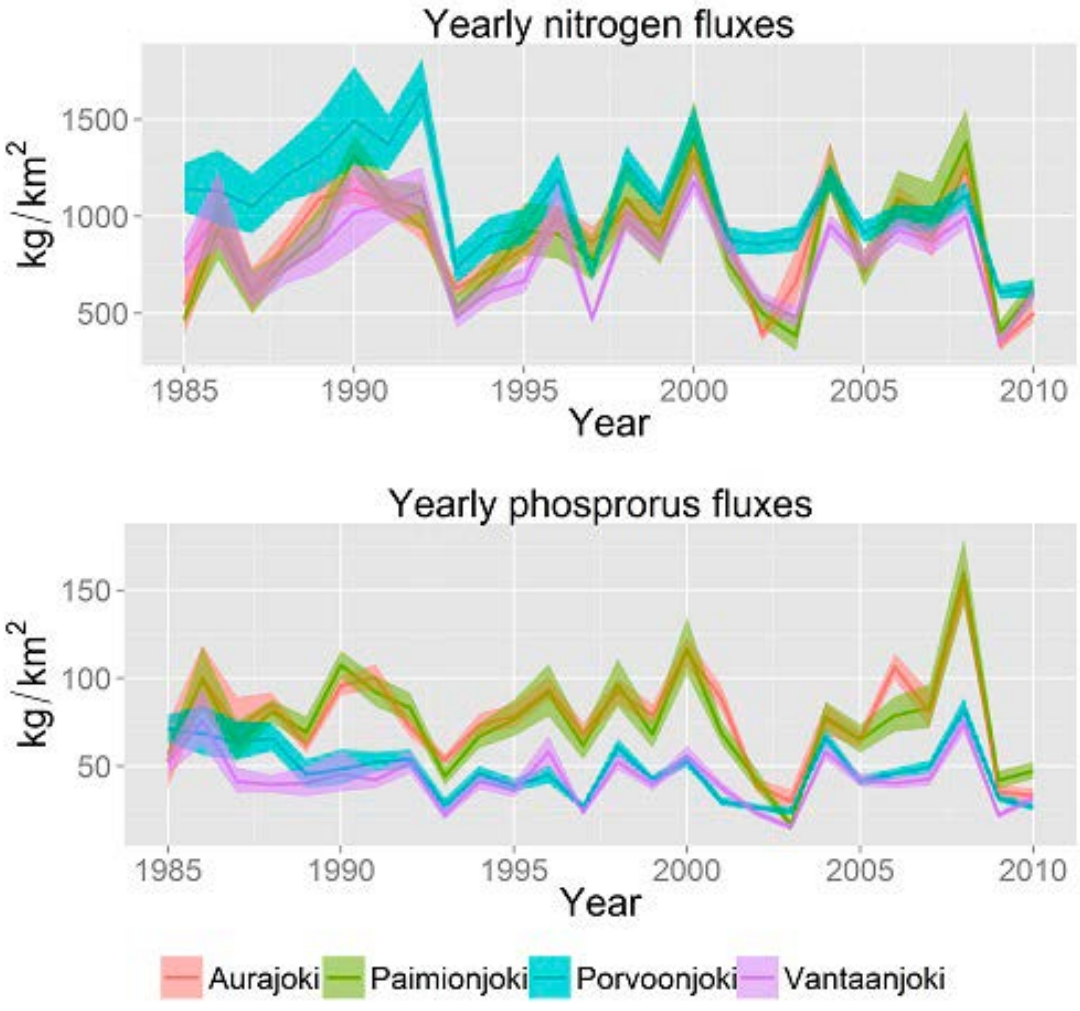


Figure 2: The point estimates and 95% prediction intervals for the yearly nutrient fluxes.

patterns without a clear trend. Especially yearly phosphorus fluxes, but also nitrogen fluxes clearly peak in 2008, followed by an even larger drop in 2009. The effect of varying sampling frequency is visible in the widths of prediction intervals.

Based on the simulation experiments (Helske et al., 2013), ordinary regression model gives somewhat biased flux estimates and produces the coefficients of variation that are often substantially smaller than our time varying model. As can be seen from Figure 3, the coefficients of variation from our model depend on the yearly sample sizes, whereas results from the ordinary regression model are overoptimistic and counterintuitive: uncertainty in the yearly flux estimate is independent from the amount of measurements in a given year.

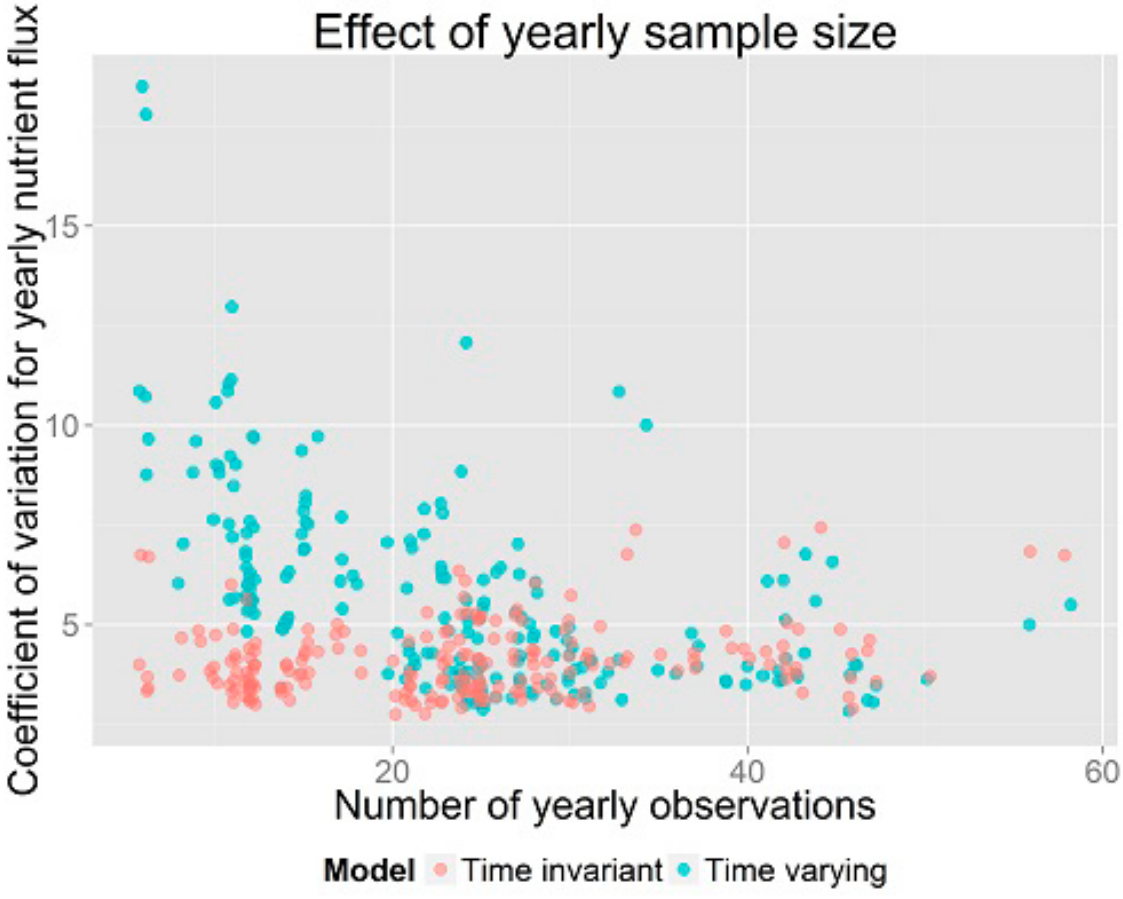


Figure 3: The relationship between coefficients of variation for yearly nutrient flux and the number of yearly nutrient concentration observations.

5 Conclusions

During the observational period 1985–2010, Finnish agricultural farmlands experienced a substantial decrease in phosphorus and nitrogen balance (OECD, 2012). Despite this drastic decrease in nutrient balance, we did not observe any corresponding trends in nutrient fluxes over the last 25 years for any of the four rivers examined. Our results seem to indicate that when daily flow data are available, relatively sparse data on nutrient concentrations can be used to estimate yearly fluxes, when the dynamic aspect of the phenomenon is taken into account.

References

J. Helske, J. Nyblom, P. Ekholm, and K. Meissner. Estimating aggregated nutrient fluxes in four Finnish rivers via Gaussian state space models. *Environmetrics*, 24(4):237–247, 2013.

OECD. Follow-up study of the impacts of agri-environmental measures in Finland. In *Evaluation of Agri-environmental Policies: Selected Methodological Issues and Case Studies*. OECD Publishing, 2012.

K. Rankinen, P. Ekholm, H. Sjöblom, and H. Rita. Nutrient losses from catchments and the governing factors. In J. Aakkula, T. Manninen, and M. Nurro, editors, *Follow-up study on the impacts of agri-environment measures (MYTVAS3) - Mid-term report*, Reports of Finland Ministry of Agriculture and Forestry 1, pages 122–131. 2010. In Finnish.

J. Wartiovaara. Amounts of substances discharged by rivers off the coast of Finland. *Publications of the Water Research Institute*, 13, 1975. In Finnish.

Application of mixed-eects models in forest sciences

Lauri Mehtätalo
University of Eastern Finland, School of Computing
May 20, 2014

1 Introduction

Forest datasets are often hierarchical. For example, they may consist of needles within branches, branches within trees, trees within sample plots or aerial images, sample plots within forest stands, forest stand within regions, repeated observations of trees in successive years or on different images. Also crossed grouping structures are common, for example, in datasets with tree increments for different calendar years, or forest stands on different aerial images. In these datasets, it is clear that the groups in the observed data represent a sample of a larger population of groups. Therefore, these datasets are naturally modeled using random effect models.

There is long tradition in using mixed-effects models in forest sciences, and foresters were among the first researchers in Finland to use them models (Lappi 1986). The use of random effects models over the fixed-effects models is justified, because (i) they provide more reliable inference on the model parameters than fixed-efects models would do. In addition, (ii) they enable predictions at different levels of the dataset and (iii) provide estimates of covariances between observations.

In an analysis of a forest dataset, the focus may either be in inference or in prediction (Harrell 2001). If the main interest is the inference (e.g. the effects of certain treatments on individuals) the first property is more important. If the main interest is prediction, then greatest benefit may arise from the possibility to make predictions at different levels of hierarchy. The prediction is possible also for groups from outside the modeling data either (i) by using the fixed part of the model or (ii) by augmenting the fixed part with predicted random effects based on some measurement data from the group. Even one observation per group is enough for such calibrated prediction.

In this paper, I will first present a simple linear mixed-effects model and some extensions of it. Thereafter, I will demonstrate and discuss the use of mixed-effects models in four different forestry situations listed below. The main benefit of mixed-effects models arsing either from prediction (P), inference (I) or estimated variance-covariance structure of the data (C).

1. Using a previously fitted linear mixed-effects model for tree height prediction (P)

2. Using a linear mixed-effect model with crossed grouping structure to predict a treatment-free response in a dataset of a thinning experiment (P).

3. Using nonlinear mixed-effect-models to analyse the previously extracted tree-level thinning effects (I)

4. Using a multivariate linear mixed-effects model system with crossed grouping structure to estimate the variance-covariance structure of repeated aerial observations of tree reflectance to aid in species classification (C).

2 Mixed-effects models

Let y_{ki} be the observed response for individual i in group k, and let x_{ki} be a fixed predictor. In a linear mixed-effects model, one may have both fixed (population level) parameters and random parameters, e.g.,

$$y_{ki} = a + bx_{ki} + \alpha_k + \epsilon_{ki} \,, \tag{1}$$

where we usually assume that $\alpha_k \sim N(0, \sigma_a^2)$ and $\epsilon_{ki} \sim N(0, \sigma^2)$; a and b are fixed parameters, which are estimated using GLS after first estimating σ_a^2 and σ^2 using (Restricted) Maximum Likelihood (RE)ML.

The model allows population level predictions as

$$\widetilde{y} = \widehat{a} + \widehat{b}x_{ki} \,,$$

and group-level predictions as

$$\widetilde{y}_k = \widehat{a} + \widetilde{\alpha}_k + \widehat{b}x_{ki} \,,$$

where $\widehat{a}$ and $\widehat{b}$ are the GLS estimates of the fixed parameters and $\widetilde{\alpha}_k$ is the predited random effect for group k, based on Empirical Best Linear Unbiased Prediction (EBLUP).

An extension of Equation 1 is the model with random slope

$$y_{ki} = a + bx_{ki} + \alpha_k + \beta_k x_{ki} + \epsilon_{ki} \tag{2}$$

where $(\alpha_k, \beta_k)' \sim N(0, \boldsymbol{D})$. Furthermore, for two nested groups, one may specify

$$y_{kti} = f(x_{kti}, \boldsymbol{b}) + \alpha_k + \alpha_{kt} + \epsilon_{kti}$$

with $\alpha_k \sim N(0, \sigma_1^2)$ and $\alpha_{kt} \sim N(0, \sigma_2^2)$. We assume where σ_1^2 and σ_2^2 are the variances of the random effects at the first and second level of grouping. For data with two crossed goups, the model with random constants at both levels of grouping is slightly different:

$$y_{kt} = f(x_{kt}; \boldsymbol{b}) + \alpha_k + \alpha_t + \epsilon_{kt} \,,$$

with $\alpha_k \sim N(0, \sigma_1^2)$ and $\alpha_t \sim N(0, \sigma_2^2)$. For nonlinear responses one may specify

$$y_{ki} = f(x_{ki}; \boldsymbol{B}_{ki}) + \epsilon_{ki} \,,$$

where

$$\boldsymbol{B}_{ki} = \boldsymbol{X}_{ki}\boldsymbol{b} + \boldsymbol{Z}_{ki}\boldsymbol{\beta}_k$$

specify the parameters of the nonlinear function as a linear function of fixed predictors and random effects; $\boldsymbol{\beta}_k \sim N(0, \boldsymbol{D})$.

The most simple varsion of a multivariate mixed-effects models is the bivariate mixed-effects model with random constants, which is specified by

$$\begin{aligned} y1_{ki} &= f1(x_{ki}; \boldsymbol{b}1) + \alpha 1_k + \epsilon 1_{ki} \\ y2_{ki} &= f2(x_{ki}; \boldsymbol{b}2) + \alpha 2_k + \epsilon 2_{ki} \end{aligned}$$

where $(\alpha 1_k, \alpha 2_k)' \sim N(0, \boldsymbol{D})$ and $(\epsilon 1_k, \epsilon 2_k)' \sim N(0, \boldsymbol{R})$.

Different combinations of the above-specified extensions are naturally possible, but fitting algorithms are not necessarily available for all situations in widely used software packages.

3 Applications

3.1 Prediction of tree height on diameter

The Height-Diameter (H-D) relationship of trees varies much among sample plots (Figure 1). However, field measurement of tree height is time-consuming. Therefore, diameter is usully recorded for all trees of a sample plot, whereas height is measured only for $0-5$ trees per plot in a forest inventory. A question arises, how to impute the heights for the rest of the trees, by using effectively the few sample trees of the plot to calibrate the model for the plot. Especially, if a previously fitted H-D model is available, it can be localized, or calibrated, for the new plot by predicting the random effects using the sampled tree heights.

In our case, the previously fitted H-D model is taken form Mehtätalo (2005), see also Lappi (1997) and Mehtätalo (2004). The logarithmic heigth H_{kti} for tree i in stand k at time t with diameter D_{kti} at the breast height is expressed by

$$\ln(H_{kti}) = a(DGM_{kt}) + \alpha_k + \alpha_{kt} + (b(DGM_{kt}) + \beta_k + \beta_{kt})D_{kti} + \epsilon_{kti} \quad (3)$$

where $a(DGM_{kt})$ and $b(DGM_{kt})$ are estimated fixed functions of plot-specific mean diameter DGM_{kt} (not shown), $(\alpha_k, \beta_k)'$ and $(\alpha_{kt}, \beta_{kt})'$ are the plot and measurement occasion -level random effects with variances and covariances (correlations in parentheses)

$$\text{var}\begin{bmatrix} \alpha_k \\ \beta_k \end{bmatrix} = \begin{bmatrix} 0.108^2 & (0.269) \\ 0.0028 & 0.0958^2 \end{bmatrix}$$

$$\text{var}\begin{bmatrix} \alpha_{kt} \\ \beta_{kt} \end{bmatrix} = \begin{bmatrix} 0.0168^2 & (-0.681) \\ -0.0003 & 0.0223^2 \end{bmatrix}$$

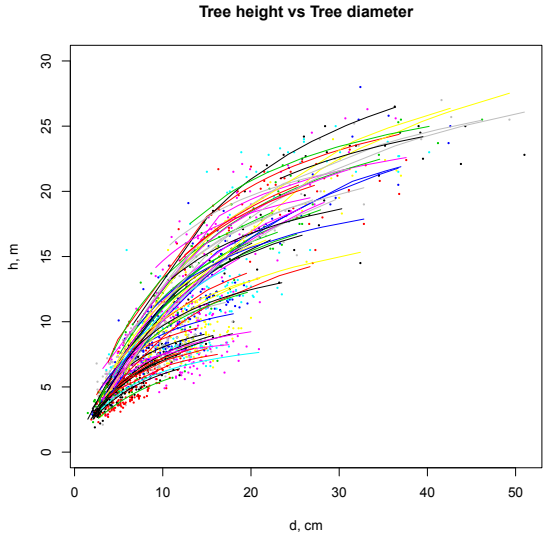


Figure 1: Tree height on tree diameter in a Scots Pine dataset in North Carelia. The lines represent the plot-specific H-D curves.

The residuals ϵ_{kti} are independent normal with $\mathrm{var}(\epsilon_{kti}) = 0.401^2\big(\max(D_{kti}, 7.5)\big)^{-1.068}$

The sample tree heights in a new stand can be described by

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Z\beta} + \boldsymbol{\epsilon}\,, \tag{4}$$

where $\boldsymbol{y}$ includes the observed sample tree heights, $\boldsymbol{\mu}$ is the predicted height based on the fixed part of model (3), $\boldsymbol{\beta} = (\ \alpha_k \quad \beta_k \quad \alpha_{k1} \quad \beta_{k1} \quad \alpha_{k2} \quad \beta_{k2} \quad \ldots)'$ includes the random effects at this particular plot with varaince $\mathrm{var}(\boldsymbol{\beta}) = \boldsymbol{D}$, $\boldsymbol{Z}$ is the corresponding design matrix, and $\boldsymbol{\epsilon}$ includes the residuals with varaince $\mathrm{var}(\boldsymbol{\epsilon}) = \boldsymbol{R}$.

The variances and covariances between random effects and observed heights can be written as

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{y} \end{bmatrix} \sim \left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \boldsymbol{D} & \boldsymbol{DZ'} \\ \boldsymbol{ZD} & \boldsymbol{ZDZ'} + \boldsymbol{R} \end{bmatrix} \right)$$

The Empirical Best Linear Unbiased Predictor (EBLUP) of random effects is

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{DZ'}(\boldsymbol{ZDZ'} + \boldsymbol{R})^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\,,$$

and the variance of prediction errors is

$$\mathrm{var}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \boldsymbol{D} - \boldsymbol{DZ'}(\boldsymbol{ZDZ'} + \boldsymbol{R})^{-1}\boldsymbol{ZD}$$

Consider a sample plot, where one tree was measured 5 years ago and 2 trees at the current year. The matrices and vectors in model (4) are

$$\boldsymbol{\mu} = \begin{bmatrix} 2.59 \\ 2.11 \\ 2.99 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 2.77 \\ 2.35 \\ 3.19 \end{bmatrix}$$

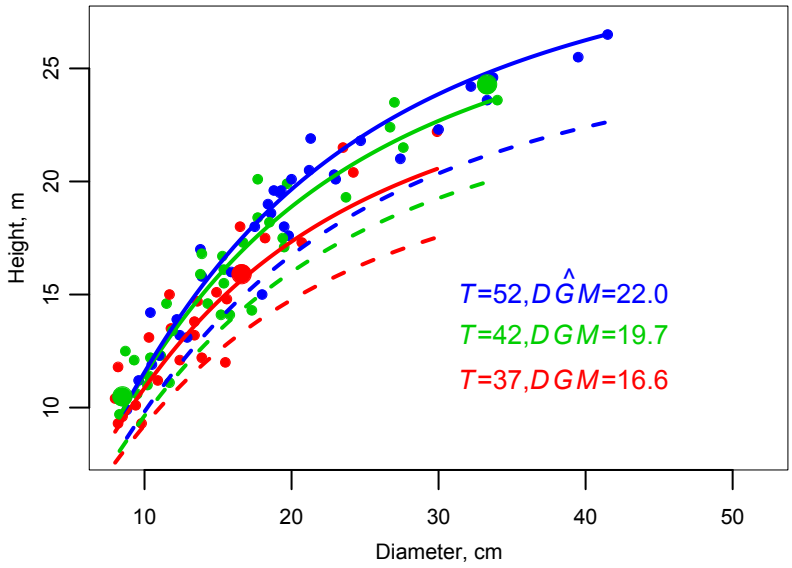


Figure 2: Tree heights on diameter in a sample plot at three different points in time, shown by different colors. The dashed lines show the predicted H-D relationship based on teh fixed part of model (3). The solid lines show the calibrated curve uisng predicted random effects based on the three observations shown using the large symbols.

$$\boldsymbol{Z} = \begin{bmatrix} 1 & -0.36 & 1 & -0.36 & 0 & 0 \\ 1 & -1.22 & 0 & 0 & 1 & -1.22 \\ 1 & 0.058 & 0 & 0 & 1 & 0.058 \end{bmatrix} \quad \boldsymbol{R} = \begin{bmatrix} 0.008 & 0 & 0 \\ 0 & 0.016 & 0 \\ 0 & 0 & 0.004 \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \alpha_k \\ \beta_k \\ \alpha_{k1} \\ \beta_{k1} \\ \alpha_{k2} \\ \beta_{k2} \end{bmatrix} \quad \boldsymbol{D} = \begin{bmatrix} 0.0118 & 0.0028 & 0 & 0 & 0 & 0 \\ 0.0028 & 0.0092 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0003 & 0.0004 & 0 & 0 \\ 0 & 0 & 0.0004 & 0.0005 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0003 & 0.0004 \\ 0 & 0 & 0 & 0 & 0.0004 & 0.0005 \end{bmatrix}$$

Predicting the random effects using EBLUP and using them to predict the plot-specific curves yields plot-specific H-D curves, which are shown in Figure 2 for three different points in time by the solid lines.

3.2 Extracting effects of silvicultural thinnings

Forest managers use silvicultural thinnings to decrease the competition of neighboring trees and, consequently, to increase the growth rate of the remaining trees for faster production of sawtimber. To understand the dynamics of thinning, one may wish to analyse the effect of thinnings on tree growth. However, the growth is affected also by other factors, especially by the site productivity, tree age, and annual weather. Mehtätalo et al. (2014) used mixed-effects models to remove these nuisance effects.

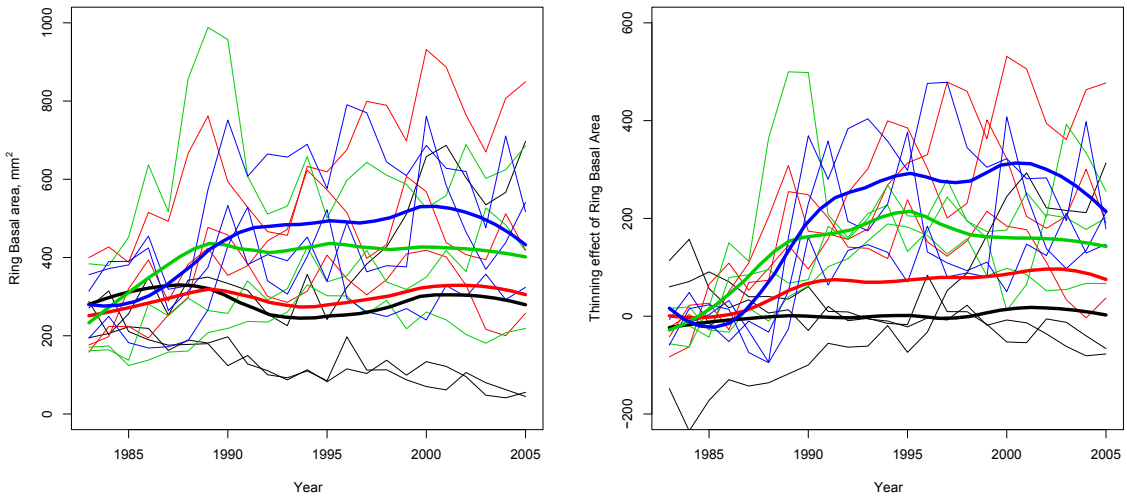


Figure 3: The observed ring basal area on calendar year on selected trees of the thinning experiment data (left) and the extracted thinning effects (right). The thick lines show the treatment-specific trend curves as follows: black: Control; red:Light; green: Moderate; blue: Heavy.

The data were collected on a thinning experiment, established in naturally generated Scots pine stands at the age of $\sim$ 25 years in Mekrijärvi, Finland in 1986. One of the four following thinning treatments were applied to each plot: No thinning (I, Control), light (II), moderate (III), and heavy (IV) thinnings. A total of 88 trees were felled in 2006, and the complete time series of diameter increments between 1983 and 2006 was measured for each tree using an X-ray densiometer. The diameter growths were transformed to basal area growths, because the interest lies in the volume growth, which may be better corrlated with the volume growth due to the dimensionality of the tree (approximately $Volume = aDiameter^2Height$). The raw data are shown in the left plot of Figure 3. The figure shows a possible weak age trend, as well as year and tree specific effects.

To model the growth of trees without the thinnings, a dataset without thinning treatments was produced by including from the original data the control treatment for whole follow-up period and the thinned treatments until the year of thinning (1986). Thereafter, a linear mixed effect model with random year and tree effects was fitted to the unthinned data

$$y_{kt} = f(T_{kt}; \boldsymbol{b}) + \alpha_k + \alpha_t + \epsilon_{kt} \tag{5}$$

where y_{kt} is the basal area growth of tree k at year t, $f(T_{kt}; \boldsymbol{b})$ is the age trend (modeled using a spline), α_k is a tree effect, α_t is a year effect and ϵ_{kt} is a residual, which are all normally distributed, and i.i.d.

Using the estimated age trend and BLUP's of year and tree effects, the growth without thinning, $\widetilde{y}_{kt}$ was predicted for treatments II -IV after the thin-

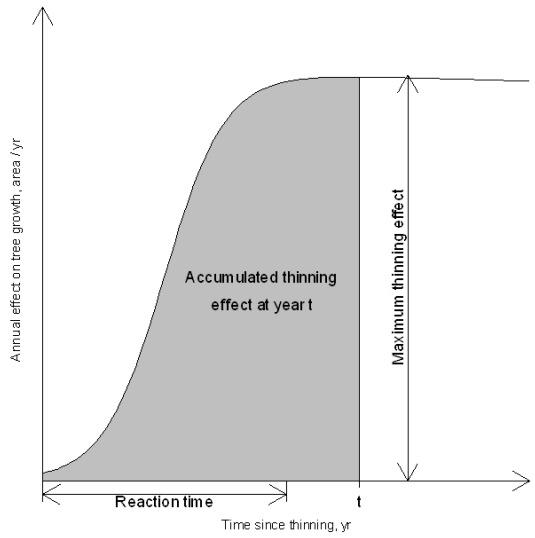


Figure 4: Illustration of the logistic curve which was used for modeling the dynamics of the thinning effects.

ning year. The pure thinning effects were estimated by subtracting the prediction from the observed growth

$$d_{kt} = y_{kt} - \widetilde{y}_{kt} \tag{6}$$

The right plot of Figure 3 shows the resulting extracted thinning effect.

3.3 Modeling thinning effects using nonlinear mixed-effects models

The thinning effects seem to switch on during a short time called **Reaction time** and stabilize thereafter at a level of **Maximum thinning effect**. To explore what predictors control these two parameters, the thinning effects of the thinned treatments 2-4 were modeled using a nonlinear mixed-effects model. The random effects were used to take into account the lack of independence resulting from the data hierarchy.

The thinning effect of tree k at time t was modeled using a logistic curve

$$d_{kt} = \frac{M_k}{1 + \exp\left(4 - 8\frac{x_{kt}}{R_k}\right)} + e_{kt}\,,$$

where d_{kt} is the thinning effect, x_{kt} is time since thinning, $M_k = \mu_0 + \mu_1 T_2 + \mu_2 T_3 + \mu_4 x_{kt} + m_k$ is the maximum thinning effect, $T_2, \ldots, T_3$ are the treatments, $R_k = \rho_0 + \rho_1 z_k + r_k$ is the reaction time and z_k is the standardized diameter (standardixzed using the plot-specific mean and standard deviation). For the random effects, we assume $\begin{bmatrix} m_k \\ r_k \end{bmatrix} \sim N(\mathbf{0}, \boldsymbol{D}_{2x2})$. The residual e_{kt} are normal,

Table 1: Parameter estimates for the thinning effect model.

Fixed parameters	Estimate	s.e.	p-value
μ_0	112.8	23.29	0.0000
μ_1	91.91	30.45	0.0026
μ_2	169.2	32.14	0.0000
μ_3	-3.214	1.006	0.0014
ρ_0	5.749	0.4458	0.0000
ρ_1	-1.461	0.4568	0.0014
Random parameters			
var(r_k)	93.012		
var(m_k)	2.0852		
cor(r_k,m_k)	0.203		
Residual			
σ^2	8.157*10$^-$4		
δ_1	8.746*10^4		
δ_2	1.886		
δ_3	0.5888		

with incocstant variance (modelled using a power function) and AR(1) structure within a tree (Mehtätalo et al. 2014).

The reaction time was 6 years (Table 1). It did not significantly vary among treatments but was shorter for large trees. The maximum thinning effect increased with thinning intensity, being 282 mm/yr for treatment IV, which indicates a 87% increase in the basal area growth compared to the control. The main conclusion from the analysis was that was that the reaction time is affected by the relative size of the tree in the stand, whereas the maximum thinning effect is affected by thinning intensity.

3.4 Modelling tree-level reflectance on aerial images

The reflectance (color) of a tree on an aerial image can be used to classify tree species. However, the viewing direction with respect to sunlight affects the spectral characteristics of a tree. This effect is species-specific. Therefore, observing a certain tree from multiple directions (=images) may provide more accurate species classification than an observation on one aerial image only (Korpela et al. 201X).

The study is based on 20 partially overlapping aerial images of a forest area. The raw data was postprocessed to provide (atmospherically corrected) reflectance data on four channels: RED, GRN, BLU and NIR. $N = 15188$ dominant trees discernible in 2-7 images formed the reference tree data (5914 Scots pines, 7105 Norway spruces, 2169 birches) Individual trees on different images were using automatically matched. The individual pixels within tree

owns were divided to sunlit and self-shaded pixels. The mean reflectances these parts were analyzed separately, resulting to a system of 8 models (4 annels, shaded and sunlit) for each of the three tree species.

In our dataset, observations from a given image are similar due to e.g. the operties of the atmosphere at the time of imaging and the atmospheric correcon. On teh other hand, repeated measurements of a certain tree are correlated ue to tree-specific properties. Therefore, the model for each response and tree ecies has the following structure

$$y_{it} = f(\boldsymbol{x}_{it}|\boldsymbol{b}) + \alpha_i + \alpha_t + \epsilon_{it}\,,$$

here i and t refer to image and tree effects, respectively. σ_1^2 and σ_2^2 are the rresponding variances. The predictors are trigonometric transformations of ue horizontal and vertical viewing and Sun angles.

The random effects at different levels of grouping are independent, therefore

$$\begin{aligned}
\mathrm{var}(y_{it}) &= \sigma_1^2 + \sigma_2^2 + \sigma^2 \\
\mathrm{cov}(y_{it}, y_{i't'}) &= 0 \\
\mathrm{cov}(y_{it}, y_{it'}) &= \sigma_1^2 \\
\mathrm{cov}(y_{it}, y_{i't}) &= \sigma_2^2
\end{aligned}$$

The multivariate model for a tree species is

$$\begin{aligned}
y1_{it} &= f1(\boldsymbol{x}_{it}|\boldsymbol{b}1) + \alpha1_i + \alpha1_t + \epsilon1_{it} \\
y2_{it} &= f2(\boldsymbol{x}_{it}|\boldsymbol{b}2) + \alpha2_i + \alpha2_t + \epsilon2_{it} \\
&\vdots \\
y8_{it} &= f8(\boldsymbol{x}_{it}|\boldsymbol{b}8) + \alpha8_i + \alpha8_t + \epsilon8_{it}
\end{aligned}$$

simply

$$\boldsymbol{y}_{it} = \boldsymbol{f}(\boldsymbol{x}_{it}|\boldsymbol{b}) + \boldsymbol{\alpha}_i + \boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_{it}$$

here the responses 1-8 refer to the sunlit and self-shaded pixels of the four annels and $(\alpha1_i, \alpha2_i, \ldots, \alpha8_i)' = \boldsymbol{\alpha}_i \sim N(0, \boldsymbol{A}_{8\times8})$ include the random image-fects, $(\alpha1_t, \alpha2_t, \ldots, \alpha8_t)' = \boldsymbol{\alpha}_t \sim N(0, \boldsymbol{B}_{8\times8})$ include the random tree-effects, nd $(\epsilon1_{it}, \epsilon2_{it}, \ldots, \epsilon8_{it})' = \boldsymbol{\epsilon}_{it} \sim N(0, \boldsymbol{E}_{8\times8})$ include the random residuals. Furnermore,

$$\begin{aligned}
var(\boldsymbol{y}_{it}) &= \boldsymbol{A} + \boldsymbol{B} + \boldsymbol{E} \\
cov(\boldsymbol{y}_{it}, \boldsymbol{y}_{i't'}) &= 0 \\
cov(\boldsymbol{y}_{it}, \boldsymbol{y}_{it'}) &= \boldsymbol{A} \\
cov(\boldsymbol{y}_{it}, \boldsymbol{y}_{i't}) &= \boldsymbol{B}
\end{aligned}$$

lodel fitting (based on REML) yielded the estimates $\widehat{\boldsymbol{b}}$, $\widehat{\boldsymbol{A}}$, $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{E}}$.

Variance components, real data, 200 000 observations (%)

	sunlit	shade	sunlit	shade	sunlit	shade	sunlit	shade
Fixed (Xβ)-%	33	11	32	13	45	29	7	-0
Tree-%	42	42	43	41	18	13	62	64
Image-%	4	12	5	14	27	46	6	2
Residual-%	21	35	20	32	10	13	25	34
Total	100	100	100	100	100	100	100	100

* Fixed part: The anisotropy trends explained SL >> SS, BLU > GRN > RED > NIR. In NIR, anisotropy is low.

* Tree-effect: The correlations are strong, both for SL and SS. A bright tree is bright across views and bands. In NIR > 60% of variance explained!!

* Image-effect: Substantial in BLU, SS > SL. Includes effects from solar elevation changes (07-09 GMT), atmospheric correction errors.

Ilkka Korpela, Oct 2012

Figure 5: The distribution of the total varaince in the reflectance data into parts explained by the fixed part of the model, tree effects, image effects, and residual varaince.

The estimated variance-covariance matrices can be used in tree species classification. Let $\boldsymbol{y}_{it}$ be an observed vector (length=8) of the reflectances of one tree t on the 8 channels on one image i. The squared Mahalanobis distance between $\boldsymbol{y}_{it}$ and $\boldsymbol{\mu}_{it}$ is

$$d_{it}^2 = (\boldsymbol{y}_{it} - \boldsymbol{\mu}_{it})'(\boldsymbol{A} + \boldsymbol{B} + \boldsymbol{E})^{-1}(\boldsymbol{y}_{it} - \boldsymbol{\mu}_{it})$$

This distance takes into account the correlation of reflectance among different channels, and is (at least under multivariate normality of the reflectance data) in a way optimal for single tree on single image. For multiple images, the squared Mahalanobis distance between $\boldsymbol{y}_{\cdot t}$ and $\boldsymbol{\mu}_{\cdot t}$ is

$$d_{\cdot t}^2 = (\boldsymbol{y}_{\cdot t} - \boldsymbol{\mu}_{\cdot t})'\boldsymbol{D}_{\cdot t}^{-1}(\boldsymbol{y}_{\cdot t} - \boldsymbol{\mu}_{\cdot t})\,,$$

where $\boldsymbol{y}_{\cdot t} = (\boldsymbol{y}'_{1t}, \ldots, \boldsymbol{y}_{mt})$ is an observed vector (with length of $8m$) of the reflectances of tree t on the 8 channels of m images. The $8m \times 8m$ variance-covariance matrix is

$$\boldsymbol{D}_{\cdot t} = \begin{bmatrix} \boldsymbol{A}+\boldsymbol{B}+\boldsymbol{E} & \boldsymbol{B} & \cdots & \boldsymbol{B} \\ \boldsymbol{B} & \boldsymbol{A}+\boldsymbol{B}+\boldsymbol{E} & & \boldsymbol{B} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{B} & \boldsymbol{B} & \cdots & \boldsymbol{A}+\boldsymbol{B}+\boldsymbol{E} \end{bmatrix}$$

This distance takes into account the correlation arising from the common tree effects. Further extension to many trees and images would be possible as well.

4 Conclusion

Mixed-effects models are useful tools for analyzing grouped datasets in different contexts. The benefit from the use of mixed-effects models depends on the application, but may be related to

- inference (Case 3: Modelling the thinnig effect),
- prediction (Case 1: H-D, Case 2: Extraction of thinning effect), or
- estimated variance-covariance structure of the data (Case 4: Species classification).

The prediction of random effects for a new group is a powerful tool to localize models afterwards using very limited datasets. It might be useful also on other fields than forest sciences.

References

Harrell, F.J. 2001. Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. Springer, New York, USA.

Korpela, I., L. Mehtätalo, A. Seppänen, and L. Markelin. 201X. Tree species classification using directional reflectance anisotropy signatures in multiple aerial images. Journal XX(X):XXX–XXX.

Lappi, J. 1986. Mixed linear models for analyzing and predicting stem form variation of scots pine. Communicationes Instituti Forestalis Fenniae 134, FFRI. 69 p.

Lappi, J. 1997. A longitudinal analysis of height/diameter curves. Forest Science 43:555–570.

Mehtätalo, L. 2004. A longitudinal height-diameter model for norway spruce in finland. Canadian Journal of Forest Research 34(1):131–140.

Mehtätalo, L. 2005. Height-diameter models for scots pine and birch in finland. Silva Fennica 39(1):55–66.

Mehtätalo, L., H. Peltola, A. Kilpeläinen, and V.P. Ikonen. 2014. The response of basal area growth of scots pine to thinning: A longitudinal analysis of tree-specific series using a nonlinear mixed-effects model. Forest Science xx(x):xx – xx.

Leo Törnqvist -palkinto:
Paras tilastotieteen
pro gradu -tutkielma
(2011–2012)

Kahden tason rakenneyhtälömallinnus ordinaalisille muuttujille: Esimerkkinä köyhyysindikaattorit Laosissa

Tytti Pasanen
TAMPEREEN YLIOPISTO
Informaatiotieteiden yksikkö
Tilastotiede

Tiivistelmä

Rakenneyhtälömallit kattavat laajan määrän erilasia tilastollisia malleja, joista klassisin tapaus lienee konfirmatorisen faktorianalyysin ja polkumallinnuksen yhdistelmä. Parametrien estimointiin on olemassa lukuisia menetelmiä, joista painotetun pienimmän neliösumman (WLS) menetelmä on kehitetty järjestysasteikollisten tai dikotomisten muuttujien mallinnukseen. Täten WLS sopii kyselytutkimusten analyysiin silloin, kun kysymykset pohjaavat Likert-asteikollisiin mittareihin. Monitasomallit taas mahdollistavat klusteroidun aineiston eri tasojen samanaikaisen mallinnuksen. Havaitut muuttujat jaetaan tällöin kahteen osaan. Klusterikeskiarvot mallinnetaan ryhmätasolla ja yksilöiden erotukset klusterikeskiarvoistaan mallinnetaan yksilötasolla.

Pro gradu -työssä kahden tason rakenneyhtälömallinnusta sovellettiin Laosin kotitalouskyselyyn. Tuloksista voi päätellä, että Laosin kotitalouksien köyhyys on sidoksissa kulutukseen sekä kylän yleiseen köyhyystasoon ja infrastruktuuriin. Monitasoisia rakenneyhtälömalleja järjestysasteikollisille muuttujille voidaan soveltaa vastaaviin tutkimusongelmiin, joiden tutkiminen perinteisin suurimman uskottavuuden menetelmin ei ole mielekästä.

1 Johdanto

Pro gradu -työ tarkasteli ordinaalisten havaittujen muuttujien monitasoista rakenneyhtälömallinnusta. Menetelmä on suhteellisen uusi ja vähän sovellettu erityisesti kehitystutkimuksen puolella, johon tutkielman aineisto ja tutki-

muskysymykset sijoittuvat. Tämä selostus pro gradu -työstä keskittyy kahden tason rakenneyhtälömallien matemaattisen perustaan, ja soveltavan esimerkin tulokset esitellään vain lyhyesti. Sovelletun aineiston ryväsotanta oli kuitenkin merkittävä tekijä menetelmän valinnassa, sillä klusteroitujen otosten tapauksessa moniin tilastollisiin menetelmiin olennaisesti kuuluva oletus havaintojen riippumattomuudesta ei päde.

Rakenneyhtälömallit kattavat laajan valikoiman erilaisia tilastollisia malleja. Näille tyypillistä on latenttien muuttujien käyttö, monimutkaiset kausaalisuhteet sekä useampi selitettävä muuttuja. Rakenneyhtälömalleissa käytetyimmän ja tunnetuimman estimointimenetelmän, suurimman uskottavuuden (*maximum likelihood* - ML), käyttö on mielekästä kuitenkin vain normaalijakautuneiden muuttujien tapauksessa. Kyselytutkimuksille tyypillisten Likert-asteikollisten tai kaksiluokkaisten muuttujien mallinnus ei täten onnistu suurimman uskottavuuden menetelmällä. Tämän lisäksi kyselyt voidaan käytännössä toteutettaa ryväsotannalla ajan tai muiden resurssien säästämiseksi, jolloin yleinen oletus havaintojen riippumattomuudesta ei päde. Monitasomallinnuksella havaintojen klusteroituminen voidaan ottaa estimoinnissa huomioon ja mallintaa sekä klusteritason että yksilötason ominaisuuksia samanaikaisesti. Tyypillisimmät monitasomallit ovat olleet regressiomallien laajennuksia (Raudenbush & Bryk 2002), joihin ei ole voinut sisällyttää latentteja muuttujia tai muita rakenneyhtälömalleille tyypillisiä ominaisuuksia.

Tässä katsauksessa esiteltävä menetelmä, kahden tason rakenneyhtälömallinnus ei-normaalijakautuneille muuttujille (Asparouhov & Muthén 2007), tarjoaa ratkaisun yllämainittuihin ongelmiin, joita kyselytutkimusten analysoinnissa voi kohdata. Menetelmä on ollut sovellettavissa tilastollisissa sovellusohjelmistoissa vasta suhteellisen lyhyen aikaa, ja ainakin gradun tekohetkellä vuonna 2012 tämä oli saatavilla ainoastaan Mplus-ohjelmistossa.

2 Rakenneyhtälömallinnus

Rakenneyhtälömallit (*structural equation models* - SEM) analysoivat malliin sisällytettyjen muuttujien kovarianssirakennetta. Parhaassa tapauksessa otoskovarianssin sekä konstruoidun mallin pohjalta ennustetun mallin kovarianssirakenteet ovat lähellä toisiaan. Kovarianssimatriisin oletetaan koostuvan parametreihin liittyvistä funktioista. Yksinkertaisimmillaan testattava hypoteesi on muotoa

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}), \tag{2.1}$$

jossa $\boldsymbol{\Sigma}$ on koko populaation havaittu kovarianssimatriisi, $\boldsymbol{\theta}$ on mallin parametrit sisältävä vektori ja $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ on mallin kovarianssimatriisi, joka on ilmaistu vektorin $\boldsymbol{\theta}$ funktiona. (Bollen 1989). Käytännössä populaation kovarianssimat-

riisi on tuntematon, ja hypoteesin testaus pohjaa otoskovarianssimatriisiin $\mathbf{S}$ ja estimoituun kovarianssimatriisiin $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$.

Nimensä mukaisesti rakenneyhtälömalli koostuu rakenneparametreja sekä satunnaismuuttujia sisältävistä yhtälöistä. Populaation rakenneparametrit estimoidaan havaittujen muuttujien tunnettujen ominaisuuksien eli varianssien ja kovarianssien avulla. Kaikkia mahdollisia parametreja ei välttämättä tarvitse estimoida, sillä tutkija voi kiinnittää näitä mielekkääksi katsomallaan tavalla. Perinteinen jako riippumattomiin $\mathbf{x}$ ja riippuviin $\mathbf{y}$ muuttujiin ei päde rakenneyhtälömalleihin, sillä näissä muuttuja voi edustaa molempia. Täten muuttujat jaetatan *eksogeenisiin* and *endogeenisiin*. Yleensä eksogeeniset muuttujat ovat kausaalisesti riippumattomia muista mallin muuttujista, kun taas endogeenisien muuttujien oletetaan riippuvan ainakin jossain määrin muista muuttujista.

Tässä luvussa esittelen yleisen rakenneyhtälömallin kaavan sekä perusteet estimointimenetelmästä järjestysasteikollisille muuttujille.

2.1 Mallin määritys

Rakenneyhtälömalli koostuu kahdesta osasta. Latenttien muuttujien *rakennemalli* määrittelee latenttien muuttujien keskinäiset suhteet, ja *mittamalli* puolestaan määrittelee latenttien muuttujien rakenteet eli niihin kuuluvat indikaattorit. Soveltavan tutkijan näkökulmasta rakennemalli on usein kiinnostavin. Kuitenkin myös mittamallin ominaisuudet ovat olennaisia mallin estimoinnin kannalta. Tässä luvussa esitetyt kaavat ja ajatukset perustuvat pitkälti Bollenin perinpohjaiseen teokseen vuodelta 1989.

Rakennemalli ilmaistaan usein muodossa

$$\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}, \tag{2.2}$$

jossa vektori $\boldsymbol{\eta}$ kuvaa endogeenisiä ja vektori $\boldsymbol{\xi}$ eksogeenisiä latentteja satunnaismuuttujia, joiden kerroinmatriisit ovat vastaavasti $\mathbf{B}$ ja $\boldsymbol{\Gamma}$. Toisin sanoen $\mathbf{B}$ ja $\boldsymbol{\Gamma}$ sisältävät kerrottaviensa yhteyksien voimakkuudet latentteihin endogeenisiin muuttujiin. Esimerkiksi kerroin β_{ij} edustaa muuttujan η_{j} yhden yksikön suuruisen muutoksen aiheuttamaa muutosta muuttujassa η_{i} olettaen, että mallin muut tekijät pysyvät vakiona. Vastaavasti kerroin γ_{ji} kuvaa muuttujan ξ_{i} yhden yksikön muutoksen vaikutusta muuttujaan η_{j}. Malleissa oletetaan, että muuttuja ei voi olla itsensä suoraa seurausta, joten matriisin $\mathbf{B}$ diagonaali kiinnitetään aina nollaksi. Tämän kiinnityksen myötä matriisi ($\mathbf{I}$-$\mathbf{B}$) on epäsingulaarinen, joten sen voi kääntää.

Muiden kaavan 2.2 oletusten mukaan vektorien $\boldsymbol{\eta}$ ja $\boldsymbol{\xi}$ sekä residuaalimatriisin $\boldsymbol{\zeta}$ odotusarvot ovat 0, ja $\boldsymbol{\zeta}$ ei voi korreloida vektorin $\boldsymbol{\xi}$ eksogeenisten muuttujien kanssa. Jatkossa vektorien $\boldsymbol{\xi}$ ja $\boldsymbol{\zeta}$ kovarianssimatriiseja merkitään symbolein $\boldsymbol{\Phi}$ ja $\boldsymbol{\Psi}$.

Mittamalli yhdistää latentit satunaismuuttujat niiden havaittuihin indi-

kaattorimuuttujiin. Näiden rakenne ilmaistaan yleensä muodossa

$$\boldsymbol{x} = \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta} \tag{2.3}$$

$$\boldsymbol{y} = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{2.4}$$

joissa $\boldsymbol{x}$ ja $\boldsymbol{y}$ sisältävät latenttien muuttujien vektoreiden $\boldsymbol{\xi}$ ja $\boldsymbol{\eta}$ indikaattorit, ja $\boldsymbol{\delta}$ sekä $\boldsymbol{\epsilon}$ kuvaavat näiden mittavirhettä. Molemmat $\boldsymbol{\Lambda}$-matriisit sisältävät latenttien vektoreiden indikaattorien lataukset, jotka kuvaavat sitä muutosta, jonka yhden yksikön muutos latentissa muuttujassa odotetaan aiheuttavan sen havaitussa indikaattorissa olettaen, että kaikki muu mallissa pysyy vakiona.

Mittamallin oletukset muistuttavat latentin mallin oletuksia. Sekä vektorien $\boldsymbol{\eta}$ ja $\boldsymbol{\xi}$ että residuaalivektorien $\boldsymbol{\delta}$ ja $\boldsymbol{\epsilon}$ odotusarvot ovat 0. Residuaalivektorit eivät voi korreloida keskenään eivätkä latenttien muuttujien kanssa. Vektoreiden $\boldsymbol{\delta}$ ja $\boldsymbol{\epsilon}$ kovarianssimatriiseja merkitään symbolein $\boldsymbol{\Theta}_\epsilon$ ja $\boldsymbol{\Theta}_\delta$.

Mikäli $\boldsymbol{\eta}$ ratkaistaan yhtälöstä 2.2, saadaan yhtälö muotoon $\boldsymbol{\eta} = (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta})$. Yksinkertaisuuden vuoksi merkitään jatkossa matriisia $(\mathbf{I}-\mathbf{B})^{-1}$ symbolilla $\mathbf{A}$. Näiden merkintöjen avulla kovarianssimatriisi voidaan ilmaista seuraavien neljän elementin avulla:

$$\begin{aligned}
\boldsymbol{\Sigma}(\boldsymbol{\theta}) &= \begin{bmatrix} \boldsymbol{\Sigma}_{yy}(\boldsymbol{\theta}) & \boldsymbol{\Sigma}_{yx}(\boldsymbol{\theta}) \\ \boldsymbol{\Sigma}_{xy}(\boldsymbol{\theta}) & \boldsymbol{\Sigma}_{xx}(\boldsymbol{\theta}) \end{bmatrix} \\
&= \begin{bmatrix} E(\boldsymbol{y}\boldsymbol{y}') & E(\boldsymbol{y}\boldsymbol{x}') \\ E(\boldsymbol{x}\boldsymbol{y}') & E(\boldsymbol{x}\boldsymbol{x}') \end{bmatrix} \\
&= \begin{bmatrix} E\left[(\boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\epsilon})(\boldsymbol{\eta}'\boldsymbol{\Lambda}_y' + \boldsymbol{\epsilon}')\right] & E\left[(\boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\epsilon})(\boldsymbol{\xi}'\boldsymbol{\Lambda}_x' + \boldsymbol{\delta}')\right] \\ E\left[(\boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta})(\boldsymbol{\eta}'\boldsymbol{\Lambda}_y' + \boldsymbol{\epsilon}')\right] & E\left[(\boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta})(\boldsymbol{\xi}'\boldsymbol{\Lambda}_x' + \boldsymbol{\delta}')\right] \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\Lambda}_y\mathbf{A}\left[\boldsymbol{\Gamma}E(\boldsymbol{\xi}\boldsymbol{\xi}')\boldsymbol{\Gamma}' + E(\boldsymbol{\zeta}\boldsymbol{\zeta}')\right]\mathbf{A}'\boldsymbol{\Lambda}_y' + E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') & \boldsymbol{\Lambda}_y E\left[(\mathbf{A}\boldsymbol{\Gamma}\boldsymbol{\xi})\boldsymbol{\xi}'\right]\boldsymbol{\Lambda}_x' \\ \boldsymbol{\Lambda}_x E\left[\boldsymbol{\xi}(\boldsymbol{\xi}'\boldsymbol{\Gamma}'\mathbf{A}')\right]\boldsymbol{\Lambda}_y' & \boldsymbol{\Lambda}_x E(\boldsymbol{\xi}\boldsymbol{\xi}')\boldsymbol{\Lambda}_x' + E(\boldsymbol{\delta}\boldsymbol{\delta}') \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\Lambda}_y\mathbf{A}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})\mathbf{A}'\boldsymbol{\Lambda}_y' + \boldsymbol{\Theta}_\epsilon & \boldsymbol{\Lambda}_y\mathbf{A}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Lambda}_x' \\ \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Gamma}'\mathbf{A}'\boldsymbol{\Lambda}_y' & \boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}_x' + \boldsymbol{\Theta}_\delta \end{bmatrix}.
\end{aligned}$$

2.2 Järjestysasteikolliset havaitut muuttujat

Käytännössä oletus havaittujen muuttujien yhteisjakauman normaalijakautuneisuudesta ei aina toteudu. Jakaumat voivat teoriassa olla normaalijakautuneita, mutta mittavirheet tai otoksen vinous voi vääristää havaittua jakaumaa. Mikäli havaittu muuttuja on järjestysasteikollinen tai kaksiluokkainen, normaalijakaumaoletus ei oletusarvoisesti päde. Tällaisessa tapauksessa mallin havaittuja muuttujia tulee muokata, jotta mallin parametrit voi estimoida.

Järjestysasteikollisten muuttujien käsittelyyn kehitetty metodi olettaa, että havaitut järjestysasteikolliset muuttujat ovat pohjimmiltaan normaalijakautuneiden lantenttien muuttujien (merkitään yläindeksillä '*', esimerkiksi $\boldsymbol{x}^*$) indikaattoreita. Tällöin kaavan 2.1 yhtälö $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ ei enää päde, vaan päähypoteesi tulee esittää muodossa

$$\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}(\boldsymbol{\theta}),$$

jossa $\boldsymbol{\Sigma}^*$ on jatkuvien muuttujien kovarianssirakenne koko populaatiossa. (Bollen 1989).

Kun mallinnettavat havaitut muuttujat ovat dikotomisia tai järjestysasteikollisia, mittamalleja 2.3 ja 2.4 kuvaavat kaavat eivät välttämättä päde. Järjestysasteikollisten muuttujien $\boldsymbol{x_q}$ tai $\boldsymbol{y_p}$ jakaumat voivat erota oleellisesti niiden taustalla piilevien jatkuvien muuttujien $\boldsymbol{x_q^*}$ tai $\boldsymbol{y_p^*}$ jakaumista. (Bollen 1989). Esimerkiksi järjestysasteikollisten muuttujien luokkavälit eivät välttämättä ole yhtä pitkät, jolloin normaalijakautuneisuusoletukset eivät päde. Mikäli piilevät jatkuvat muuttujat oletetaan moniulotteisesti normaalijakautuneiksi, kaavan 2.3 voi korjata korvaamalla vektorin $\boldsymbol{x}$ vektorilla $\boldsymbol{x}^*$ ja kaavan 2.4 voi korjata korvaamalla vektorin $\boldsymbol{y}$ vektorilla $\boldsymbol{y}^*$ (Bollen 1989).

Järjestysasteikollinen havaittu muuttuja ei voi olla lineaarisessa suhteessa tätä vastaavan piilevän jatkuvan muuttujan kanssa, joten näiden kahden välille on sovitettava epälineaarinen funktio (Bollen 1989). Tämän seurauksena myöskään havaittujen ja latenttien muuttujien varianssit eivät voi olla lineaarisessa suhteessa toisiinsa. Yleensä tulkinnan helpottamiseksi latentti jatkuva muuttuja standardoidaan, jolloin kovarianssimatriisi ja korrelaatiomatriisi ovat samat.

Kahden järjestysasteikollisen muuttujan jatkuvien vastineiden keskinäistä korrelaatiota kutsutaan *polykoriseksi* korrelaatioksi. Mikäli havaitut muuttujat ovat dikotomisia, tätä korrelaatiota kutsutaan *tetrakoriseksi*, kun taas järjestysasteikollisen ja jatkuvan muuttujan vastaavaa korrelaatiota kutsutaan *polyseriaaliseksi*. (Bollen 1989).Kunkin näistä voi estimoida yhdessä tai kahdessa vaiheessa, joskin kaksivaiheinen menetelmä on teknisesti tehokkaampi ja täten rakenneyhtälömallinnukseen soveltuvat ohjelmistot soveltavat ainoastaan tätä. (Maydeu-Olivares, García-Forero, Gallardo-Pujol & Renom 2009).

Kaksivaiheinen estimaatiomenetelmä Olsson (1979) esitteli ensimmäisenä kaksivaiheisen menetelmän polykoristen korrelaatioiden laskemiseen. Menetelmässä piilevän jatkuvan muuttujan arvot määritetään niitä vastaavista diskreeteistä, järjestysasteikollisista muuttujista monotonisen funktion avulla. Ensin estimoidaan kynnysarvot kullekin standardoidulle reunajakaumalle $\boldsymbol{x}_q^*$ ja $\boldsymbol{y}_p^*$ ($q = 1, 2, \ldots,Q$ ja p=$1, 2, \ldots,P$, jossa Q ja P kuvaavat eksogeenisten ja endogeenisten havaittujen, ei-jatkuvien muutujien määrää matriiseissa $\boldsymbol{x}$ ja $\boldsymbol{y}$). Jotta malli olisi toimiva, täytyy kullekin muuttujalle p määrittää sellaiset kynnysarvot a_k, joille seuraava kaava pätee:

$$y_{pi} = k \Leftrightarrow a_{pk-1} < y_{pi}^* < a_{pk}. \tag{2.5}$$

Kaavassa 2.5, alaindeksi i viittaa yksilöihin, k=1, 2, $\ldots, c$, ja c on vektorin $\boldsymbol{y}_p$ kategorien lukumäärä. Vastaavasti vektorin $\boldsymbol{x}_q$ kategorien määrä on d. Yksinkertaisuuden vuoksi jatkossa a_{pk} on lyhennetty muotoon a_k. Alin ja ylin kynnysarvo ovat aina a_0=$-\infty$ ja $a_c = \infty$, ja loput $c - 1$ kynnysarvot estimoidaan

seuraavan kaavan avulla:

$$a_k = \Phi^{-1}\left(\sum_{i=1}^{k} \frac{N_i}{N}\right), \tag{2.6}$$

jossa Φ kuvaa standardoidun normaalijakauman kertymäfunktioita ja N_i on havaintojen määtä i:nnessä kategoriassa. (Bollen 1989; Asparouhov & Muthén 2007).

Seuraavan vaiheen tavoite on estimoida polykorinen korrelaatio suurimman uskottavuuden menetelmällä olettaen kaavalla 2.6 määritetyt kynnysarvot (Jöreskog 1990). Olkoon π_{kl} todennäköisyys sille, että havainto kuuluu soluun (k, l), joka kaksiulotteisen normaalijakauman tapauksessa, korrelaatiokertoimen ollessa ρ, on

$$\begin{aligned}\pi_{kl} &= Pr[x_q^* = k, y_p^* = l] = \int_{a_{k-1}}^{a_k}\int_{b_{l-1}}^{b_l} \phi_2(u,v)dvdu \\ &= \Phi_2\left(a_k, b_l\right) - \Phi_2\left(a_{k-1}, b_l\right) - \Phi_2\left(a_k, b_{l-1}\right) + \Phi_2\left(a_{k-1}, b_{l-1}\right),\end{aligned}$$

jossa ϕ_2 on kaksiulotteisen normaalijakauman tiheysfunktio, Φ_2 vastaava kertymäfunktio ja a_k sekä b_l ovat muuttujien $\boldsymbol{x}_q^*$ ja $\boldsymbol{y}_p^*$ kynnysarvot. Täten uskottavuusfunktio, joka maksoimoidaan, on muotoa

$$L = C\prod_{k}^{d}\prod_{l}^{c}\pi_{kl}^{N_{kl}},$$

jossa C on tässä yhteydessä epärelevantti vakio, joka voidaan tiputtaa kaavasta tässä vaiheessa, ja N_{kl} kuvaa havaintojen määrää solussa (k, l). Kun yhtälö logaritmoidaan molemmin puoliin, saadaan yhtälö helpommin ratkaistavaan muotoon

$$\ln L = \sum_{k}^{d}\sum_{k}^{c} N_{kl}\ln(\pi_{kl}),$$

jonka maksimi löydetään derivoimalla yhtälö tuntemattoman parametrin ρ suhteen ja ratkaisemalla nollakohta. (Olsson 1979). Lopulta täytyy enää ratkaista yhtälö

$$\begin{aligned}\frac{\partial l}{\partial \rho} &= \sum_{k=1}^{d}\sum_{l=1}^{c}\frac{N_{kl}}{\pi_{kl}}\left(\frac{\partial \pi_{kl}}{\partial \rho}\right) \\ &= \sum_{k=1}^{d}\sum_{l=1}^{c}\frac{N_{kl}}{\pi_{kl}}\left[\phi_2\left(a_k, b_l\right) - \phi_2\left(a_{k-1}, b_l\right) - \phi_2\left(a_k, b_{l-1}\right) + \phi_2\left(a_{k-1}, b_{l-1}\right)\right] = 0.\end{aligned}$$

Olssonin (1979) mukaan absoluuttiset erot todellisten ja estimoitujen korrelaatioiden välillä ovat käytännössä hyvin pieniä. Teoriassa polykorinen korrelaatiomatriisi onkin matriisin $\boldsymbol{\Sigma}^*$ tarkentuva estimaattori, minkä vuoksi hypoteesia $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ voi testata tässä alaluvussa kuvatulla menetelmällä (Bollen 1989.)

2.3 Painotettu pienimmän neliösumman estimaattori

Käytännössä sekä mallin parametrit että populaation kovarianssit ja varianssit ovat tuntemattomia. Oletetut populaatioarvot johdetaan siis otoskovarianssimatrisiista $\mathbf{S} = \hat{\mathbf{\Sigma}}$. Tuntemattomat rakenneparametrit vektorissa $\boldsymbol{\theta}$ estimoidaan minimoimalla rakennekovarianssimatriisin $\mathbf{\Sigma}(\boldsymbol{\theta})$ ja matriisin $\mathbf{S}$ erot sopivalla sovitusfunktiolla F. (Bollen 1989).

On olemassa lukuisia estimaattifunktiota $F(\mathbf{S}, \mathbf{\Sigma}(\boldsymbol{\theta}))$. Suurimman uskottavuuden estimointi on suositeltava, jos havaitut muuttujat noudattavat normaalijakaumaa. Sopivia estimointivaihtoehtoja ovat myös yleistetty pienimmän neliösumman (*generalised least squares* - GLS) ja painottamaton pienimmän neliösumman (*unweighted least squares* - ULS) menetelmät, joiden matemaattiset ominaisuudet tunnetaan, mutta joiden estimointi vaatii vähemmän informaatiota kuin ML-estimointi. Mikäli mallin havaitut muuttujat eivät ole jatkuvia ja normaalijakautuneita, suositellaan sovitusfunktioksi painotettua pienintä neliösummaa (*weighted least squares* - WLS). Lein (2009) tutkimuksessa WLS oli verrattain harhaton yli sadan havainnon otoksilla. Jos estimointi toteutetaan korrelaatiomatriisin avulla, WLS ja sen johdannaiset ovat ainoita mielekkäitä estoimointivaihtoehtoja (Bollen 1989; Boulton 2011; Hox, Maas &Brinkhuis 2010; Olsson 1979; Yu 2002).

WLS-menetelmän sovitusfunktiossa on olennaista määrittää vektori $\hat{\boldsymbol{\rho}}$, joka sisältää kaikki otoskovarianssi- tai -korrelaatiomatriisin elementit duplikaatteja lukuunottamatta. Vektori $\boldsymbol{\sigma}(\boldsymbol{\theta})$ sisältää matriisin $\mathbf{\Sigma}(\boldsymbol{\theta})$ elementit ilman duplikaatteja. Täten sovitusfunktio WLS-menetelmällä on muotoa

$$F_{\mathrm{WLS}} = \left[\hat{\boldsymbol{\rho}} - \boldsymbol{\sigma}\left(\boldsymbol{\theta}\right)\right]' \mathbf{W} \left[\hat{\boldsymbol{\rho}} - \boldsymbol{\sigma}\left(\boldsymbol{\theta}\right)\right]. \tag{2.7}$$

Vektorin $\boldsymbol{\theta}$ arvot valitaan siten, että ne minimoivat vektoreiden $\hat{\boldsymbol{\rho}}$ ja $\boldsymbol{\sigma}(\boldsymbol{\theta})$ erotuksen painotetun summan. Tämän seurauksena $\hat{\boldsymbol{\theta}}$ on vektorin $\boldsymbol{\theta}$ tarkentuva estimaattori sillä ehdolla, että alkuperäinen oletus $\mathbf{\Sigma} = \mathbf{\Sigma}(\boldsymbol{\theta})$ on tosi.

Jotta painomatriisi $\mathbf{W}$ voitaisiin määrittää, täytyy esitellä matriisi $\mathbf{G}$, joka on matriisin $\hat{\boldsymbol{\rho}}$ asymptoottinen kovarianssimatriisi tai sen tarkentuva estimaattori. On todistettu, että mikäli $\mathbf{W}{=}\mathbf{G}^{-1}$, niin funktion F_{WLS} avulla määritetty $\hat{\boldsymbol{\theta}}$ on asymptoottisesti tehokas kaikille funktioille, joille kaava 2.7 pätee. (Bollen 1989). Jos painomatriisi on sen sijaan identiteettimatriisi ($\mathbf{W}{=}\mathbf{I}$), on kyseessä ULS- estimaattori. Jos taas painomatriisiksi valitaan ainoastaan matriisin $\mathbf{G}^{-1}$ diagonaalilla olevat elementit ja loput kiinnitetään nolliksi ($\mathbf{W}{=}\mathbf{G}_0^{-1}$), saadaaan estimaattoriksi vino painotettu pienin neliösumma (*diagonal weighted least squares* - DWLS). (Asparouhov & Muthén 2007.)

Seuraavassa luvussa esiteltävässä monitasomallinnuksessa käytetään usein funktiota F_{DWLS}, sillä se on laskennallisesti tehokkaampi ja joustavampi kuin F_{WLS} (Hox 2010). Näiden kahden sovitusfunktion asymptoottiset ominaisuudet eroavat kuitenkin toisistaan, jolloin samat taustaoletukset eivät päde ja harhaa tulee korjata (Asparouhov & Muthén 2007). Tässä katsauksessa ei kuitenkaan esitellä näitä harhaa korjaavia toimenpiteitä.

3 Monitasoinen rakenneyhtälömallinnus

Yhteiskuntatieteissä aineistot ovat usein hiearkkisesti rakentuneet. Vaikka pääkiinnostuksen kohde tutkimuksessa olisikin yksilössä, ryväsotokset voivat säästää aikaa ja olla kustannustehokkaampia kuin täysin satunnainen otos suuresta populaatiosta. Mikäli havaintojen ei voi olettaa olevan riippumattomia toisistaan, täytyy analyysi suorittaa useammalla tasolla esimerkiksi monitasomallinnuksen keinoin. Monitasomenetelmin voi tarkastella lukuisia erilasia tilastollisia malleja, joista tässä esittelen rakenneyhtälömallinnuksen ei-normaalijakautuneille muuttujille.

Hoxin (2002) opasta mukaillen monitasoinen kovarianssirakenne voidaan ilmaista yksinkertaisimmillaan seuraavanlaisesti. Olkoon kuvattava aineisto p-ulotteisessa matriisissa $\boldsymbol{y}_{ij}$, jossa i viittaa yksilöihin ja j klustereihin. Matriisin $\boldsymbol{y}_{ij}$ havaittujen arvojen (*total score*) oletetaan koostuvan kahdesta osasta:

$$\begin{aligned} \boldsymbol{y}_{\mathrm{t}} &= \boldsymbol{y}_w + \boldsymbol{y}_b \\ &= \boldsymbol{\Lambda}_w \boldsymbol{\eta}_{wij} + \boldsymbol{\epsilon}_{wij} + \boldsymbol{\Lambda}_b \boldsymbol{\eta}_{bj} + \boldsymbol{\epsilon}_{bj} \end{aligned} \tag{3.1}$$

jossa $\boldsymbol{y}_b$ on ryhmien välinen komponentti eli ryhmäkeskiarvo $\overline{\boldsymbol{y}}_j$, ja $\boldsymbol{y}_w$ on ryhmän sisäinen komponentti eli yksilön erotus ryhmäkeskiarvostaan $\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_j$. Erona edellisessä luvussa kuvattujen rakenneyhtälömallien merkintöihin monitason rakenneyhtälömalleissa matriisi $\boldsymbol{x}$ kuvaa vain yhdellä tasolla vaihtelevia muuttujia ja $\boldsymbol{y}$ kuvaa useammalla tasolla vaihtelevia muuttujia. Tarkempi erittely rakenteesta esitellään alaluvussa 3.1.

Monitasomalleja voidaan kutsua myös satunnaisten vaikutusten malleiksi, joka terminä kuvaa hyvin matriisin $\boldsymbol{y}$ monitasoista rakennetta. Yksittäisestä muuttujasta $\boldsymbol{y}_p$ erotetaan klusteritason ($\boldsymbol{y}_{bp}$) ja yksilötason ($\boldsymbol{y}_{wp}$) vaikutus, jolloin saadaan kaksi uutta riippumatonta latenttia muuttujaa. Ryhmien välinen komponentti eli klusterivaikutus on muuttuja, joka sisältää kunkin klusterin yksilöiden keskiarvon. Näiden klusterikeskiarvojen oletetaan olevan satunnainen otos ryhmävakiotermien populaatiosta. Yksilötasolla nämä ryhmäkeskiarvot ajatellaan satunnaisiksi vakiotermeiksi, kun taas klusteritasolla ne ovat mallinnettavia satunnaismuuttujia. Täten voidaan samanaikaisesti tarkastella sekä ryhmien että yksilöiden välisiä riippuvuuksia.

Kuten havaitut arvot kaavassa 3.1, myös monitasoinen populaatiokovarianssimatriisi koostuu kahdesta osasta:

$$\boldsymbol{\Sigma}_{\mathrm{t}} = \boldsymbol{\Sigma}_w + \boldsymbol{\Sigma}_b. \tag{3.2}$$

Kaavassa 3.2 $\boldsymbol{\Sigma}_b$ on matriisin $\boldsymbol{y}_b$ ryhmäkeskiarvojen populaatiokovarianssimatriisi ja $\boldsymbol{\Sigma}_w$ on yksilöiden ja ryhmäkeskiarvojen erotusten $\boldsymbol{y}_w$ kovarianssimatriisi. (Hox 2002).

Populaatiokovarianssien $\boldsymbol{\Sigma}_w$ ja $\boldsymbol{\Sigma}_b$ estimaattoreina toimivia otoskovarianssimatriiseja $\mathbf{S}_b$ ja $\mathbf{S}_w$ ei voi laskea suoraviivaisesti etenkään ei-normaalijakautuneiden

havaittujen muuttujien tapauksessa. (Hox 2002). Seuraavaksi esittelen näiden matriisien $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_w$ estimoinnissa käytetyn robustin menetelmän. Ainakin pro gradu -työn tekohetkellä syksyllä 2012 menetelmää sovelsi ainoastaan Mplus-ohjelmisto, jota kehittävien Asparouhovin ja Muthénin julkaisuun vuodelta 2007 esiteltävä estimointimenetelmä pohjaa.

Estimointi toteutetaan pääosin samoissa vaiheissa kuin perinteiset rakenneyhtälömallit. Ensin määritellään testattava malli, ja toiseksi lasketaan otosestimaatit aineistosta. Ei-normaalijakautuneiden hiearkkisesti rakentuneiden muuttujien tapauksessa otosestimaattien laskennassa käytetään muun muassa suurimman uskottavuuden menetelmää EM-algoritmin ja numeerisen integroinnin ohella. Ensin estimoidaan yksiulotteiset parametrit ja sen jälkeen kaksiulotteiset parametrit. Vasta tämän jälkeen rakenneyhtälömallin parametrit voidaan estimoida minimoimalla WLS-menetelmän sovitusfunktio. Lopulliset estimaatit saadaan, kun malli esitetään otosestimaattien funktiona. Menetelmän on kehittänyt Muthén (1984), ja tämä on ollut verrattain vähän aikaa saatavilla kahden tai useamman tason malleille. Menetelmä on teknisesti tehokas, tarkka ja joustava havaittujen muuttujien jakaumien suhteen (Asparouhov & Muthén 2007).

3.1 Mallin määritys

Tässä alaluvussa määrittelen rakennemallin muodon. Ensimmäisessä vaiheessa määritetään piilevät latentit muuttujat $\boldsymbol{y}^*$. Mikäli havaitut muuttujat $\boldsymbol{y}_p$ ovat normaalijakautuneet, $\boldsymbol{y}_p = \boldsymbol{y}_p^*$ pitää paikkansa. Muutoin teoreettiset kynnysarvot muodostetaan samoin perustein kuin alaluvussa 2.2:

$$y_{pij} = k \Leftrightarrow \tau_{pk-1} < y^*_{pij} < \tau_{pk}. \tag{3.3}$$

Piilevä normaalijakautunut latentti $\boldsymbol{y}_p^*$ koostuu sekin kahdesta normaalijakautuneesta itsenäisestä latentista osasta:

$$y^*_{pij} = y_{wpij} + y_{bpj}, \tag{3.4}$$

jossa $j = 1, \ldots, C$ edustaa klustereita, $i = 1, \ldots, N_j$ havaintoja kussakin klusterissa ja $p = 1, \ldots, P$ niitä havaittuja muuttujia, jotka ovat olemassa molemmilla tasoilla. Yksilöefekti y_{wpij} sekä klusteriefekti y_{bpj} ovat riippumattomia normaalijakautuneita latentteja muuttujia.

Rakennemallia varten määritellään normaalijakautuneet latentit vektorimuuttujat $\boldsymbol{\eta}_{wij} = (\eta_{w1ij}, \ldots, \eta_{wM_1ij})$ yksilötasolla ja $\boldsymbol{\eta}_{bj} = (\eta_{b1j}, \ldots, \eta_{bM_2j})$ klusteritasolla. Näissä M_1 viittaa yksilötason ja M_2 klusteritason latenttien muuttujien määrään. Riippumattomat muuttujat, Q_1-ulotteinen $\boldsymbol{x}_{wij}$ sekä Q_2-ulotteinen $\boldsymbol{x}_{bj}$, ilmaistaan muodossa x_{wq_1ij} sekä x_{bq_2j}, joissa $q_1 = 1, \ldots, Q_1$ ja $q_2 = 1, \ldots, Q_2$.

Täten yksilö- ja ryhmätason rakennemallit ovat seuraavaa muotoa:

$$\begin{cases} \boldsymbol{y}_{wij} & = \boldsymbol{\Lambda}_w \boldsymbol{\eta}_{wij} + \boldsymbol{\varepsilon}_{wij} \\ \boldsymbol{\eta}_{wij} & = \mathbf{B}_w \boldsymbol{\eta}_{wij} + \boldsymbol{\Gamma}_w \mathbf{x}_{wij} + \boldsymbol{\xi}_{wij} \end{cases} \tag{3.5}$$

$$\begin{cases} \boldsymbol{y}_{bj} & = \boldsymbol{\nu}_b + \boldsymbol{\Lambda}_b \boldsymbol{\eta}_{bj} + \boldsymbol{\varepsilon}_{bj} \\ \boldsymbol{\eta}_{bj} & = \boldsymbol{\alpha}_b + \mathbf{B}_b \boldsymbol{\eta}_{bj} + \boldsymbol{\Gamma}_b \mathbf{x}_{bj} + \boldsymbol{\xi}_{bj}, \end{cases} \tag{3.6}$$

joissa matriisit $\boldsymbol{\Lambda}_w$, $\mathbf{B}_w$, $\boldsymbol{\Gamma}_w$, $\boldsymbol{\nu}_b$, $\boldsymbol{\Lambda}_b$, $\boldsymbol{\alpha}_b$, $\mathbf{B}_b$ ja $\boldsymbol{\Gamma}_b$ sisältävät estimoitavat parametrit. Ryhmätason vakiotermivektorit $\boldsymbol{\nu}_b$ ja $\boldsymbol{\alpha}_b$ eivät ole tulkinnallisesti kiinnostavia. Residuaalimatriisit $\boldsymbol{\varepsilon}_{wij}$, $\boldsymbol{\xi}_{wij}$, $\boldsymbol{\varepsilon}_{bj}$ ja $\boldsymbol{\xi}_{bj}$ ovat riippumattomia ja normaalijakautuneita, ja näiden keskiarvo on 0. Residuaalimatriisien vastaavia kovariansseja merkitään matriisein $\boldsymbol{\Theta}_w$, $\boldsymbol{\Psi}_w$, $\boldsymbol{\Theta}_b$ ja $\boldsymbol{\Psi}_b$. Identifioituvuuden takaamiseksi residuaalin $\boldsymbol{\varepsilon}_{wpij}$ varianssi kiinnitetään luvuksi 1, jos p:nnes muuttuja ei ole jatkuva. Myös muita kiinnityksiä voidaan joutua tekemään identifioituvuuden takaamiseksi mallista riippuen.

3.2 Otosestimaatit

Otosestimaattien laskenta on monimutkaisempaa, kun aineisto on hierarkkinen ja havaitut muuttujat eivät ole jatkuvia. Tässä alaluvussa esittelen datamatriisin esitysmuodon ja tiivistelmän laskennallisista menetelmistä, joita estimoinnissa käytetään.

Aineisto esitetään niin sanotun saturoidun mallin muodossa. Tällöin malliin ei oleteta latentteja muuttujia $\boldsymbol{\eta}_{wij}$ tai $\boldsymbol{\eta}_{bj}$, ja täydet kovarianssimatriisit sovitetaan sekä yksilö- että ryhmätason muuttujille. Kategoristen muuttujien kynnysarvot muodostetaan samalla tavoin kuin yhden tason rakenneyhtälömalleissa 2.5, ja piilevä latentti $\boldsymbol{y}_p^*$ määritellään kuten rakennemallissa kaavassa 3.4:

$$y_{pij} = k \Leftrightarrow a_{pk-1} < y^*_{pij} < a_{pk} \tag{3.7}$$

$$y^*_{pij} = y_{wpij} + y_{bpj}. \tag{3.8}$$

Kuten aiemmin kuvattiin, saturoitu malli on muotoa

$$\begin{aligned} \boldsymbol{y}_{wij} &= \boldsymbol{\Pi}_w \mathbf{x}_{wij} + \boldsymbol{\epsilon}_{wij} \\ \boldsymbol{y}_{bj} &= \boldsymbol{\mu}_b + \boldsymbol{\Pi}_b \mathbf{x}_{bj} + \boldsymbol{\epsilon}_{bj}. \end{aligned} \tag{3.9}$$

Residuaalivektorit $\boldsymbol{\epsilon}_{wij}$ ja $\boldsymbol{\epsilon}_{bj}$ oletetaan normaalijakautuneiksi nollan keskiarvoilla, ja näiden kovarianssimatriisit ovat $\boldsymbol{\Sigma}_w$ ja $\boldsymbol{\Sigma}_b$. Mikäli p:nnes muuttuja on kategorinen, täytyy identifioituvuuden vuoksi residuaalin ϵ_{wpij} varianssi kiinnittää luvun 1 suuruiseksi ja keskiarvoparametri μ_{pb} luvun 0 suuruiseksi.

Kaavan 3.9 parametrit estimoidaan kahdessa vaiheessa. Ensin estimoidaan yksiulotteiset parametrit: ryhmätason keskiarvot μ_{bp}, kynnysarvot a_{pk}, kertoimet Π_{wpq} ja Π_{bpq} ja residuaalikovarianssimatriisin diagonaalilla olevat elementit Σ_{wpp} ja Σ_{bpp}. Seuraavassa vaiheessa residuaalien kovarianssit eli diagonaalin ulkopuoliset elementit matriiseista $\boldsymbol{\Sigma}_w$ ja $\boldsymbol{\Sigma}_b$ estimoidaan kaksiulotteisilla uskottavuusmenetelmillä olettaen yksiulotteiset estimaatit tunnetuiksi.

Yksi- ja kaksiulotteisen suurimman uskottavuuden estimoinnin menetelmä on yleistetty kasvukäyräsekamalleille (*growth mixture models*) kehitetystä menetelmästä (Asparouhov & Muthén 2008). Metodi perustuu EM-algoritmiin,

jossa latentteja muuttujia $\boldsymbol{y}_{wij}$ ja $\boldsymbol{y}_{bj}$ käsitellään kuten puuttuvaa tietoa. EM-algoritmi on iteratiivinen menetelmä, jota käytetään suurimman uskottavuuden estimoinnissa, kun aineistoa puuttuu tai se on rakenteeltaan monimutkainen, esimerkiksi hierarkkinen. Tässä yhteydessä algoritmia ei esitellä sen tarkemmin, mutta sanottakoon, että EM-algoritmi yksinkertaistaa suurimman uskottavuuden estimointia. (McLachlan & Krishnan 1997.)

Kun kaksivaiheinen estimointi on suoritettu, yksi- ja kaksiulotteiset estimaatit kootaan vektoriin $\mathbf{s}$, jolle lasketaan asymptoottinen kovarianssimatriisi $\mathbf{G}$. Muthén ja Satorra (1995) ovat näyttäneet, miten matriisi $\mathbf{G}$ muodostetaan siten, että se on estimoitujen parametrien tarkentuva kovarianssiimatriisi. Lyhyesti sanoen kovarianssit perustuvat yksi- ja kaksiulotteisten estimaattien uskottavuuksien ensimmäisiin derivaatteihin sillä ehdolla, että kaksiulotteiset estimaatit ovat ehdollisia yksiulotteisille estimaateille. Matriisin $\mathbf{G}$ käänteismatriisia käytetään painomatriisina WLS-menetelmän estimoinnin viimeisessä vaiheessa.

3.3 Rakennematriisin estimointi

Kahden tason WLS-estimoinnin kolmannessa vaiheessa rakennemallin parametrit estimoidaan kuten Muthénin julkaisussa vuodelta 1984, mutta laajentamalla metodia kahden tason tapaukseen. Määritellään ensin malli. Kynnysarvot sekä monitasoinen rakenne esitetään kuten saturoidun mallin tapauksessa (3.7 - 3.9):

$$
\begin{aligned}
y_{pij} = k \Leftrightarrow a^*_{pk-1} < y^*_{pij} < a^*_{pk} \\
\boldsymbol{y}_{wij} = \boldsymbol{\Pi}^*_w \mathbf{x}_{wij} + \boldsymbol{\epsilon}_{wij} \\
\boldsymbol{y}_{bj} = \boldsymbol{\mu}^*_b + \boldsymbol{\Pi}^*_b \mathbf{x}_{bj} + \boldsymbol{\epsilon}_{bj}.
\end{aligned}
$$

Rakennemalli on saturoidun mallin sisäkkäinen malli (*nested*). Indeksillä '*' merkitään niitä osia mallista, jotka sisältävät estimoitavat parametrit. Näiden esitysmuoto esitetään seuraavaksi. Kaavoissa 3.5 ja 3.6 kuvattujen teoreettisten rakennemallien yhdistetyt muodot ovat

$$
\begin{aligned}
\boldsymbol{y}_{wij} = \boldsymbol{\Lambda}_w (\mathbf{I} - \mathbf{B}_w)^{-1} (\boldsymbol{\Gamma}_w \boldsymbol{x}_{wij} + \boldsymbol{\xi}_{wij}) + \boldsymbol{\epsilon}_{wij} \\
\boldsymbol{y}_{bj} = \boldsymbol{\nu}_b + \boldsymbol{\Lambda}_b (\mathbf{I} - \mathbf{B}_b)^{-1} (\boldsymbol{\alpha}_b + \boldsymbol{\Gamma}_b \boldsymbol{x}_{bj} + \boldsymbol{\xi}_{bj}) + \boldsymbol{\epsilon}_{bj}.
\end{aligned}
$$

Yllä olevat kaavat ovat hyödyllisiä, kun määritetään saturoidun mallin (3.7 - 3.9) ja rakennemallin (3.3 - 3.6) välisiä suhteita. Esimmäisenä yhtälöt ratkaistaan standardoimattomien estimaattien (merkitään yläindeksillä '**') suhteen. Näitä ovat kertoimet $\boldsymbol{x}$ molemmilla tasoilla, ryhmätason vakiotermit sekä resi-

duaalimuuttujien kovarianssit molemmilla tasoilla:

$$\boldsymbol{\Pi}_w^{**} = \boldsymbol{\Lambda}_w(\mathbf{I} - \mathbf{B}_w)^{-1}\boldsymbol{\Gamma}_w \tag{3.10}$$

$$\boldsymbol{\Pi}_b^{**} = \boldsymbol{\Lambda}_b(\mathbf{I} - \mathbf{B}_b)^{-1}\boldsymbol{\Gamma}_b \tag{3.11}$$

$$\boldsymbol{\mu}_b^{**} = \boldsymbol{\nu}_b + \boldsymbol{\Lambda}_b(\mathbf{I} - \mathbf{B}_b)^{-1}\boldsymbol{\alpha}_b \tag{3.12}$$

$$\begin{aligned}\boldsymbol{\Sigma}_w^{**} &= \mathrm{Cov}(\boldsymbol{\epsilon}_w) = \mathrm{Cov}(\boldsymbol{\Lambda}_w(\mathbf{I} - \mathbf{B}_w)^{-1}\boldsymbol{\xi}_w + \boldsymbol{\varepsilon}_w) \\ &= \boldsymbol{\Lambda}_w(\mathbf{I} - \mathbf{B}_w)^{-1}\boldsymbol{\Psi}_w \left[(\mathbf{I} - \mathbf{B}_w)^{-1}\right]^T \boldsymbol{\Lambda}_w^T + \boldsymbol{\Theta}_w\end{aligned} \tag{3.13}$$

$$\begin{aligned}\boldsymbol{\Sigma}_b^{**} &= \mathrm{Cov}(\boldsymbol{\epsilon}_b) = \mathrm{Cov}(\boldsymbol{\Lambda}_b(\mathbf{I} - \mathbf{B}_b)^{-1}\boldsymbol{\xi}_b + \boldsymbol{\varepsilon}_b) \\ &= \boldsymbol{\Lambda}_b(\mathbf{I} - \mathbf{B}_b)^{-1}\boldsymbol{\Psi}_b \left[(\mathbf{I} - \mathbf{B}_b)^{-1}\right]^T \boldsymbol{\Lambda}_b^T + \boldsymbol{\Theta}_b.\end{aligned} \tag{3.14}$$

Saturoidun mallin kategoriset muuttujat oli määritelty siten, että klusteritason keskiarvot μ_{bp} olivat nollia ja yksilötason residuaalien varianssit ϵ_{wpij} kiinnitettiin yhden suuruisiksi. Jotta saturoidun ja rakennemallin parametreja voisi vertailla, täytyy rakennemallin parametrit 3.10 - 3.14 standardoida.

Estimaattien standardointia varten parametrivektorien ja -matriisien ne solut, jotka liittyvät kategorisiin muuttujiin, tulee painottaa. Olkoon $\boldsymbol{\Delta}_w$ p-ulotteinen diagonaalimatriisi, jonka diagonaalilla on luku $1/\sqrt{\Sigma_{wpp}^{**}}$, mikäli p:nnes muuttuja on kategorinen, ja muussa tapauksessa luku 1. Samalla tavoin olkoon $\boldsymbol{\delta}_b$ p-ulotteinen vektori, jonka p:nnes elementti on suuruudeltaan μ_{bp}^{**}, mikäli p:nnes muuttuja on kategorinen ja muuten 0. Standardoitujen kynnysarvojen sekä kaavojen $3.10 - 3.14$ parametrien määrittämiseksi seuraavien määritelmien tulee päteä:

$$\begin{aligned}\boldsymbol{a}_k^* &= \boldsymbol{\Delta}_w(\boldsymbol{\tau}_k - \boldsymbol{\delta}_b) \\ \boldsymbol{\mu}_b^* &= \boldsymbol{\Delta}_w(\boldsymbol{\mu}_b^{**} - \boldsymbol{\delta}_b) \\ \boldsymbol{\Pi}_w^* &= \boldsymbol{\Delta}_w\boldsymbol{\Pi}_w^{**} \\ \boldsymbol{\Pi}_b^* &= \boldsymbol{\Delta}_w\boldsymbol{\Pi}_b^{**} \\ \boldsymbol{\Sigma}_w^* &= \mathrm{Cov}(\boldsymbol{\Delta}_w\boldsymbol{\epsilon}_w) = \boldsymbol{\Delta}_w\boldsymbol{\Sigma}_w^{**}\boldsymbol{\Delta}_w \\ \boldsymbol{\Sigma}_b^* &= \mathrm{Cov}(\boldsymbol{\Delta}_w\boldsymbol{\epsilon}_b) = \boldsymbol{\Delta}_w\boldsymbol{\Sigma}_b^{**}\boldsymbol{\Delta}_w.\end{aligned}$$

Kategoristen muuttujien estimaatit siis jaetaan niiden yksilötason virhetermien keskihajonnoilla, ja ne keskitetään vähentämällä estimaateista ryhmätason keskiarvot. Lopuksi standardoidut estimaatit yhdistetään vektoriin $\mathbf{s}^*$ samassa järjestyksessä kuin saturoidut estimaatit yhdistettiin vektoriin $\mathbf{s}$.

WLS-estimoinnin sovitusfunktio on samankaltainen kuin yhden tason rakenneyhtälömallien estimoinnissa 2.7:

$$F_{\mathrm{WLS}} = (\boldsymbol{s} - \boldsymbol{s}^*)'\mathbf{W}(\boldsymbol{s} - \boldsymbol{s}^*). \tag{3.15}$$

Lopulliset estimaatit saadaan minimoimalla 3.15 rakennemallien parametrien (3.3)−(3.6) suhteen.

Käytännössä estimoinnissa painomatriisin diagonaalia $\mathbf{G}_0^{-1}$ käytetään useammin kuin matriisia $\mathbf{G}^{-1}$, sillä diagonaalimatriisiin liittyy vähemmän vaatimuksia klusterien määrän ja otoskoon suhteen (Hox 2010). Kyseenomainen menetelmä, diagonaalinen painotettu pienimmän neliösumman menetelmä (DWLS), on

suhteellisen harhaton silloinkin, kun havaitut muuttujat ovat jatkuvia. DWLS-estimointia on kahta eri tyyppiä, WLSM ja WLSMV, jotka erovat toisistaan sen suhteen, miten mallin sopivuuden arvioinnissa käytettävää F_{DWLS}-arvoa korjataan. Molemmat menetelmät tuottavat samat estimaatit ja keskivirheet, mutta eri yhteensopivuusarvot. WLSM korjaa F_{DWLS}-suureen keskiarvon suhteen ja WLSMV keskiarvon ja varianssin suhteen. (Hox et al. 2010.) Mallin yhteensopivuuden arviointia ei käsitellä tässä katsauksessa tämän tarkemmin.

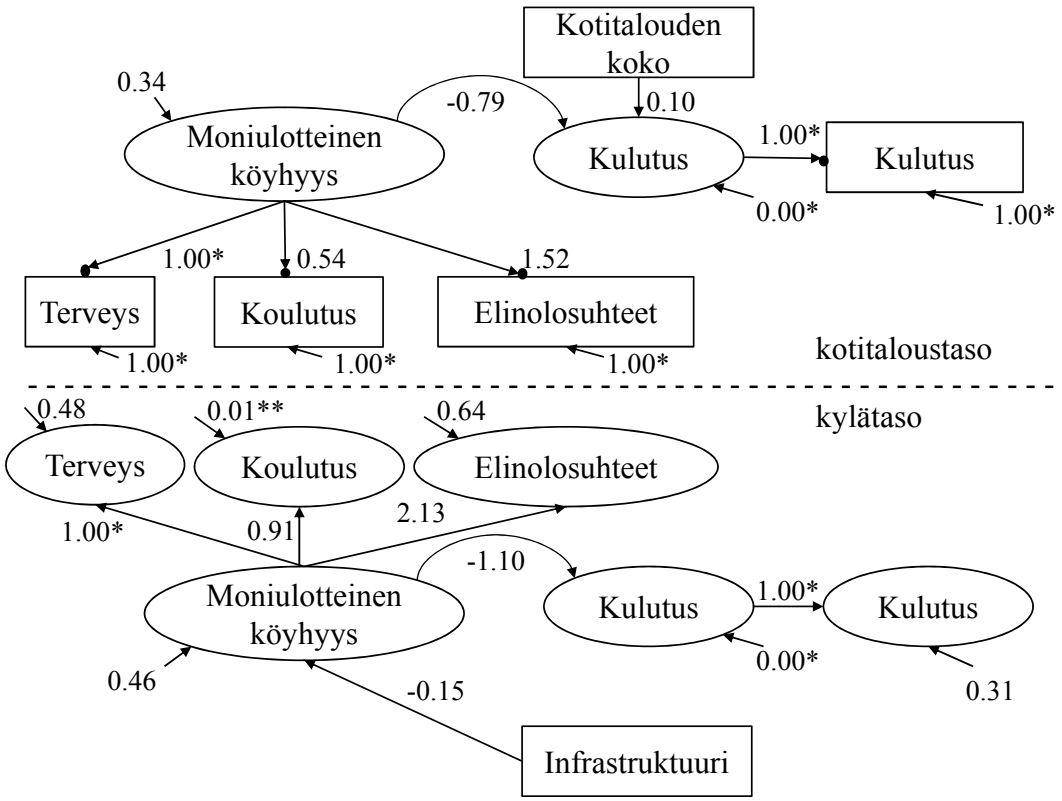
4 Tulokset

Aiemmissa luvuissa kuvattua menetelmää sovellettiin Laosista vuonna 2011 kerättyyn kotitalouskyselyyn (n=1261), joka toteutettiin klusteriotoksena 123 kylässä (Tuominen et al. 2013). Mallissa vertailtiin kahden köyhyyden määritelmän suhdetta sekä kylä- että kotitaloustasolla. Köyhyysmuuttujia olivat rahallinen kulutus sekä moniulotteinen köyhyys, jonka indikaattorit muodostettiin UNDP:n (2010) määritelmien avulla. Näiden kahden köyhyyden määritelmän suhteen tutkimista on pidetty tärkeänä jatkotutkimuksen kohteena (UNDP 2010). Kotitaloustasolla huomioitiin myös kotitalouden koon mahdollinen yhteys köyhyyteen ja kylätasolla puolestaan infrastruktuurin yhteys kylän keskimääräiseen köyhyystasoon.

Malli estimoitiin useissa vaiheissa suositusten mukaisesti. Viimeisimmän mallin sopivuus aineistoon oli hyvä (χ^2=21.4, df=10, p=0.018; RMSEA=0.027; CFI=0.98; TLI=0.95; SRMR within=0.020, between 0.062; WRMR=0.55) ja kaikki paitsi yksi parametri olivat merkitseviä 99% luottamustasolla. Kuviossa 4.1 esiteltävässä mallissa kaikki muuttujat, joihin suuntaavan nuolen päässä on piste, on estimoitu esitetyllä menetelmällä ei-jatkuville muuttujille. Muuten kuvion merkinnät ovat samat kuin rakenneyhtälömalleissa yleensä (Bollen 1989).

Tulosten mukaan merkittävä osa kotitalouksien köyhyydestä Laosissa määrittyy jo asuinkylän perusteella, sillä 21–61 % indikaattorien vaihtelusta määrittyi kylätasolla ja loput kotitaloustasolla. Myös moniulotteisen köyhyyden ja rahallisen kulutuksen suhde oli voimakkaampi kylä- kuin kotitaloustasolla. Suuremmissa kotitalouksissa kulutus oli suurempaa kuin pienemmissä kotitalouksissa, mutta yhteyttä moniulotteiseen köyhyyteen ei ollut. Kylätasolla taas nissä kylissä, joissa oli laajemmin infrastruktuuria, oli keskimääräinen köyhyystaso matalampi, mutta kulutukseen infrastruktuuri ei ollut yhteydessä.

Tulokset kannustavat tutkimaan moniulotteista köyhyyttä latenttina faktorina yhden indikaattorin sijaan ja analysoimaan köyhyyttä sekä yksilö- että yhteisötasolla samanaikaisesti. Esitelty menetelmä soveltuu erityisen hyvin klusteroitujen kyselyaineistojen tutkimiseen. Gradun tekohetkellä menetelmä rajoittui kahden tason malleihin, mutta Mplus-ohjelmiston uudemmassa ver-



Kuvio 4.1. Lopullisen mallin parametrien estimaatit, joista kaikki paitsi yksi ovat merkitseviä 99 % luottamustasolla. χ^2=21.4 (df=10, p=0.018), RMSEA=0.027, CFI=0.98, TLI=0.95, SRMR within=0.020 ja between=0.062, WRMR=0.55. *kiinnitetty **p≥0.01

siossa 7 on mahdollista mallintaa kolmea tasoa samanaikasesti vastaavin periaattein.

Lähteet

Asparouhov, T. & Muthén, B. (2007), ”Computationally Efficient Estimation of Multilevel High-Dimensional Latent Variable Models”, proceedings of the 2007 JSM meeting in Salt Lake City, Utah, *Section on Statistics in Epidemiology*, 2531–2535.

——— (2008), ”Growth Mixture Modeling: Analysis of non-Gaussian Random Effects”, in *Longitudinal Data Analysis*, eds. Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G., Boca Raton: Chapman & Hall/CRC Press.

Bollen, K. A. (1989), *Structural Equations with Latent Variables*, New York: Wiley.

Boulton, A. (2011), ”Fit Index Sensitivity in Multilevel Structural Equation Modeling”, MA thesis, University of Kansas.

Hox, J. (2002), *Multilevel analysis: Techniques and Applications*, New Jersey: Lawrence Erlbaum Associates Inc.

——— (2010), *Multilevel analysis: Techniques and Applications* (2nd ed.), New Jersey: Lawrence Erlbaum Associates Inc.

Hox, J., Maas, C. & Brinkhuis, J. (2010), "The Effect of Estimation Method and Sample Size in Multilevel Structural Equation Modeling", *Statistica Neerlandica*, 65, 157–170.

Jöreskog, K. G. (1990), "New Developments in LISREL: Analysis of Ordinal Variables Using Polychoric Correlations and Weighted Least Squares", *Quality and Quantity*, 24, 387–404.

Lei, P-W. (2009), "Evaluating Estimation Methods for Ordinal Data in Structural Equation Modeling", *Quality and Quantity*, 43, 495–507.

Maydeu-Olivares, A., García-Forero, C., Gallardo-Pujol, D. & Renom, J. (2009), "Testing Categorized Bivariate Normality with Two-Stage Polychoric Correlation Estimates", *Methodology*, 5, 131–136.

McLachlan, G. J. & Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.

Muthén, B. 0. (1984), "A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators", *Psychometrika*, 49, 115–132.

Muthén, B. 0. & Satorra, A. (1995), "Complex Sample Data in Structural Equation Modeling", *Sociological Methodology*, 25, 267-316.

Olsson, U. (1979), "Maximum Likelihood Estimation of the Polychoric Correlation Coefficient", *Psychometrika*, 44, 443–460.

Raudenbush, S. W. & Bryk, A. S. (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods*, California: Sage Publications, Inc.

Tuominen, V., Pasanen, T., Keskiväli, I., Lakkala, H., Akgün, O., Luukkanen, J., Kaivo-Oja, J. & Panula-Ontto, J. (2013), *Energy, Environment and Livelihoods in the Lao PDR. Results from a 2011 Household Survey*, Tutu e-julkaisuja 4/2013, Turun yliopisto: Tulevaisuuden tutkimuskeskus.

United Nations Development Programme (2010), *Human Development Report 2010: 20th Anniversary Edition*, New York: United Nations Development Programme.

Yu, C-Y. (2002), "Evaluation Cutoff Criteria of Model Fit Indices for Latent Variable Models with Binary and Continuous Outcomes", Ph.D. thesis, University of California.

Luonnonpopulaatioiden geneettisen rakenteen ja historian tilastollinen päättely

Jukka Sirén
Helsingin yliopisto

1 Johdanto

Populaatiogenetiikan historia on vahvasti sidoksissa tilastollisten menetelmien kehitykseen. Yhteisen perinnön selkein keulakuva on R.A. Fisher (1890-1962), joka oli uranuurtaja niin tilastotieteessä kuin populaatiogenetiikassa ja evoluutiotutkimuksessa. Fisherin roolista tilastotieteen kehityksessä onkin sanottu, että olisi helpompaa luetella tilastotieteen osa-alueet, joihin hänellä ei ollut kiinostusta, kuin ne joihin hänellä oli (Savage, 1976). Vastaavasti Fisher kehitti yhdessä J.B.S. Haldanen ja Sewall Wrightin populaatiogenetiikan matemaattisen perustan.

Väitöskirjani voikin katsoa olevan jatkoa tälle tilastotieteen ja populaatiogenetiikan väliselle yhteydelle. Oman lisänsä tähän yhteyteen tarjoavat nykyaikaisen sekvensointitekniikan tuottamat laajat geneettiset aineistot, joiden analysoimiseen tarvitaan laskennallisten ja tilastollisten menetelmien lisäksi ymmärrystä taustalla olevista biologisista ilmiöistä.

Työssä on kehitetty tilastollisia menetelmiä kahteen erilliseen, mutta toisiinsa liittyvään, populaatiogenetiikan ongelmaan. Ensimmäinen näistä käsittelee populaatioiden geneettisen rakenteen oppimista, joka muotoillaan tilastollisena ryhmittelyongelmana. Toinen käsittelee tämän havaitun rakenteen historiaa, ja siinä tavoitteena on populaatioiden yhteyksiä kuvaava puun estimointi. Väitöskirjassani (Sirén, 2012) ja sen viidessä osajulkaisussa (Corander, Sirén and Arjas, 2008; Corander et al., 2008, 2013; Sirén, Marttinen and Corander, 2011; Sirén, Hanage and Corander, 2013) on esitelty näihin ongelmiin ratkaisuja, jotka perustuvat taustalla olevien biologisten prosessien mallintamiseen.

Eri biologian osa-alueilla on populaatiolle useita toisistaan poikkeavia määritelmiä (Waples and Gaggiotti, 2006). Tässä työssä populaatiolla tarkoitetaan saman lajin yksilöiden muodostamaa ryhmää, jonka jäsenet asuvat samalla maantieteellisellä alueella ja jonka jäsenillä on mahdollisuus pariutua keskenään.

Populaatiorakenteen ja -historian oppimisen perustana on yksilöiden ja populaatioiden välinen geneettinen vaihtelu. Eliöiden perimä on koodattuna deoksiribonukleiinihappoon (DNA), joka koostuu neljästä eri emäksestä. Näiden emästen järjestys sisältää perinnöllisen informaation, jonka perusteella eliön muoto ja toiminta määräytyy. Eri eliöiden ja eliöryhmien välillä on suuria eroja perimän koossa ja rakenteessa, mikä on seurausta miljoonien vuosien vähittäisestä muutoksesta. Eliöiden välisiä sukulaisuussuhteita pystytäänkin tunnistamaan tämän vähittäisen vaihtelun perusteella niin yksilö-, populaatio-, laji- kuin laajemmillakin tasoilla.

Populaatiotason geneettistä vaihtelua tutkittaessa tärkeimmäksi työkaluksi ovat viimeisten parin vuosikymmenen aikana muodostuneet merkkigeenit. Nämä ovat lyhyitä

kohtia perimässä, jossa nähdään useita muotoja eli alleeleja eri yksilöillä ja populaatioilla. Usein merkkigeeneille ei tunneta mitään varsinaista biologista funktiota, vaan ne oletetaan satunnaisten mutaatioiden luomiksi muutoksiksi. Nykyään ehkä laajimmin käytettyjä merkkigeenejä ovat yhden emäksen muutokset (single nucleotide polymorphism, SNP), joissa yksi DNA:n emäs on muuttunut mutaation seurauksen toiseksi. Vaikka teoriasssa yhdessä SNP:ssä voisi havaita neljä eri alleelia, mutaatiot ovat käytännössä niin harvinaisia, että lähes kaikki tunnetut SNP:t ovat kaksialleelisia. Toinen varsinkin aiemmin laajalti käytetty merkkigeenityyppi on mikrosatelliitti, jossa lyhyt muutaman emäksen jakso toistuu kymmeniä tai satoja kertoja. Toistojen määrässä esiintyy suurta vaihtelua, ja toistojen lukumäärät muodostavatkin mikrosatelliitin eri alleelit.

Geneettisen vaihtelun mekanismit ovat tiettyjä poikkeuksia lukuunottamatta lähes samanlaisia riippumatta siitä, mistä eliöistä on kyse. Mutaatio, satunnaisvaihtelu ja luonnonvalinta toimivat geneettisellä tasolla yhtenevästi kaikilla eliöillä. Tässä työssä kehitettyjä menetelmiä onkin sovellettu niin bakteerien kuin ihmisten populaatioiden tutkimiseen.

Väitöskirjassa kehitettyjen menetelmien perusteena on ollut biologisten prosessien ymmärtäminen ja niiden mallintaminen todennäköisyyden sääntöjen mukaisesti. Vaikka joissain tilanteissa saataisiin lähes yhtä hyviä tuloksia yksinkertaisemmilla tilastollisilla menetelmillä, niin taustalla olevien prosessien mallintaminen auttaa ymmärtämään ongelmaa ja saatuja tuloksia paremmin. Esimerkiksi alleelifrekvenssien korrelaatioihin perustuvilla menetelmillä saadaan estimoitua populaatioiden historia lähes yhtä hyvin kuin väitöskirjani osajulkaisussa Sirén, Marttinen and Corander (2011) esitellyllä menetelmällä ja paljon pienemmällä laskennallisella vaativuudella. Yksinkertaisemmalla menetelmällä ei kuitenkaan saada samaa tulkintaa puun oksien pituuksille kuin esittelemällämme mallilla, jossa oksan pituus on suhteessa ajanjakson pituuteen ja populaation kokoon. Erilaisilla menetelmillä on paikkansa eri tutkimuksongelmissa. Monesti kuitenkin geneettisen aineiston kerääminen vaatii niin suurta työmäärää, että siitä kannattaa yrittää saada mahdollisimman paljon irti käyttäen kehittyneitä tilastollisia menetelmiä.

2 Bayesläinen tilastollinen laskenta

Kaikki väitöskirjasasa esitellyt tilastolliset menetelmät perustuvat Bayesläisiin tilastollisiin malleihin, jotka on johdettu kussakin ongelmassa taustalla olevien biologisten prosessien perusteella. Bayesläiset menetelmät tarjoavat lukuisia tehokkaita mahdollisuuksia monimutkaisten tilastollisten mallien päättelyyn. Nämä ovat kuitenkin vain välineitä, ja merkittävämpiä esitellyissä menetelmissä ovat ongelmiin johdetut uskottavuusfunktiot. Kaikilla malleilla itse päättelyn voisi tehdä myös käyttäen muita kuin Bayesläisiä menetelmiä ilman, että tulokset merkittävästi muuttuisivat. Joissain tapauksissa käytetyt menenelmät ovat identtisiä suurimman uskottavuuden menetelmän kanssa, mutta vain tulosten tulkinta on hieman erilainen.

Bayesläisten menetelmien suurin vahvuus tilastollisten mallien päättelyssä on todennäköisyyden käyttäminen kaiken epävarmuuden mittaamiseen, mikä mahdollistaa todennäköisyyslaskennan tehokkaan koneiston käyttämisen. Parametrien posteriorijakauma tiivistää kaiken epävarmuuden yhteen todennäköisyysjakaumaan, jonka perusteella voidaan johtaa estimaatteja eri suureista.

Laskennan kannalta Bayesläisen päättelyn voi muotoilla yhtenä ongelmana, joka kattaa lähes kaikki tilanteet. Useimmissa tilanteissa kiinostuksen kohteena oleva estimaatti on määritelty jonkinlaisena posterioriodotusarvona:

$$E\left(h\left(\theta\right) \mid X\right) = \int h\left(\theta\right) p\left(\theta \mid X\right) d\theta. \tag{1}$$

Tässä θ on parametrivektori, X havaittu data, $p\left(\theta \mid X\right)$ parametrien posteriorijakauma ja $h\left(\theta\right)$ funktio, joka määrittää estimaatin. Sopivalla funktion h valinnalla tällä määritelmällä voidaan laskea odotusarvoja, variansseja, välejä, reunajakaumia ja lähes mitä tahansa suureita.

Monissa käytännön ongelmissa kaavan (1) käyttöä rajoittaa se, ettei posteriorijakauma $p\left(\theta \mid X\right)$ ole mitään standardimuotoa eikä odotusarvoa pysty siten laskemaan analyyttisesti. Näissä tilanteissa joudutaankin turvautumaan erilaisiin numeerisiin approksimaatioihin, joiden kehittämisestä onkin muodostunut oma tieteenalansa tilastotieteen sisään. Tyypillisesti nämä perustuvat Monte Carlo laskentaan, jossa posteriorijakaumasta arvotaan suuri määrä parametrivektoreita $\theta_1, \ldots, \theta_N$ ja lasketaan estimaatti

$$\hat{I} = \frac{1}{N} \sum_{i=1}^{N} h(\theta_i).$$

Suurten lukujen lain perusteella estimaatti suppenee kohti posterioriodotusarvoa (1).

Näytteiden tuottaminen posteriorijakaumasta on harvoin suoraviivaista. Viimeisen 20 vuoden aikana laajaa suosiota ovat saaneet Markovin ketjuihin perustuvat Monte Carlo menetelmät (Markov chain Monte Carlo, MCMC), joissa parametrivektoreita ei arvota riippumattomasti, vaan simuloidaan Markovin ketjua, jonka tasapainojakaumana on haluttu posteriorijakauama. Yksinkertaisimmat MCMC algoritmit ovat helppo toteuttaa muutamilla koodiriveillä ja niillä voidaan laskea odotusarvoja korkeadimensioisista ja monimutkaisista jakaumista. Monte Carlo - ja erityisesti MCMC-menetelmien vahvuus onkin niiden skaalautuvuus korkeisiin dimensioihin.

Käytännössä monissa tilanteissa, joissa monimutkaisia tilastollisia malleja sovitetaan laskennallisilla menetelmillä, algoritmien ei käyttö yleensä onnstu suoraviivaisesti. Tässäkin väitöskirjassa esiteltyjen menetelmien kehittämisessä suurinta työtä on vaatinut laskennallisten työkalujen sovitus kuhunkin ongelmaan. Osa menetelmien perustana olevista tilastollisista malleista on korkeadimensioisia käsittäen tuhansia eri parametreja. Toisaalta joissan menetelmissä ongelmana on ollut parametriavaruuden monimutkaisuus, kuten populaatiorakennetta kuvaavien joukon ositusten kohdalla.

3 Geneettinen vaihtelu populaatiossa

Sekä populaatioiden geneettisen rakenteen että historian oppimisen perustana on populaation sisällä vallitseva geneettinen vaihtelu. Yksinkertaisuuden vuoksi tarkastellaan yhtä merkkigeeniä, jossa havaitaan kahta eri muotoa eli alleelia. Useampi alleelisten merkkigeenien kohdalle tilanne yleistyy suoraviivaisesti. Käytännön ongelmissa merkkigeenejä on yleensä useita, vaihdellen muutamasta jopa yli miljoonaan, mutta ne voidaan tietyin ehdoin olettaa toisistaan riippumattomiksi. Tämä toteutuu, mikäli merkkigeenit sijaitsevat eri kromosomeissa tai samassa kromosomissa mutta riittävän kaukana toisistaan.

Populaatioiden välisiä eroja geneettisessä vaihtelussa tarkastellaan tässä työssä alleelifrekvenssien perusteella. Alleelifrekvenssi p kertoo tietyn alleelin suhteellisen osuuden merkkigeenissä populaation yksilöillä. Tässä työssä käytetyissä matemaattisissa malleissa oletetaan populaatioiden sisällä pariutumisen tapahtuvan satunnaisesti. Tästä seuraa se, että populaatiosta satunnaisesti valitulla yksilöllä kukin alleeli on todennäköisyydellä p määrättyä tyyppiä. Useimmat eläimet ja monet muutkin eliöt ovat diploidisia, eli niillä on kaksi erilaista kopiota perimästä: toinen äidiltä ja toinen isältä. Tällöin jos merkitään alleeleja kirjaimin a ja A, voi yksilöllä olla kolme eri yhdistelmää (genotyyppiä) näistä alleeleista: aa, aA ja AA. Satunnaisen pariutumisen vallitessa populaatiossa näiden todennäköisyydet ovat:

$$P(aa) = p^2, P(aA) = 2p(1-p), P(AA) = (1-p)^2. \tag{2}$$

Jakauma (2) tunnetaan Hardy-Weinberg -tasapainona, joka toimii perustana populaatiorakenteen oppimisessa.

4 Populaatioiden geneettinen rakenne: ryhmittely- ja luokitteluongelma

Populaatioiden geneettinen rakenne on tässä työssä ymmärretty yksilöiden jakona erillisiin populaatioihin. Tämä ei ole kaikissa tilanteissa täysin realistinen oletus, sillä useissa tutkimuksissa on osoitettu luonnonpopulaatioiden geneettisen vaihtelun olevan enemmän jatkuvaa kuin diskreettiä. Esimerkiksi Novembre et al. (2008) näytti kuinka ihmisten välinen geneettinen vaihtelu seuraa melko tarkasta maantieteellistä vaihtelua asuinpaikoissa. Yksinkertaistus erillisiin populaatioihin on kuitenkin paikallaan monissa tilanteissa, joissa ainesto on peräisin useista erillisistä paikoista eikä kata laajaa yhtenäistä aluetta.

Matemaattisesti ongelma voidaan kuvata siten, että meillä on N ryhmiteltävää kappaletta, jotka haluamme ryhmitellä K:hon homogeeniseen ryhmään. Ryhmien lukumäärä K on tuntematon ja ryhmät itsessään määriytyvät vain niiden jäsenten perusteella. Tällöin kappaleiden ryhmittely on joukon $\{1, \ldots, N\}$ ositus S, joka on kokoelma $\{s_1, \ldots, s_K\}$, missä kukin kappale kuuluu täsmälleen yhteen joukkoon s_c.

Ryhmittelyn perusteena on jokaiselta kappaleelta on mitattu L ominaisuutta ja tavoitteena on ryhmitellä ominaisuuksiltaan samanakaltaiset kappaleet samaan ryhmään. Kunkin ryhmän sisällä kappaleiden ominaisuuksien oletetaan tulevan samasta jakaumasta, jolloin ryhmän sisällä ominaisuuksille saadaan todennäköisyys

$$p(x_c|S) = \prod_{j=1}^{L} \prod_{i=1}^{n_c} p_{c,j}(x_{c,j,i}),$$

missä $x_{c,j,i}$ on ryhmän s_c yksilön i havainto omisaisuudessa j, x_c sisältää kaikkien ryhmän s_c yksilöiden ominaisuudet, n_c on yksilöiden määrä ja $p_{c,j}$ on ominaisuuden j jakauma ryhmässä s_c. Ryhmäkohtaiset jakaumat määritetään usein ryhmäkohtaisen parametrin avulla, joka kuitenkin marginalisoidaan pois:

$$p_{c,j}(x) = \int p_{c,j}(x \mid \gamma) p(\gamma) d\gamma.$$

Marginalisointi tehdään, koska parametrit itsessään eivät ole mielenkiinnon kohteena, vaan niiden rooli on vain luoda eroja ryhmien välille. Populaatiorakenteen oppimisessa ominaisuudet ovat yksilöiden genotyyppejä L:ssä eri merkkigeenissä ja niiden ryhmäkohtaiset todennäköisyydet määräytyvtä jakauman (2) mukaan, missä alleelifrekvenssi p on tuntematon parametri.

Ryhmittely tehdään ositusten posterioritodennäköisyyksien perusteella. Määrittämällä osituksille priorijakauma $p(S)$ saadaan posterioritodennäköisyydet kaavalla

$$p\left(S \mid x\right) = \frac{p\left(S \mid x\right) p(S)}{\sum_{S'} p\left(S' \mid x\right) p(S')}. \tag{3}$$

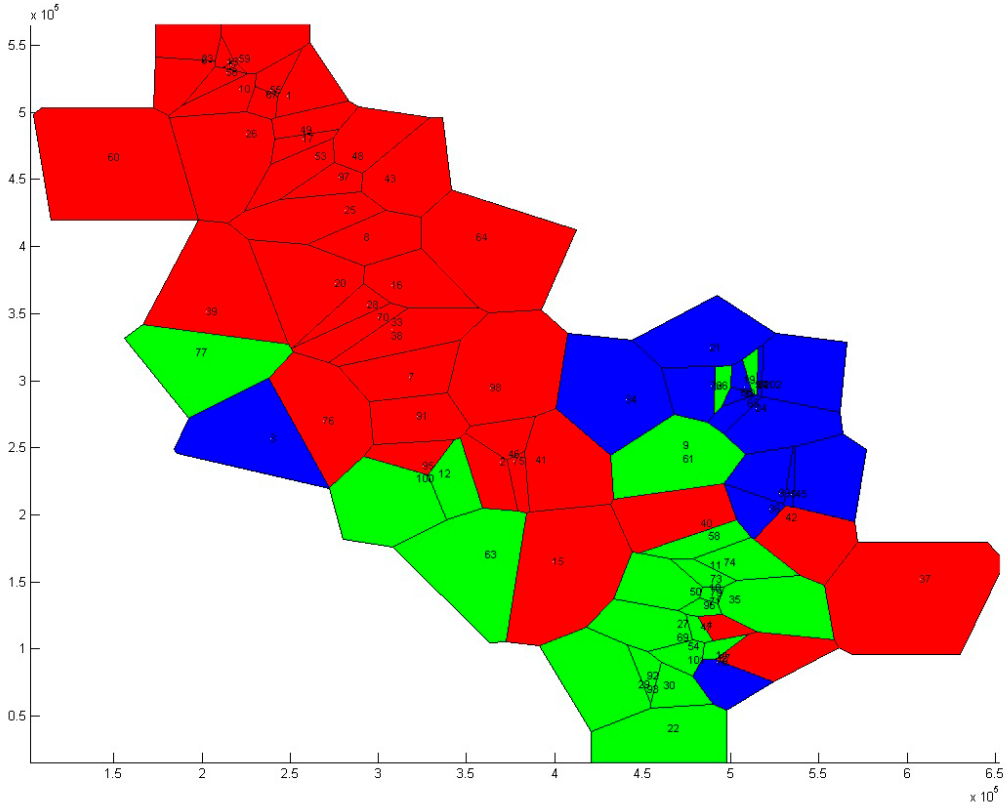
Käytännössä mikäli kappaleita on enemmän kuin 20, ositusten määrä kasvaa niin suureksi, ettei nimittäjässä olevan summan laskeminen ole enää mahdollista. Koska summa on sama kaikille osituksille, voidaan kuitenkin etsiä todennäköisyyden (3) maksimoiva ositus S vertaamalla vain nimittäjässä olevia termejä. Väitöskirjan osajulkaisuissa, joissa tätä menetelmää on sovellettu, havaitun geneettisen vaihtelun parhaiten selittävää ositusta on etsitty heuristisilla algoritmeilla.

Jotta todennäköisyydelle (3) voisi antaa Bayseläisen paradigman mukainen tulkinta subjektiivisen uskomuksen asteena aineiston havaitsemisen jälkeen, täytyisi myös priorijakauma $p(S)$ kuvata uskomuksen astetta ennen aineistoa. Ryhmittelyistä olevien ennakkokäsitysten, tai niiden puutteen, muotoileminen priorijakaumaksi on kuitenkin vaikeaa, sillä ositusten määrä kasvaa nopeammin kuin eksponentiaalisesti kappaleiden määrän suhteen ja eri ositusten väliset suhteet ovat monimutkaisia. Tasainen jakauma yli kaikkien mahdollisten luokkien on usein standardivalinta priorijakaumaksi, ja se onkin perusteltavissa, mikäli etsitään vain yksittäistä ainestoon parhaiten sopivaa ryhmittelyä.

Osajulkaisussa Corander, Sirén and Arjas (2008) esiteltiin vaihtoehtoinen tapa määrittää populaatiorakenteen priorijakauma, mikäli havaintojen maantieteelliset sijainnit ovat tiedossa. Saman lajin yksilöt, jotka ovat maantieteellisesti lähellä toisiaan, ovat usein myös geneettisesti lähempänä toisiaan, kuin kaukana toisistaan olevat yksilöt. Näin ollen lähellä toisiaan havaittujen yksilöiden voidaan myös olettaa kuuluvan samaan populaatioon. Tätä havaintoa hyödynnettiin rakentamalla ykslöiden välille naapurustoverkosto niiden sijaintien perusteella. Populaatiorakenteen priorijakauma määritettiin suosimaan osituksia, joissa vierekkäin olevat yksilöt kuuluvat samaan populaatioon. Kuvassa 1 on pohjoisamerikkalaisille ahmoille estimoitu populaatiorakenne, jossa on hyödynnetty yksilöiden maantieteellisiä sijainteja.

Osituksien priorijakaumaan voidaan myös lisätä rajoitteita sille, minkälaisia ryhmittelyjä sallitaan. Yksinkertaisimmillaan tämä voi olla yläraja ryhmien määrälle K, mikä on perusteltua, koska useimmiten voidaan olettaa ryhmien määrän olevan paljon pienempi kuin ryhmiteltävien kappaleiden määrä. Äärimmilleen vietynä rajoite voi määrittää osalle kappaleista ryhmän ja loppujenkin osalta sen, että ne tulevat jostain näistä ryhmistä. Tässä tapauksessa kyse on luokittelusta, johon tämänkaltainen määrittely antaa uudenlaisen lähestymistavan. Osajulkaisussa Corander et al. (2013) onkin tutkittu ryhmittelyn ja luokittelun suhdetta tästä näkökulmasta, ja vertailtu erilaisten valintojen tehokkuutta luokittelussa.

Väitöskirjassa kehitetyt menetelmät populaatiorakenteen oppimiseen on liitetty osaksi Bayesian Analysis of Population Structure (BAPS) -ohjelmaa, jonka tarjoamia analyysimahdollisuuksia on esitelty osajulkaisussa Corander, Sirén and Arjas (2008). Menetelmien



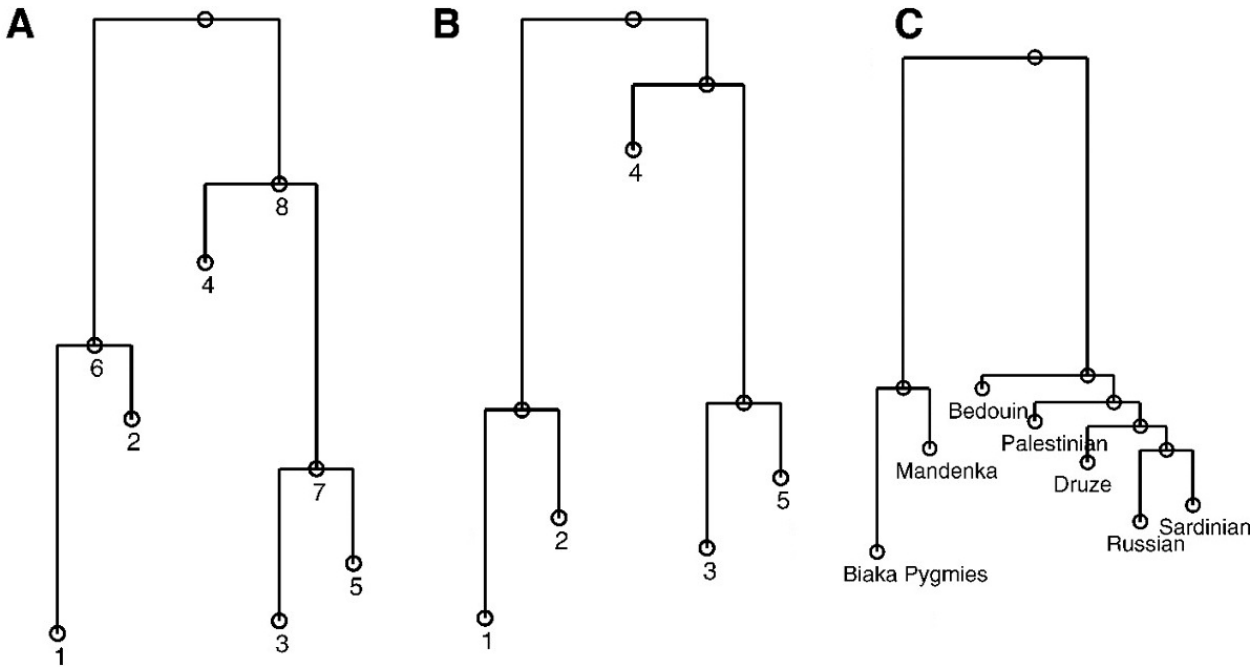
Kuva 1: Ahmojen geneettinen populaatiorakenne yksilöiden maantieteellisiä koordinaatteja hyödyntäen. Numerot ova yksilöitä ja niiden ympärillä olevan polygonin väri kertoo, mihin populaatioon yksilö kuuluu.

julkaiseminen osana vapaasti saatavaa ohjelmistopakettia on merkittävästi lisännyt niiden käyttöä ja auttanut biologeja käytännön populaatiogenetiikan tutkimusongelmissa.

5 Populaatioiden geneettinen historia

Populaatioiden tämän hetkisen geneettisen rakenteen lisäksi on usein kiinostuksen kohteena ymmärtää, kuinka tämä geneettinen rakenne on muotoutunut. Yksinkertaistettuna tässä tilanteessa halutaan oppia puu, jonka kärkinä on tämän hetkiset populaatiot ja juuressa hypoteettinen historiallinen populaatio ja joka selittää kuinka populaatiot ovat erkaantuneet toisistaan. Kuvassa 2 on esimerkkejä populaatioiden historiaa kuvastavista puista.

Populaatioiden väliset geneettiset erot merkkigeenien osalta ovat suurimmaksi ovat seurausta ajan yli vähitellen tapahtuvasta alleelifrekvenssien muutoksesta. Tämä erottaakin populaatioiden historiaa estimoivat menetelmät perinteisistä fylogeneettisistä menetelmistä, jotka mallintavat lajien välisiä geneettisiä eroja. Lajitasolla aikaskaala on pidempi, minkä takia mutaatiot selittävät suurimman osan vaihtelusta, kun taas popu-



Kuva 2: Esimerkkejä populaation historiaa kuvaavista puista. **A** Puu, jota käytettiin datan simuloinnissa. **B** Simuloidusta datasta 320 SNP:n perusteella estimoitu puu. **C** 7 ihmispopulaation historia 200 SNP:n perusteella. Julkaisusta Sirén, Marttinen and Corander (2011).

laatiotasolla mutaatioiden merkitys on pienempi.

Populaatiogenetiikassa laajalti käytetty malli alleelifrekvenssien muutokselle on Wright-Fisher -malli , jonka R.A. Fisher ja Sewall Wright kehittivät 1920- ja 1930-luvuilla. Yksinkertaisimmassa Wright-Fisher -mallissa seurataan alleelien määrän muutosta kiinteäkokoisessa populaatiossa ajan yli. Merkitään X_t:llä alleelin lukumäärää sukupolvessa t ja N populaatiokokoa (alleelien kokonaismäärä). Tällöin seuraavan sukupolven $t+1$ alleelien määrälle saadaan binomijakauma

$$p(X_{t+1} \mid X_t) \sim Bin\left(N, \frac{X_t}{N}\right). \tag{4}$$

Malli on yksinkertainen, mutta sen on havaittu selittävän hyvin luonnossa havaittua geneettistä vaihtelua.

Vaikka yhden sukupolven yli tapahtuvalle muutokselle on yksinkertainen esitys (4), mallin dynamiikka on kuitenkin monimutkainen pidemmällä aikavälillä. Populaatioiden historiaan sovellettaessa pitäisi pystyä laskemaan jakauma alleelien lukumäärälle sukupolvessa $t+s$ ehdolla X_t, missä s voi olla kymmeniä, satoja tai vaikka tuhansia, mutta sen laskeminen analyyttisesti on liian monimutkaista.

Vaihtoehtoinen lähestymistapa saadaan tarkastelemalla muutosta skaalatulla aikaasteikolla käyttäen sukupolvien sijasta populaatiokoolla jaettua aikaa $\tau = \frac{t}{N}$ ja mallintamalla alleelifrekvenssia $\theta_\tau = \frac{X_\tau}{N}$ kokonaismäärän sijaan. Mikäli tällöin annetaan populaatiokoon kasvaa rajatta saadaan mallille niin sanottu diffuusioapproksimaatio. Tämä

käytännössä tarkoittaa sitä, että lyhyen ajan yli allelifrekvenssien muutos seuraa normaalijakaumaa

$$p(\theta_{\tau+\epsilon} \mid \theta_\tau) \sim N\left(\theta_\tau, \epsilon\theta_\tau(1-\theta_\tau)\right),$$

missä $\epsilon \to 0$. Normaalijakaumaa onkin käytetty alleelifrekvenssien muutoksen mallintamiseen 1960-luvulta lähtien, vaikka sen tarkkuus heikkenee pidemmillä aikaväleillä ja lähellä rajoja (0 ja 1). Se tarjoaa kuitenkin erittäin nopean tavan estimoida suurenkin populaatiojoukon historia, minkä takia se on edelleen laajalti sovellettu menetelmä.

Normaalijakaumaa tarkempi approksimaatio diffuusiolle saadaan käyttämällä Beta-jakaumaa, jonka parametrit on valittu siten, että odotusarvo ja varianssi ovat samat kuin normaalijakaumalla. Osajulkaisussa Sirén, Marttinen and Corander (2011) esittelimme menetelmän, jossa alleelifrekvenssien muutos kussakin puun oksassa ylhäältä alas mallinnettiin Beta-jakaumilla. Alleelifrekvenssien muutos mallinnettiin jokaiselle frekvenssille erikseen, jolloin tuntemattomia parametreja oli mallissa ongelman koosta riippuen satoja tai tuhansia. Mallia käytettiin ihmispopulaatioiden historian oppimiseen SNP-aineistojen perusteella ja siitä saadut tulokset olivat hyvin linjassa kirjallisuudessa esitettyjen arvioiden kanssa. Kuvassa 2 on esitetty menetelmällä estimoituja puita simuloidusta ja oikeasta geneettisestä aineistosta.

Muille merkkigeeneille kuin SNP:ille, joissa havaitaan käytännössä vain kahta alleelia, yksinkertaisin Wright-Fisher -malli ei enää riitä kuvaamaan alleelifrekvenssien muutosta. Esimerkiksi mikrosatelliittien kohdalla yhdessä merkkigeenissä voi olla kymmeniä eri alleeleja ja lisäksi alleelien väliset mutaatiot ovat yleisiä, eikä niitä voida ohittaa, kuten SNP:ien kohdalla.

Väitöskirjassa kehitin tapoja approksimoida myös näitä monimutkaisempia Wright-Fisher -malleja, johtamalla analyyttisesti pitkän aikavälin muutoksen kaksi ensimmäistä momenttia. Valitsemalla sopivan jakauman, jonka momentit asetetaan samoiksi kuin analyyttisesti lasketut, pystytään approksimoimaan hyvin joitain Wright-Fisher -malleja. Osajulkaisussa Sirén, Hanage and Corander (2013) esiteltiin menetelmä, jossa tätä sovellettiin bakteerien populaatiogenetiikkaan.

6 Yhteenveto

Väitöskirjassa on kehitetty populaatiogenetiikan ongelmiin useita eri tilastollisia menetelmiä, joiden yhteisenä piirteenä on taustalla olevien biologisten ilmiöiden mallinnus ja Bayesläinen lähestymistapa tilastolliseen päättelyyn. Osa menetelmistä on suunniteltu sellaisiksi, että muut tutkijat voivat niitä työssään kohtaamissa ongelmissa käyttää, ja julkaistu tästä syystä osana vapaasti saatavaa ohjelmistopakettia. Toiset ovat sen sijaan kehitetty enemmän osoittaakseen, minkälaista analyysiä on mahdollista tehdä, kuin laajamittaiseen yleiseen käyttöön. Esimerkiksi artikkelin (Sirén, Marttinen and Corander, 2011) julkaisun jälkeen on kirjallisuudessa esitelty uusia menetelmiä, jotka parantavat tai tarjoavat vaihtoehdon siinä esitellylle mallille.

Viitteet

Corander J, Cui Y, Koski T, Sirén J. 2013. Have I seen you before? Principles of Bayesian predictive classification revisited. Statistics and Computing. pp. 1–15. 10.1007/s11222-011-9291-7.

8

Corander J, Marttinen P, Sirén J, Tang J. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics. 9:539.

Corander J, Sirén J, Arjas E. 2008. Bayesian spatial modeling of genetic population structure. Comput. Stat. 23:111–129.

Novembre J, Johnson T, Bryc K, et al. (11 co-authors). 2008. Genes mirror geography within Europe. Nature. 456:98–101.

Savage LJ. 1976. On rereading R. A. Fisher. The Annals of Statistics. 4:441–500.

Sirén J. 2012. Statistical models for inferring the structure and history of populations from genetic data. Ph.D. thesis, University of Helsinki.

Sirén J, Hanage WP, Corander J. 2013. Inference on population histories by approximating infinite alleles diffusion. Molecular Biology and Evolution. 30:457–468.

Sirén J, Marttinen P, Corander J. 2011. Reconstructing population histories from single nucleotide polymorphism data. Molecular Biology and Evolution. 28:673–683.

Waples RS, Gaggiotti O. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Molecular Ecology. 15:1419–1439.

ILTAPÄIVÄSEMINAARI 23.4.2013

Helsingin kaupunki Tietokeskus

SUOMEN TILASTOSEURA

Aihe: Mitä tilastot ja rekisterit kertovat tämän päivän nuorista?
Aika: 23.4.2013 klo 14–17
Paikka: Helsingin taloushallintopalvelun Saldo-auditorio, Sörnäisten rantatie 27 A
Järjestäjät: Helsingin kaupungin Tietokeskus ja Suomen Tilastoseura

Ohjelma:

Tilaisuuden avajaissanat, Asta Manninen, johtaja, Helsingin kaupungin tietokeskus

Suomen tilastoseuran tervehdys, Kimmo Vehkalahti, Suomen tilastoseura

Nuorten hyvinvointi kansallisen syntymäkohortti 1987 -aineiston valossa, Reija Paananen, erikoistutkija, THL

Nuoret toimeentulotuen saajat, Elise Haapamäki, tutkija, Helsingin kaupungin tietokeskus

Ketä nuoret työttömät ovat ja mitä eroja nuorten työttömyydessä on eri EU-maissa?, Liisa Larja, yliaktuaari, Tilastokeskus

Mitä tilastot kertovat nuorisotakuun piiriin kuuluvista nuorista Helsingissä?, Seija Saari, projektitutkija, Helsingin kaupungin tietokeskus

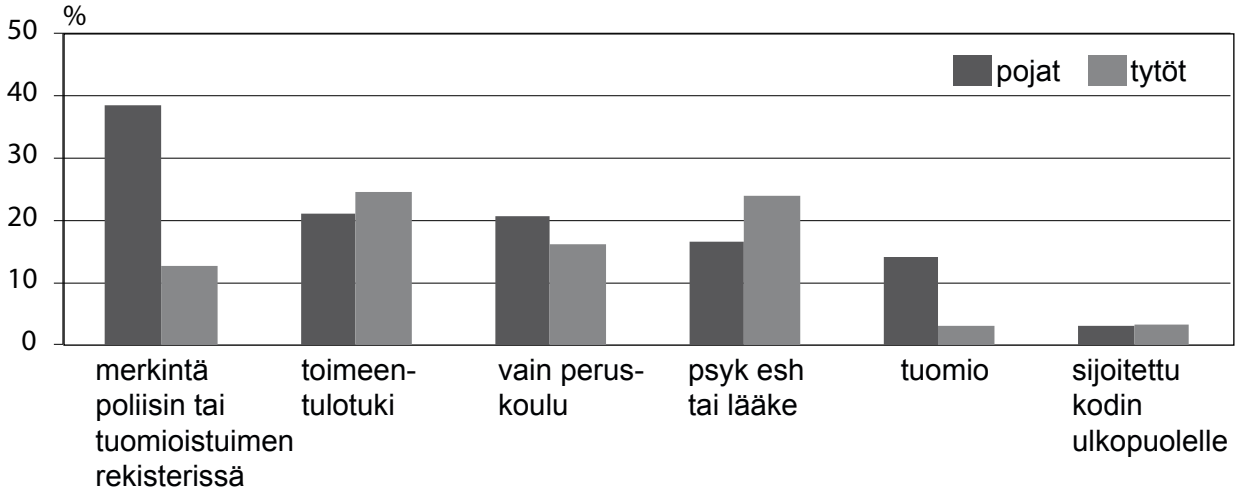
Mitä nuorten koululaisten arkeen kuuluu? Nuoret Helsingissä 2011 -tutkimuksen tuloksia, Vesa Keskinen, tutkija, Helsingin kaupungin tietokeskus

Kansallinen syntymäkohortti 1987

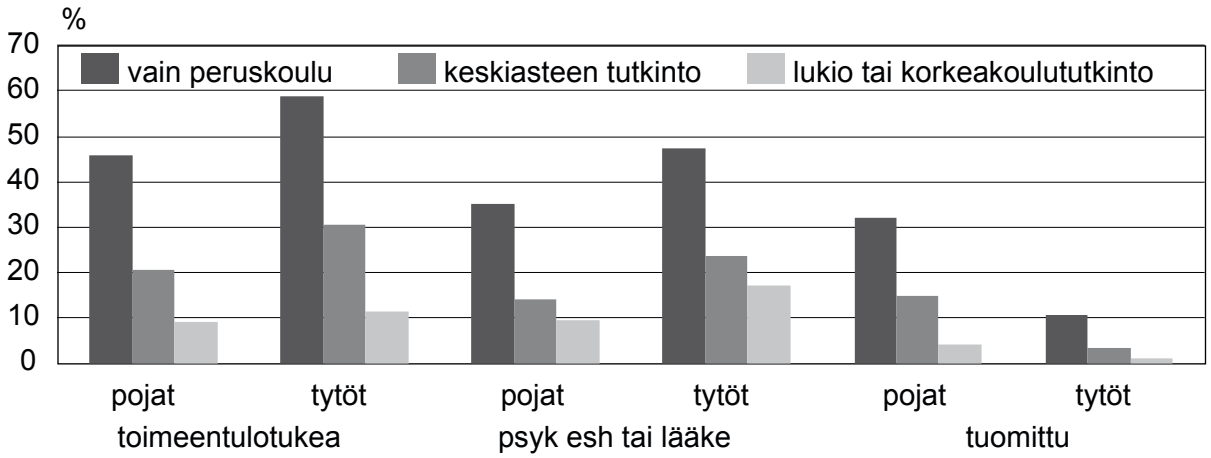
Paananen Reija & Gissler Mika
THL

Vuonna 1987 Suomessa syntyneiden pitkittäisseurannan perusteella voidaan todeta, että Suomessa suurin osa nuorista aikuisista voi hyvin. Huomattavaa on kuitenkin, että on merkittävä joukko lapsia ja nuoria, joille on 21 vuoden ikään mennessä rekisteröity erilaisia hyvinvoinnin ongelmia (kuva 1.). Viidesosa ikäluokasta on käyttänyt psykiatrian erikoissairaanhoidon palveluita tai psyykenlääkkeitä ennen aikuisikää, eikä joka viidennellä ole toisen asteen koulutusta. Toimeentulo-ongelmat sekä rikollisuus koskettavat noin joka neljättä 1987 syntynyttä. Tutkimuksemme mukaan huomattava osa mielenterveyden häiriöistä alkaa jo lapsena ja nuorena, jolloin ne vaikuttavat koulussa pärjäämiseen ja nuoren myöhempään hyvinvointiin.

Tutkimuksen perusteella käy selväksi, että hyvinvointi eriytyy ja hyvinvoinnin ongelmat, kuten kouluttamattomuus, mielenterveys- sekä toimeentulo-ongelmat kasaantuvat. Niillä nuorilla, joilla ei ole peruskoulun jälkeistä jatkotutkintoa seuranta-ajan päättyessä, on huomattavasti yleisemmin myös psykiatrisen erikoissairaanhoidon tai psyykenlääkkeiden käyttöä, toimeentulo-ongelmia ja rikollisuutta (kuva 2.).



Kuva 1. Keskeiset hyvinvoinnin osoittimet vuonna 1987 syntyneillä tytöillä ja pojilla vuoteen 2008 mennessä.



Kuva 2. Toimeentulotukea saaneiden, psyykenlääkkeitä tai psykiatrisen erikoissairaanhoidon palveluita käyttäneiden sekä rikoksista tuomittujen osuudet vuonna 1987 syntyneiden ikäluokasta sukupuolen ja koulutusasteen mukaan vuoteen 2008 mennessä.

Ongelmien ylisukupolvisuus

Kansallinen syntymäkohortti 1987 -tutkimuksen tulokset kertovat myös ongelmien periytymisestä sukupolvelta toiselle, ylisukupolvisuudesta. Lapsuuden olosuhteilla on huomattava vaikutus myöhempään hyvinvointiin. Vanhemman kuolema, vakava sairastuminen tai mielenterveyden ongelmat ovat kiinteässä yhteydessä lasten myöhempiin hyvinvoinnin ja mielenterveyden ongelmiin; vanhempien työttömyys sekä taloudelliset ja terveydelliset vaikeudet lisäävät heidän lastensa riskiä koulunkäynnin ja mielenterveyden ongelmille sekä riskiä tulla sijoitetuksi kodin ulkopuolelle. Lapsuuden perhetekijät, elinolot ja kehitysympäristöt ovat oleellisia yhteiskuntaan kiinnittymisessä. Aiempien tutkimusten perusteella tiedetään, että varhaiset elinolot vaikuttavat esimerkiksi kouluttautumiseen ja tätä kautta kokonaisvaltaisesti myöhempään hyvinvointiin. Tutkimusten mukaan kehitysympäristöjen ongelmien vaikutus on sitä suurempi mitä varhaisemmassa vaiheessa niitä esiintyy.

Tutkimuksen tulokset kertovat myös vanhempien mielenterveyden ongelmien vaikutuksista lasten kehitykseen. Toimeentulovaikeuksien ohella lasten ongelmien taustalla on usein vanhempien mielenterveysongelmia ja perhesuhteiden muutoksia. Kolmasosalla psykiatrisen diagnoosin saaneista lapsista ja nuorista on psykiatrisessa erikoissairaanhoidossa hoidettu vanhempi, kun kaikkien vuonna 1987 syntyneiden vanhemmista noin viidesosa on käyttänyt psykiatrisen erikoissairaanhoidon palveluita. Noin 62 prosenttia psykiatrisen diagnoosin saaneista lapsista ja nuorista on joko yksinhuoltajaperheestä tai on kokenut vanhempien avioeron tai vanhemman kuoleman, kun nämä muuten koskettavat 45 prosenttia koko ikäluokasta. Näin siis myös vanhempien ongelmat kasautuvat ja jättävät jälkensä lasten hyvinvointiin.

Kansallinen syntymäkohortti 1987 -tutkimusaineisto

Tutkimuksessa on seurattu kaikkia Suomessa 1987 syntyneitä lapsia sikiökaudelta 31.12.2008 saakka. Aineistoon kuuluu kaikkiaan 60 069 Suomessa syntynyttä lasta. Tässä tutkimuksessa 73 lasta ei voitu jäljittää epätäydellisen, puuttuvan tai väärän henkilötunnuksen vuoksi. Seurannassa on siis mukana 59 996 Suomessa vuonna 1987 syntynyttä lasta. Kansallinen syntymäkohortti 1987 sisältämät tiedot, sekä rekisterit, joista tiedot on kerätty, on kuvattu yksityiskohtaisesti aiemmin (Paananen ym. 2012). Rekisterien avulla tehty pitkäaikaisseuranta tarjoaa monia mahdollisuuksia hyvinvoinnin seurantaan ja hyvinvoinnin kehittymisen tutkimukseen. Kansallinen syntymäkohortti 1987 on ensimmäinen pitkittäisseuranta, joka kattaa kaikki Suomessa tiettynä vuonna syntyneet lapset. Suomalainen rekisteritieto on tutkimuksellisessakin mielessä huipputasoa, sekä laadultaan että määrältään. Valtakunnallisen rekisteriseurannan ehdoton vahvuus on aineiston kattavuus, sekä tutkimusjoukon, käytettyjen muuttujien kuin alueellistenkin tekijöiden osalta.

Johtopäätökset

Tutkimuksen tulokset eivät tue ajatusta ennalta määrätyistä tai geenien kautta siirtyvistä ylisukupolvisista ilmiöistä, vaan ennemminkin ajatusta siitä, että kehityksellä on tietty suunta, jota arki vahvistaa tai heikentää. Hyvinvointiongelmat ilmenevät usein vasta pitkän ajan kuluttua. Näin ollen hyvinvoinnin tukeminen ja ehkäisevä työ tulisikin aloittaa varhain, jo ennen ongelmien ilmaantumista, ja varhaislapsuudessa ja lapsuudessa ilmeneviin signaaleihin tulisi puuttua ennen vaikeiden oireiden ilmestymistä. Ehkäisevien ja hyvinvointia tukevien palveluiden merkitys korostuu ylisukupolvisen ongelmaketjun katkaisemisessa ja syrjäytymisen ehkäisemisessä. Koska hyvinvointi rakentuu arjessa, on aivan oleellista, että eri kasvuympäristöt kuten päivähoito, koulu ja harrastukset toimivat lasten ja nuorten hyvinvointia tukien ja suojaavia tekijöitä vahvistaen. Lisäksi peruspalveluihin ja toisaalta esimerkiksi psykiatrisiin avopalveluihin kannattaa panostaa, sillä ehkäisevä työ sekä varhainen ongelmiin puuttuminen ovat ensiarvoisen tärkeitä niin inhimilliseltä kuin taloudelliseltakin kannalta.

Lähteet

Paananen R, Ristikari T, Merikukka M, Rämö A & Gissler M. 2012. Lasten ja nuorten hyvinvointi Kansallinen syntymäkohortti 1987–tutkimusaineiston valossa. Raportti 52/2012, Terveyden ja hyvinvoinnin laitos.

Nuoret toimeentulotuen saajina Helsingissä

Tutkija Elise Haapamäki
Helsingin kaupungin tietokeskus

Tilastoseuran iltapäiväseminaarin esitys Nuoret toimeentulotuen saajina Helsingissä perustuu Helsingin kaupungin tietokeskuksen tilastoja sarjan julkaisuun Nuoret toimeentulotuen saajat – Pitkittäistarkastelu 18–20-vuotiaista helsinkiläisistä toimeentulotuen saajista vuosina 2006–2011.

Esityksessä tarkastellaan 18–20-vuotiaita helsinkiläisiä toimeentulotuen saajia vuosina 2006, 2009 ja 2011. Tavoitteena oli selvittää, keitä Helsingissä toimeentulotukea saaneet nuoret ovat ja miltä heidän toimeentulotuen polkunsa näyttäytyy. Aineistona toimi Helsingin sosiaaliviraston (nykyisin Sosiaali- ja terveysvirasto) toimeentulotuen asiakasrekisteristä muodostettu aineisto vuosilta 2006–2011.

Tutkimusasetelma oli aineistolähtöinen, eli 18–20-vuotiaita nuoria tarkasteltiin hieman eri lähtökohdista kolmena tarkastelussa olevina vuosina. Vuonna 2006 18–20-vuotiaiden helsinkiläisten toimeentulotukea saaneiden osalta katsottiin asiakkuuden mahdollista jatkuvuutta tai toistuvuutta vuosina aina vuoteen 2011 saakka. 2011 toimeentulotukea saaneista 18–19-vuotiaista nuorista taas pystyttiin palaamaan taaksepäin vuoteen 2006, ja tutkimaan ovatko nuoren vanhemmat olleet myös helsinkiläisinä toimeentulotuen saajina. Vuoden 2009 18–20-vuotiaiden tuensaajien kohdalla nuorten tukipolkua pystyttiin katsomaan muutaman vuoden kumpaakin suuntaan.

Moni toimeentulotukea saavista nuorista jää tuen piiriin useammaksi vuodeksi. Vuonna 2006 toimeentulotukea Helsingissä saaneista 18–20-vuotiaista nuorista 40 prosenttia yhä tuen piirissä 23–25-vuotiaina vuonna 2011. Neljännes kaikista vuonna 2006 tukea saaneista nuorista oli toimeentulotuen piirissä kaikkina kuutena tarkastelujakson vuotena (2006–2011). Nuoren toimeentulotuentarve oli useimmiten jatkuvaa tai toistuvaa. Ainoastaan 20 prosenttia vuonna 2006 toimeentulotukea saaneista 18–20-vuotiaista nuorista ei itse saanut tukea tai asunut tukea saavassa kotitaloudessa yhtenäkään viitenä seuraavana vuotena.

Suurin osa toimeentulotukea saaneista nuorista asui yksin. Yksinasuvista nuorista miehistä, joiden toimeentulotuen saanti jatkuu useamman vuoden ajan, lähes kaikki jatkoivat asumista yksin. Nuorten naisten osalta tilanne muuttui jonkin verran, ja tuen jatkuessa yksinhuoltajien osuus lisääntyi. Toisaalta yksinasuvat naiset jäivät miehiä harvemmin

toimeentulotuen piiriin useammaksi vuodeksi. Työttömänä olleilla nuorilla tuensaajilla oli keskimääräisistä korkeampi riski jäädä toimeentulotuen piiriin pidemmäksi aikaa. Opiskelijoilla ja koululaisilla taas riski oli hieman keskimääräistä alhaisempi.

Iso osa nuorista tuensaajista oli aikaisemmin asunut toimeentulotukea saaneessa kotitaloudessa. Vuonna 2011 18-vuotiaista tuensaajista yli puolet oli ollut osallisena toimeentulotukeen vähintään yhtenä vuotena viiden edellisen vuoden aikana. Vastaavasti vuotta vanhemmista, 19-vuotiaista, aikaisempi osallisuus löytyi lähes joka toiselta. Nuorilla, jotka asuivat aikaisemmin toimeentulotukea saaneessa kotitaloudessa, tuensaanti jatkui useammin kuin niillä, joilla aikaisempaa osallisuutta ei löytynyt.

Lisätietoja:
Helsingin kaupungin tietokeskuksen tilastoja 2013:12: Nuoret toimeentulotuen saajat – Pitkittäistarkastelu 18–20-vuotiaista helsinkiläisistä toimeentulotuen saajista vuosina 2006–2011 (luettavissa sähköisesti www.hel.fi/tietokeskus =>julkaisut)

Ketä nuoret työttömät ovat ja mitä eroja nuorten työttömyydessä on eri EU-maissa

Liisa Larja
(liisa.larja@stat.fi)

Kirjoitus perustuu artikkeliin *Hyvinvointikatsauksessa 1/2013* ja esitykseen Tilastoseuran seminaarissa "Mitä tilastot ja rekisterit kertovat tämän päivän nuorista", 23.4.2013

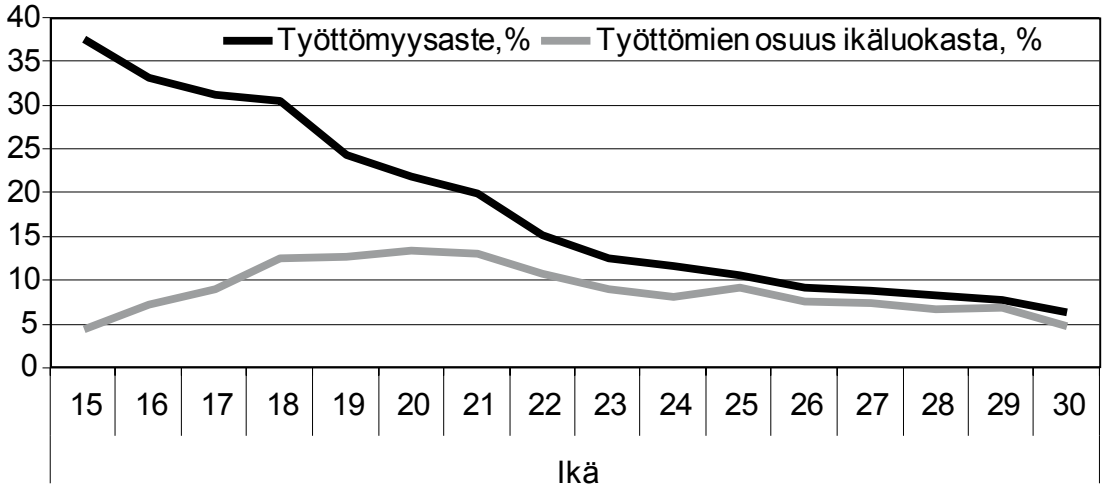
Nuorisotyöttömyydestä puhutaan huolestuneeseen sävyyn, ja usein nuorten työttömyys ja syrjäytyminen samastetaan. Työttömyysaste on kuitenkin varsin huono tapa kuvata nuorten elinoloja – puhumattakaan syrjäytymisestä. Yhdessä maassa nuorista työttömistä 70 prosenttia on opiskelijoita, mutta toisessa maassa vain 3 prosenttia. Sen vuoksi on syytä tarkastella sitä, miten eri maiden nuoret työttömät eroavat toisistaan.

Työttömyysasteen pohjalta tehdään usein liian yksioikoisia päätelmiä

Aika ajoin nuorten 20 prosentin työttömyysasteen tulkitaan tarkoittavan sitä, että joka viides nuori on työtön. Tällainen tulkinta on kuitenkin väärinkäsitys. Kyse ei ole koko ikäluokkaa koskevasta osuudesta, vaan työttömyysaste on työttömien osuus työvoimasta eli työllisten ja työttömien yhteismäärästä. Suurin osa nuorista opiskelee eikä kuulu työvoimaan, joten jakolaskun nimittäjässä on varsin pieni joukko nuoria. Kun 15–24-vuotiaiden työttömyysaste oli 20 prosenttia vuonna 2011, se tarkoitti noin 10 prosenttia koko ikäluokasta eli 66 000 henkilöä.

Nuorisotyöttömyyttä tarkasteltaessa on tärkeää ottaa huomioon ikärajaus: työttömyysaste on korkea nuorimpien ja matala vanhempien joukossa (kuvio 1). Suomessa 15–19-vuotiaiden työttömyysaste on 31 prosenttia, mutta 20–24-vuotiaiden 16 prosenttia. Yli 24-vuotiaan väestön työttömyysaste on enää vajaa 7 prosenttia.

Kuvio 1. Työttömyysaste ja työttömien osuus ikäluokasta Suomessa iän mukaan vuonna 2011. Vuosikeskiarvo.

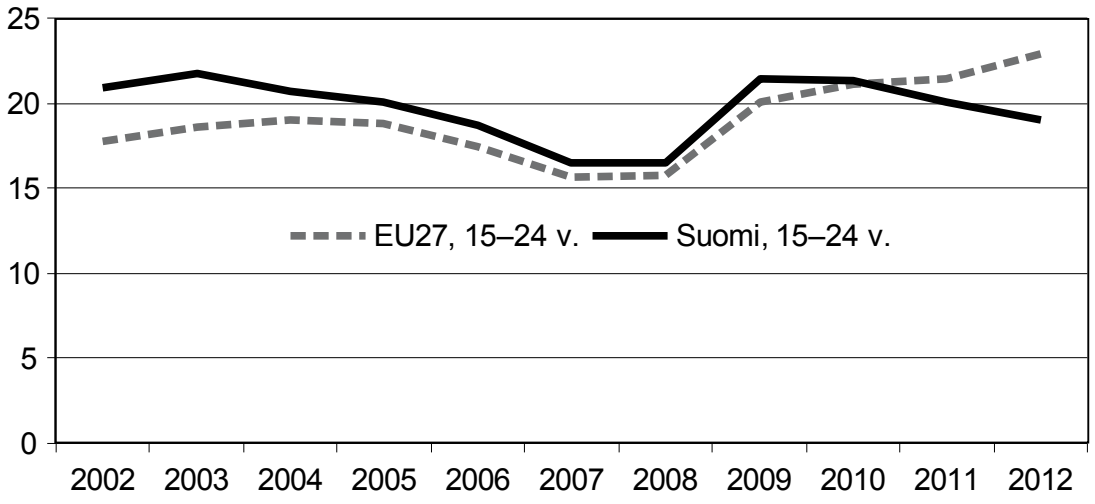


Lähde: Tilastokeskus 2013.

Suomessa talouskriisistä toipuminen on jo alkanut

Suomessa nuorten työttömyysaste pysytteli aina vuoteen 2010 asti hieman EU:n keskiarvon yläpuolella. Kun muissa EU-maissa talouskriisin vaikutukset levisivät edelleen, vuonna 2011 Suomessa tilanne kääntyi jo parempaan suuntaan ja nuorten työttömyysaste laski EU-keskiarvon alapuolelle (kuvio 2). Vuonna 2012 nuorisotyöttömyys lisääntyi EU-maissa (23 %), mutta vähentyi Suomessa (19 %).

Kuvio 2. 15–24-vuotiaiden työttömyysaste Suomessa ja EU:ssa vuosina 2002–2012. Vuosikeskiarvo.*



* Varusmiespalveluksessa olevat eivät ole mukana.
Lähde: Eurostat 2013.

Yli puolet Suomen nuorista työttömistä on opiskelijoita

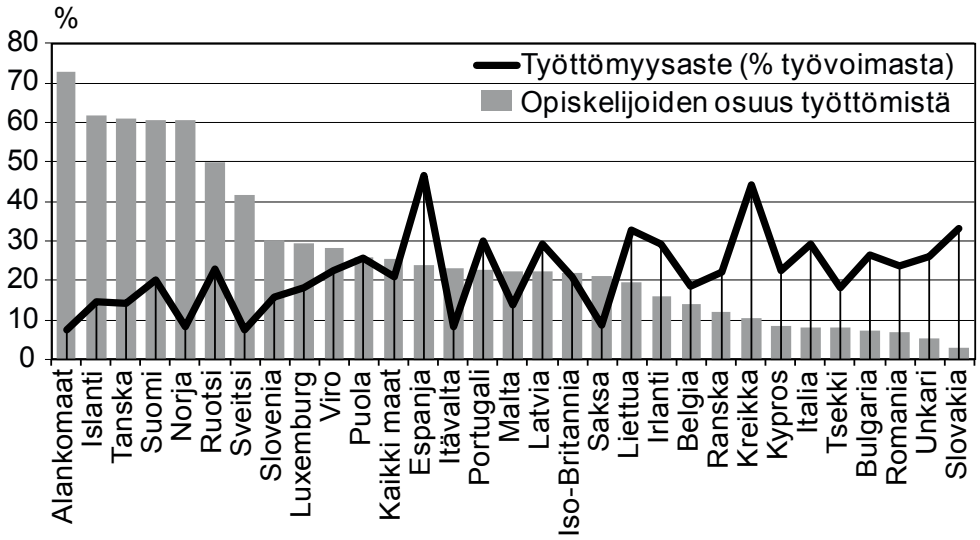
Maidenvälisissä vertailuissa on syytä analysoida hieman tarkemmin nuorisotyöttömyyden rakennetta. Yksi merkittävimpiä eroja lienee opiskelijoiden ja vain osa-aikaista työtä etsivien osuus nuorista työttömistä. Kansainvälisessä tilastoinnissa myös opiskelijat ja koululaiset lasketaan työllisiksi, jos he tekevät opintojen ohella vaikka vain satunnaista osa-aikatyötä. Jos taas opiskelijat etsivät tällaista työtä, heidät luetaan työttömiksi.

Pohjoismaissa nuorisotyöttömyys eroaa muista maista siinä, että se on suurelta osin opiskelijoiden osa-aikatyön ja kesätyön etsimistä. Pohjoismaissa ja Alankomaissa työttömistä nuorista 50–70 prosenttia on opiskelijoita. Suomessa työttömiä opiskelijoita oli 61 prosenttia 15–24-vuotiaista työttömistä. (Kuvio 3.)

Muissa Euroopan maissa työnteko opiskelun ohella ei ole yhtä yleistä, ja opiskelijoiden osuus nuorista työttömistä on keskimäärin vain 25 prosenttia ja Slovakiassa vain 3 prosenttia (kuvio 3). Tämä on hyvä ottaa huomioon nuorisotyöttömyyslukuja vertaillessa.

Alankomaissa nuorisotyöttömyys on vähäistä kaikilla mittareilla: työttömyysaste on Euroopan alhaisin ja näistä työttömistäkin lähes kaikki ovat opiskelijoita. Seuraavaksi pienin työttömyysaste on Itävallassa ja Saksassa, mutta näissä maissa työttömistä on opiskelijoita vain hieman yli 20 prosenttia. Näin ollen opiskelemattomia työttömiä nuoria on Itävallassa, Saksassa ja Suomessa täsmälleen yhtä suuri osuus koko ikäluokasta eli 4 prosenttia, vaikka suomalaisnuorten varsinainen työttömyysaste on 10 prosenttiyksikköä korkeampi kuin Itävallassa ja Saksassa. – Suomalaisnuorten työttömyystilanne ei siis ole niin huono kuin 20 prosentin työttömyysasteesta saattaa päätellä.

Kuvio 3. 15–24-vuotiaiden työttömyysaste sekä työttömien opiskelijoiden osuus työttömistä vuonna 2011. Vuosikeskiarvo.

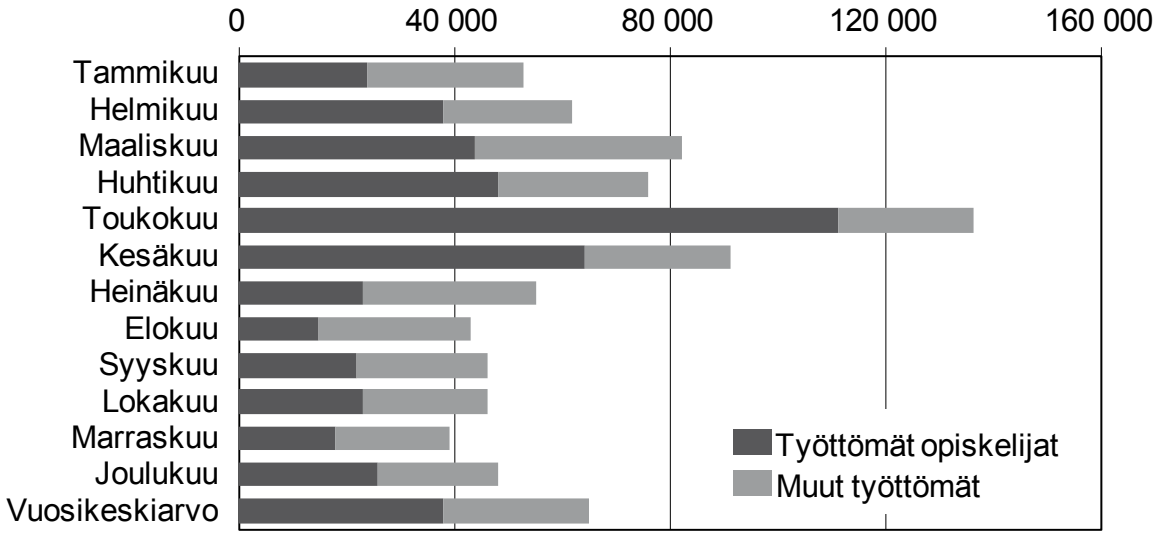


Lähde: Eurostat 2012.

Suomessa on vähiten nuoria pitkäaikaistyöttömiä

Myös työttömyyden pituuksien avulla voidaan kuvata työttömyyslukujen eroja. Suurimmalle osalle suomalaisnuorista työttömyys on väliaikainen vaihe opiskelun ja työelämän välillä. Tämän havaitsee jo nuorten työttömyyslukujen kuukausivaihtelusta: työttömien nuorten määrä kohoaa lähes 140 000:een toukokuussa, kun opiskelijat etsivät kesätöitä ja vastavalmistuneet ensimmäistä työpaikkaansa (kuvio 4). Vähiten (noin 40 000) työttömiä nuoria on syksyllä koulujen alettua. Nuorten työttömyys on yleensä huomattavasti lyhytkestoisempaa kuin muulla väestöllä (Hämäläinen& Hämäläinen 2012).

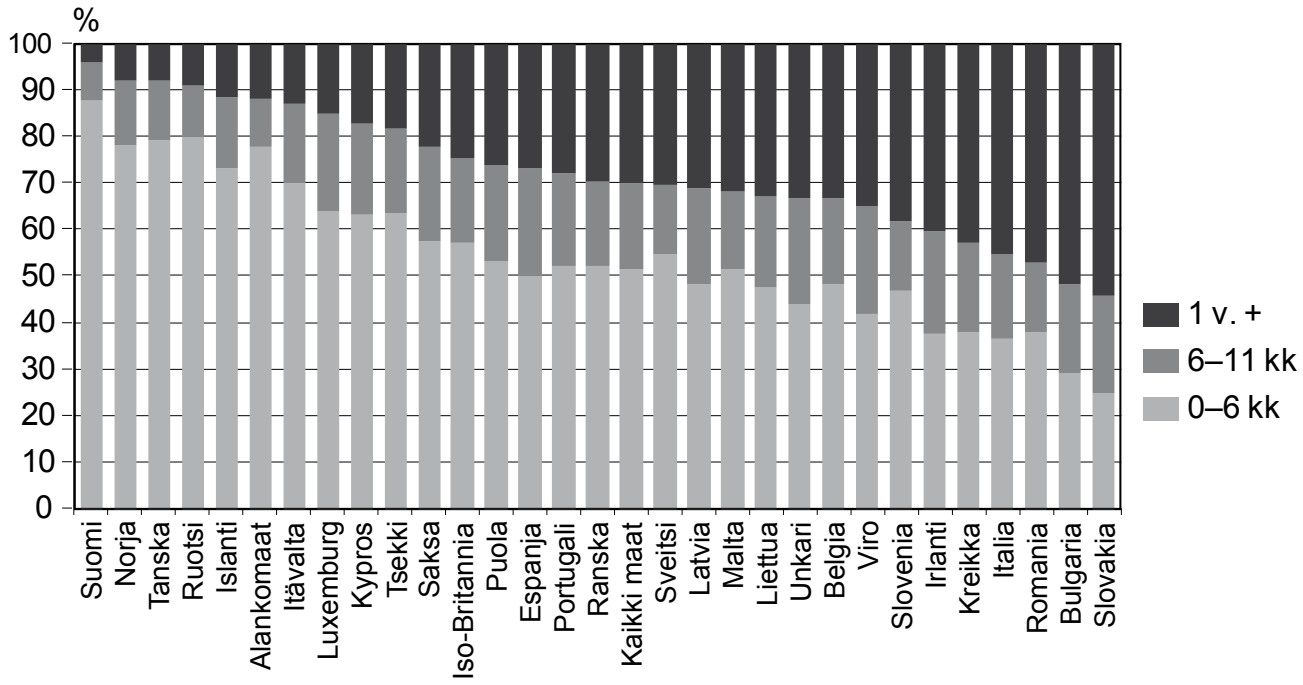
Kuvio 4. 15–24-vuotiaat työttömät opiskelijat ja muut 15–24-vuotiaat työttömät kuukausittain Suomessa vuonna 2011.



Lähde: Tilastokeskus 2013.

Nuorten pitkäaikaistyöttömyyttä on Suomessa vähemmän kuin missään muussa EU-maassa. Suomessa vain 4 prosenttia nuorista työttömistä on ollut työttömänä yhtäjaksoisesti yli vuoden, kun Slovakissa ja Bulgariassa pitkäaikaistyöttömien osuus on yli 50 prosenttia (kuvio 5). Suomalaisista nuorista työttömistä 88 prosenttia on ollut yhtäjaksoisesti työttömänä alle 6 kuukautta. Myös muissa Pohjoismaissa nuorten työttömyysjaksot ovat suhteellisen lyhyitä.

Kuvio 5. 15–24-vuotiaiden työttömyyden kesto eri maissa vuonna 2011. Vuosikeskiarvo.



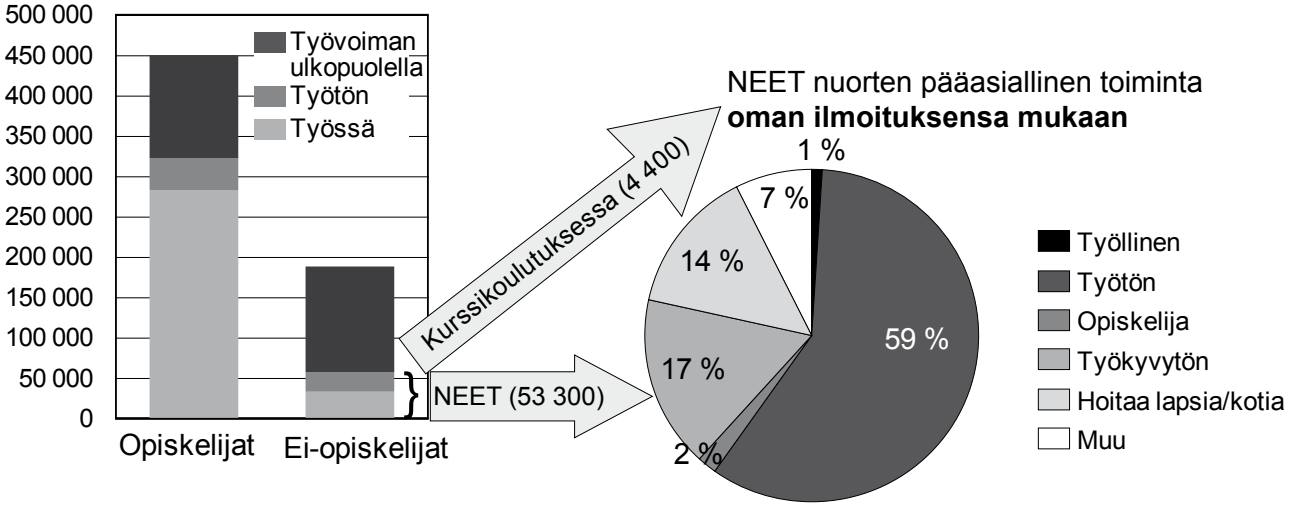
Lähde: Eurostat 2012.

Työttömyysaste ei kuvaa nuorten syrjäytymistä

Kuten edellä on käynyt ilmi, työttömyysaste ei ole järin hyvä indikaattori nuorten elinolojen kuvaajana. Se ei ota huomioon työvoiman ulkopuolella olevien suurta määrää ikäluokasta, eikä se myöskään tee eroa sen suhteen, onko työtä etsivä pääasialliselta toiminnaltaan opiskelija vai ei. Se ei myöskään kerro, onko kyse muutamien viikkojen kesätyön hausta vai vuosia kestävästä työttömyydestä. Näin ollen työttömyysaste ei kuvaa sitä, kuinka suuri osa nuorista on ”syrjäytyneitä” tai avun tarpeessa.

Jotta opiskelijoiden suuri osuus nuorista ikäluokista tulisi paremmin huomioon otetuksi, on nuorisotyöttömyyden sijaan alettu käyttää ns. NEET-astetta (Not in Employment, Education or Training). NEET tarkoittaa sellaisia nuoria, jotka eivät ole työssä, koulutuksessa tai kurssilla (esim. työväenopiston kurssia tai valmennuskurssia). Työvoimatutkimuksen mukaan vuonna 2011 Suomessa oli opiskelemattomia 15–24-vuotiaita nuoria työttömiä 4 prosenttia ikäluokasta (25 500) ja työvoiman ulkopuolella 5 prosenttia ikäluokasta (32 200) eli yhteensä 9 prosenttia ikäluokasta (57 700 nuorta) (kuvio 6).

Kuvio 6. 15–24-vuotiaat opiskelijat ja muut nuoret työmarkkina-aseman mukaan vuonna 2011.*



* Varusmiespalveluksessa olevat eivät ole mukana.

Lähde: Eurostat 2012

NEET-nuoria (tai ”ulkopuolisia”) pidetään usein ”syrjäytyneinä”, vaikka määritelmä ei kerro muuta kuin sen, että kyseiset nuoret eivät ole tutkimushetkellä työssä eivätkä opiskelemassa. NEET-määritelmä ei kerro, onko nuori masentunut, yksinäinen tai taloudellisissa vaikeuksissa (ks. myös Okkonen 2008). Osalle nuorista työn ja koulutuksen ulkopuolella olo voi olla kielteinen kokemus, osalle se voi olla lomaa pätkätöiden välillä, osa voi käyttää aikaansa järjestötoimintaan (Kojo 2012), odottaa opiskelun tai asepalveluksen alkamista, lukea pääsykokeisiin tai hoitaa omaa lastaan kotona.

Työvoimatutkimuksessa vastaajilta kysytään omaa käsitystä pääasiallisesta toiminnasta; tämä tieto on käyttökelpoinen erityisesti NEET-nuorten tilannetta tarkasteltaessa. Esimerkiksi yliopiston pääsykokeisiin valmentautuva nuori tuskin kokee itseään työttömäksi saatikka syrjäytyneeksi, vaan ehkä pikemminkin opiskelijaksi, vaikka hän rekisteritietojen mukaan olisikin sekä koulutuksen että työelämän ulkopuolella.

NEET-nuorista 59 prosenttia (31 400) ilmoitti pääasialliseksi toiminnakseen työttömyyden vuonna 2011, 17 prosenttia (8 900) kertoi olevansa työkyvytön, 14 prosenttia (7 500) hoiti kotona lapsiaan ja 7 prosenttia (4 000) sanoi tekevänsä jotain muuta (kuvio 9). Lisäksi joukossa oli joitakin henkilöitä, jotka kertoivat pääasialliseksi toiminnakseen työn tai opiskelun, vaikka heillä ei tutkimusviikolla ollutkaan työpaikkaa tai tutkintoon johtavaa opiskelupaikkaa. Näiden tietojen pohjalta on mahdotonta tietää, kuinka moni heistä on ”syrjäytynyt”.

Itsensä työttömäksi mieltävistä NEET-nuorista 30 prosenttia oli ollut työttömänä myös vuosi ennen haastattelua. Opiskelijana oli ollut noin 40 ja työssä 20 prosenttia. Työvoimatoimistoon kertoo rekisteröityneensä noin 80 prosenttia, mutta työttömyyskorvausta kertoi saavansa vain noin 50 prosenttia. Vajaa 60 prosenttia oli suorittanut lukion tai ammattikoulututkinnon, muilla oli peruskoulupohja.

Työkyvyttömiksi itsensä määritteleviä on enemmän kuin omia lapsiaan kotona hoitavia (kuvio 9). Näiden 8 900 nuoren tilannetta voi tulkita Kelan ja ETK:n tilastojen valossa. Vuonna 2011 Kelan vammaistukea sai 2 716 16–24-vuotiasta (Kela 2012). Samana vuonna ETK:n eläkkeensaajatilaston (ETK 2011) mukaan työkyvyttömyyseläkkeellä oli 6 428 16–24-vuotiasta nuorta.

Kelan sairaspäiväraha- ja eläketilastojen mukaan vuonna 2009 työkyvyttömyyseläkkeelle siirtyneistä alle 30-vuotiasta nuorista 75 prosenttia sai työkyvyttömyyseläkkeen mielenterveyden häiriön, erityisesti masennuksen, perusteella (Raitasalo & Maaniemi 2011). Masennuksen takia työkyvyttömyyseläkkeelle siirtyneitä oli vuonna 2009 enemmän kuin kertaakaan 2000-luvulla.

Kiinnostusta herättävät myös ne 4 000 nuorta, jotka kertoivat pääasialliseksi toiminnakseen ”jotakin muuta” kuin edellä mainitut työ, työttömyys, opiskelu, työkyvyttömyys tai lasten hoito. Työvoimatutkimuksen aineisto ei kerro muut-kategorian sisällöstä enempää – kyseessä voi olla esimerkiksi yliopiston pääsykokeisiin valmistautuminen tai lomailu ennen varusmiespalveluksen tai opiskelun alkua. Muu-luokkaa kuvaa se, että vuosi ennen tutkimuksen tekoa 60 prosenttia oli ollut opiskelemassa. Loput olivat olleet työssä, työttömänä, varusmiespalveluksessa, työkyvyttömänä tai tekemässä ”jotain muuta”. Lähes 80 prosenttia joukosta oli suorittanut lukion tai ammattikoulun. 70 prosenttia ei ollut rekisteröitynyt työvoimatoimistoon.

Missä nuorisotyöttömyys on vakavinta?

Työttömyys- ja NEET-astetta verrattaessa huomataan, että nuorisotyöttömyys on erilaista eri maissa. Esimerkiksi Bulgariassa, jossa nuorten työttömyysaste (kuvio 7) jää kauas Espanjan ja Kreikan luvuista, on nuorista suurempi osa ilman sekä työtä että koulutusta. Itä-Euroopassa taas erityisesti Liettuassa, Slovakiassa ja Puolassa työttömyysaste on varsin korkea, mutta ilman koulutusta ja työtä on vain hieman suurempi osuus ikäluokasta kuin Suomessa.

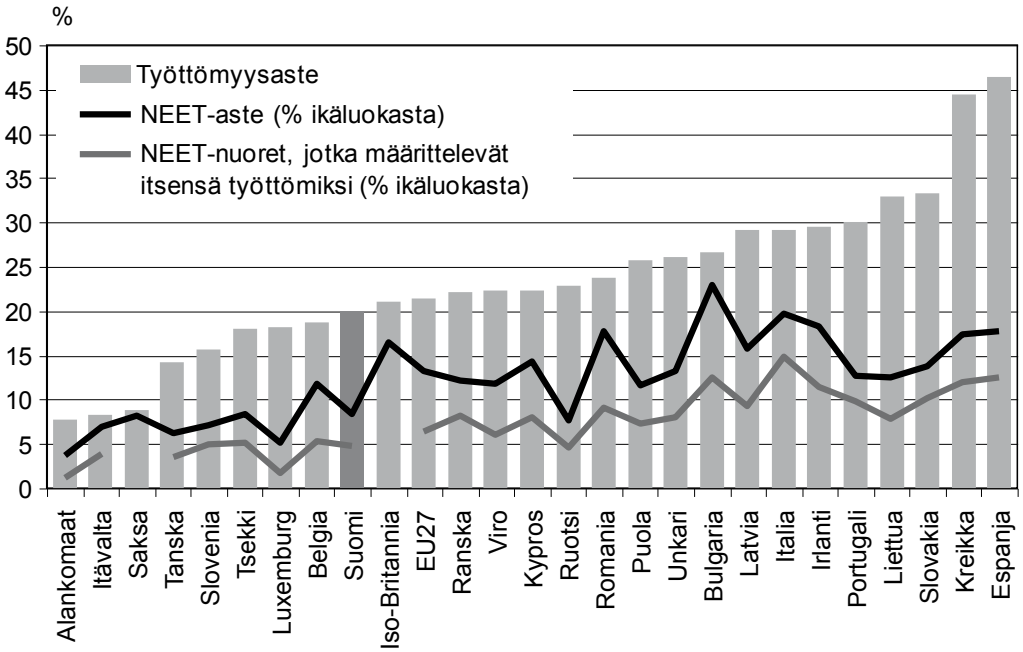
Suomessa nuorisotyöttömyys näyttää suhteellisen ”hyvälaatuiselta”. Työelämän ja koulutuksen ulkopuolella olevista nuorista niitä, jotka ovat omasta mielestään työttömiä, on Suomessa varsin vähän – selvästi vähemmän on vain Alankomaissa ja Luxemburgissa. Työssä tai koulutuksessa olemattomia nuoria on kyseisen ikäluokan nuorista

Suomessa yhtä suuri osuus kuin Saksassa, vaikka Suomen työttömyysaste on yli 10 prosenttiyksikköä korkeampi. (Kuvio 7.)

NEET-aste kertoo myös Espanjan ja Kreikan nuorten tilanteesta. Näiden maiden työttömyysasteet ovat korkeat; viime vuonna espanjalaisnuorten työttömyysaste oli jo yli 50 prosenttia. Koska työttömyysaste tarkoittaa työttömien osuutta työvoimasta, ei joka toinen nuori Espanjassa tai Kreikassa ole työtön. NEET-astetta tarkasteltaessa huomataan, että sekä työn että koulutuksen ulkopuolella olevia nuoria on Kreikassa ja Espanjassa suurin piirtein yhtä suuri osa nuorista (noin 20 %) kuin Italiassa ja Romaniassa ja jopa vähemmän kuin Bulgariassa. Vaikka tilanne näissä maissa onkin huonompi kuin EU-maissa keskimäärin (13 %), ei ero ole niin dramaattinen kuin pelkkiä työttömyysasteita vertaamalla voi olettaa.

NEET-asteiden tulkinnassa on otettava huomioon, että luvut sisältävät esimerkiksi kotona lapsiaan hoitavat nuoret naiset, joiden osuus Etelä- ja Itä-Euroopassa on huomattavasti suurempi kuin Pohjoismaissa. Niitä nuoria, joilla ei ole koulutuspaikkaa, työtä ja jotka itse kokevat olevansa työttömiä, on Espanjassa ja Kreikassa alle 13 prosenttia ikäluokasta. Luku on kaksinkertainen EU-keskiarvoon verrattuna, mutta kaukana siitä virhetulkinnasta, jonka mukaan joka toinen nuori Espanjassa ja Kreikassa olisi työtön.

Kuvio 7. 15–24-vuotiaiden työttömien osuus työvoimasta (työttömyysaste), ilman työtä, koulutusta tai kurssitusta olevien osuus ikäluokasta (NEET-aste) sekä itsensä työttömäksi määrittelevien NEET-nuorten osuus ikäluokasta EU-maissa vuonna 2011. Vuosikeskiarvo.*



* Varusmiespalveluksessa olevat eivät ole mukana.

Lähde: Eurostat 2012

Lähteet:

ETK 2011. Tilasto Suomen eläkkeensaajista. http://www.etk.fi/fi/gateway/PTARGS_0_2712_459_440_3034_43/http%3B/content.etk.fi%3B7087/publishedcontent/publish/etkfi/fi/julkaisut/tilastojulkaisut/tilastovuosikirjat/tilasto_suomen_elakkeensaajista_2011_7.pdf.

Eurostat 2012. Labour Force Survey. Mikroaineisto.

Eurostat 2013. Tietokantataulukot. Unemployment rate, annual average, by sex and age groups (%) [une_rt_a]. Supplementary indicators to unemployment, annual average, by sex and age groups (lfsi_sup_age_a). http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database.

Hämäläinen, Kati & Hämäläinen, Ulla 2012. Matkalla maailmalle. Nuorten työttömyyden esiintyvyys ja kesto. Työpoliittinen Aikakauskirja 2/2012.

Kela 2012. Kelan vammaisetuustilasto. http://www.kela.fi/it/kelasto/kelasto.nsf/NET/180612141439MR/$File/Vamm_11.pdf?OpenElement.

Kojo, Marjaana 2012. Pause päällä. Työn ja koulutuksen ulkopuoliset jaksot nuorten elämänkulussa. Janus 2/2012.

Okkonen, Kaisa-Mari 2008. Syrjäytymistä on vaikea kuvata tilastoilla. Hyvinvointikatsaus 2/2008.

Raitasalo, Raimo & Maaniemi, Kaarlo 2011. Nuorten mielenterveyden häiriöiden aiheuttamat sairauspoissaolot ja työkyvyttömyys vuosina 2004–2009. Nettityöpapereita 23/2011. Kela.

Tilastokeskus 2013. Työvoimatutkimuksen tilastotietokanta.

Netissä julkaistut artikkelit kirjaksi Nuoret Helsingissä 2011

Helsingissä on tutkittu nuorten vapaa-aikaa ja harrastuksia kymmenen vuoden välein vuosina 1982, 1990, 2000 ja 2011. Vuoden 2011 tutkimuksen aineisto kerättiin sähköisenä lomakekyselynä keväällä 2011. Tutkimukseen osallistui 42 koulua ja sen kohderyhmänä olivat peruskoulun 5.–9.-luokkalaiset sekä lukion ja ammatillisen oppilaitoksen 1. ja 2. vuositason oppilaat. Tutkimukseen vastasi reilut 1 400 iältään 11–19-vuotiasta nuorta. Kyselyyn vastasi 75 prosenttia otokseen valituista kouluista.

Tutkimuksen tulokset julkaistiin 24 artikkelina tutkimuksen omilla nettisivuilla. Ensimmäiset syyskuussa 2011, viimeisin päivitettiin elokuussa 2012.

Kirja **”Nuoret Helsingissä 2011. Vapaalla, koulussa, vaikuttamassa”** (192 s.) toimitettiin näistä artikkeleista. Kirja ilmestyi joulukuussa 2012.

Verrattuna vuoden 2000 tuloksiin:

- Nuorten tulevaisuutta koskevat huolenaiheet olivat yleistyneet
- Nuorten ympäristömyönteinen käyttäytyminen lisääntynyt
- Lukemisessa suuria muutoksia (lehdet, nuortenlehdet) – vähentynyt
- Valokuvaaminen ja kevyen musiikin soittaminen ja laulaminen yleistyneet → erityisesti tytöt
- Television katsominen vähentynyt
- Piirtäminen, maalaaminen ja keräily vähentyneet
- Lasten ja nuorten kouluviihtyvyys parantunut
- Koulutyöhön vaikuttaminen lisääntynyt
- Itseilmaisu ja rohkeus lisääntyneet

Lähde: Vesa Keskinen & Anna Sofia Nyholm: Nuoret Helsingissä 2011. Vapaalla, koulussa, vaikuttamassa. Helsingin kaupungin tietokeskus. Tutkimuksia 2012:3.

ILTAPÄIVÄSEMINAARI 5.11.2013

Tilastokeskus | Helsingin kaupunki Tietokeskus | Suomen Tilastoseura

Aihe: Tilastojen luku- ja käyttötaito
Aika: 5.11.2013 klo 14–17
Paikka: Helsingin taloushallintopalvelun Saldo-auditorio, Sörnäisten rantatie 27 A
Järjestäjät: Tilastokeskus, Helsingin kaupungin Tietokeskus ja Suomen Tilastoseura

Ohjelma:

Tilaisuuden avaus
Kimmo Vehkalahti, puheenjohtaja, Suomen Tilastoseura
ja Ari Jaakola, tilasto- ja tietopalvelupäällikkö, Helsingin kaupungin tietokeskus

Mitä on tilastojen luku- ja käyttötaito
Jussi Melkas, YTL

Lapset tutkimassa koulun hyvinvointia –
***Lasten tilastollinen vuosikirja* -hankkeen esittely**
Minna Torppa, projektipäällikkö, Forum Virium Helsinki

ISLP – *International Statistical Literacy Project* –
tilastojen lukutaidon edistäminen maailmanlaajuisena ponnistuksena
Reija Helenius, kehittämispäällikkö, Tilastokeskus, ISLP Director

ISLP:n maailmanlaajuisen kisan voittoposterin tekemisen vaiheet
Riikka Muje ja Riikka Ylikulppi, Lyseonpuiston lukio, Rovaniemi

Seminaari oli osa kansainvälisen tilastovuoden ohjelmaa *(International Year of Statistics 2013)*. Juhlavuotta vietettiin maailmanlaajuisesti 126 maassa yli 2 100 organisaation voimin. Tilastovuoden tavoitteena oli kasvattaa tietoisuutta tilastoista ja niiden vaikuttavuudesta yhteiskunnan eri alueilla, tehdä tilastoalan ammatteja tunnetuksi sekä tukea tilastojen ja tilastotieteen kehitystä.

STATISTICS2013.ORG
KANSAINVÄLINEN TILASTOVUOSI
OSALLISTUVA ORGANISAATIO

Mitä on tilastojen luku- ja käyttötaito?
Hengenpelastus ja navigointi tilastotulvassa

Jussi Melkas
Tilastoseuran seminaari 5.11.2013

Miksi tilastojen lukutaitoa tarvitaan?

Tilastojen lukutaitoa edellyttää päätöksenteon aiempaa tiiviimpi kytkeytyminen tilastoihin erityisesti globalisaation seurauksena. EU:n politiikkaohjelmat ovat tästä hyvä esimerkki. Lisäksi tilastollisten mallien lisääntyvä käyttö päätöksenteossa ja sen riskit korostavat tilastojen lukutaidon merkitystä.

Ylipäätään kansalaiset joutuvat elämään keskellä tilastojen tulvaa, jota lisäävät mm. julkisen hallinnon aineistojen avaaminen laajaan käyttöön (Open Data), ns. datajournalismi (datapohjaisten aineistojen lisääntyvä journalistinen käyttö) sekä muoti-ilmiö Big Data, jonka vaikutuksiin törmätään niin yksityiselämässä kuin töissäkin.

Tilastojen lukutaidon edellytykset ja ongelmat eivät ole vain lukijan ongelmia, vaan ne on ymmärrettävä myös tilastoja laadittaessa.

Tilastojen lukutaidon tasot

Tilaston lukeminen on sen käyttöä. **Suppea lukutaito** auttaa pysymään hengissä tilastotulvassa: se takaa tilastollisten tietojen ymmärtämisen kadunmiehen käyttötarpeisiin. **Laaja tilastojen lukemisen taito** antaa mahdollisuuden navigoida tilastovirrassa: se antaa mahdollisuuden tilastojen kriittiseen lukemiseen ja aktiiviseen käyttöön työelämässä. Esitys käsittelee ns. maallikoilta edellytettävää tilastojen luku- ja hallintataitoa. Varsinaisten tilastoammattilaisten osaamiselle on tietenkin kovemmat vaatimukset.

Läheisiä käsitteitä

Numeracy, numerolukutaito, lukulukutaito
Numeracy on kykyä ymmärtää ja soveltaa yksinkertaisia matemaattisia käsitteitä. Sen perustaidot ovat yhteen- ja vähennyslasku, kertominen ja jakaminen.

Datalukutaito
Data literacy on kyky lukea dataa, luoda ja viestiä siihen perustuvaa informaatiota.

Medialukutaito
Media literacy on joukko kykyjä, jotka mahdollistavat monenlaisissa medioissa, genreissä ja muodoissa esiintyvien viestien analyysin, arvioinnin ja luomisen.

Transliteracy
Kyky lukea, kirjoittaa ja olla vuorovaikutuksessa erilaisilla alustoilla, erilaisten työvälineiden ja medioiden avulla.

Information literacy
Kyky tunnistaa informaation tarve, kyky tunnistaa, paikallistaa, arvioida ja käyttää tehokkaasti informaatiota käsillä olevan ongelman ratkaisemiseen.

Lähde: Wikipedia (määritelmiä mukailtu)

Tilastojen suppea lukutaito: miten pysyä hengissä tilastotulvassa?

On tunnettava tilastojen peruskielioppi: miten taulukoita ja tilastografiikkaa luetaan. Se sisältää myös suoriutumisen esimerkiksi tilastografiikan lukemisen edellyttämistä keskeisimmistä tarkistuksista.

Perusmatematiikan osaaminen: laskutaito, joka sisältää vähintään prosenttilaskun, muutosprosenttien, indeksien ja keskilukujen (ja vähän myös hajonnan) ymmärtämisen. Myös otannasta on ymmärrettävät perusasiat eli perusjoukon tunnistaminen ja edustavan näytteen edellytysten tunteminen.

Tilastoissa käytettävien (substanssi-) käsitteiden ymmärtäminen: niiden avulla tietää, mistä tilasto kertoo, esim. BKT, työttömyysaste, inflaatio, syntyvyys, kuolleisuus, -kanta, - tase.

Tiedon tuotannon perusasioiden ymmärtäminen; esim. että kysymys vaikuttaa siihen, mitä vastataan, tai että vastaajat valikoituvat aiheen ja tiedonkeruutavan mukaan (hyvä esimerkki, ks. Pajunen) ja että luokitukset ovat tulkintaa.

Tilastolähteiden tuntemus. Ainakin se pitää tietää, että Tilastokeskuksen verkkosivujen kautta löytyy kattava valikoima luotettavia tilastoja.

Laaja lukutaito: miten navigoidaan tilastotulvassa?

Laaja tilastojen lukemisen taito edellyttää, että henkiinjäämistason taitoja syvennetään ja laajennetaan. Tulee hallita tilastollisia työvälineitä, ymmärtää jotain kausaliteetin selvittämisen ongelmista, todennäköisyydestä ja riskeistä, summaindekseistä. On myös hallittava käsitteiden mittaaminen. Voidakseen työskennellä tehokkaasti tilastojen kanssa on tunnettava keskeiset tilastojärjestelmät, tärkeimmät tilastolähteet ja osattava käyttää tietoaineistoja ja erilaisia datoja.

Lisäksi tarvitaan **kriittistä suhtautumista** tilastoihin. On ymmärrettävä, että vaikka tilastofaktat ovat tosia, ne ovat aina myös konstruktioita ja valintaa. On osattava kiinnittää huomiota siihen, kuka on tehnyt tilaston ja missä tarkoituksessa. On pohdittava tilaston kattavuutta - puuttuuko tilastoista jotain? On otettava selvää, miten tilasto on laadittu ja onko tilaston laadussa ongelmia.

Tarvitaan myös kykyä arvioida tilastoista tehtyjen tulkintojen pätevyyttä.

Tilastojen lukutaito ja päätöksenteko

Laajan lukutaidon tehtävänä on varmistaa, että tilastoilla tuetaan tarkoituksenmukaisesti päätöksentekoa.

Päätöksenteon käytännön ongelmissa tarvitaan evidenssiä. Kaikki evidenssi ei kuitenkaan ole tilastoa, minkä vuoksi tilastollinen evidenssi on osattava yhdistää muunlaiseen tietoon.

Lisäksi tilastollinenkin tieto edellyttää sekä tilastollista asiantuntemusta että asiantuntemusta myös substanssikysymyksissä (taloustiedettä, sosiaalitieteitä, pyskologiaa). Tilasto-osaaminen onkin osattava kytkeä joustavasti eri alojen asiantuntijoiden ja eri tieteiden väliseen yhteistyöhön. Esimerkiksi Big Data ei aukea vain tekniikan avulla, vaan tulkinnassa ja mallintamisessa on ymmärrettävä myös substanssia (taloustiede, sosiaalitieteet, psykologia).

Miten tilastojen lukutaitoa voi kehittää?

Tärkeää on tilastoja koskevan kriittisen keskustelun seuraaminen. Toinen keskeinen asia on tilastojen aktiivinen käyttö – soveltaminen, erehtyminen ja oivaltaminen. Kolmas on virheistä ja puutteista raportoiminen – se auttaa paitsi itseä myös muita lukemaan tilastoja.

Lähteet, linkkejä ym.

Tilastokoulu: Tilastojen ABC, *http://tilastokoulu.stat.fi/verkkokoulu_v2.xql?page_type=ketusivu&course_id=tkoulu_tlkt*

Wikipedia, hakusanat: numeracy, media literacy, data literacy, information literacy, transliteracy

Pajunen Jari: Tutkimukset ja pikakyselyt. Blogi 8.8.2013. Taloustutkimus Oy. Saantitapa: *http://www.taloustutkimus.fi/blogi/?x1810595=w2844852*

Simpura Jussi – Melkas Jussi: Numerot käyttöön, Gaudeamus 2013

Rasmus Daniel, W. Why Big Data Won't Make You Smart, Rich, Or Pretty. Blogi 27.1.2012, *https://magazine.fastcompany.com*

Lumley Thomas, Why big data is not enough, Blogi 23.4.2013, *http://www.statschat.org.nz/author/thomas-lumley/*

Kahan Dan M. & al. Motivated Numeracy and Enlightened Self-Government, Yale Law School, Public Law Working Paper 307, 2013. Saantitapa: *http://www.culturalcognition.net/browse-papers/motivated-numeracy-and-enlightened-self-government.html*

Suzannah Brecknell. Interview: Hetan Shah, The Royal Statistical Society Civil service & world 31.10.2013. Saantitapa: *http://www.civilserviceworld.com/interview-hetan-shah-the-royal-statistical-society/*

Hauskoja tapoja oppia tilastojen käyttöä

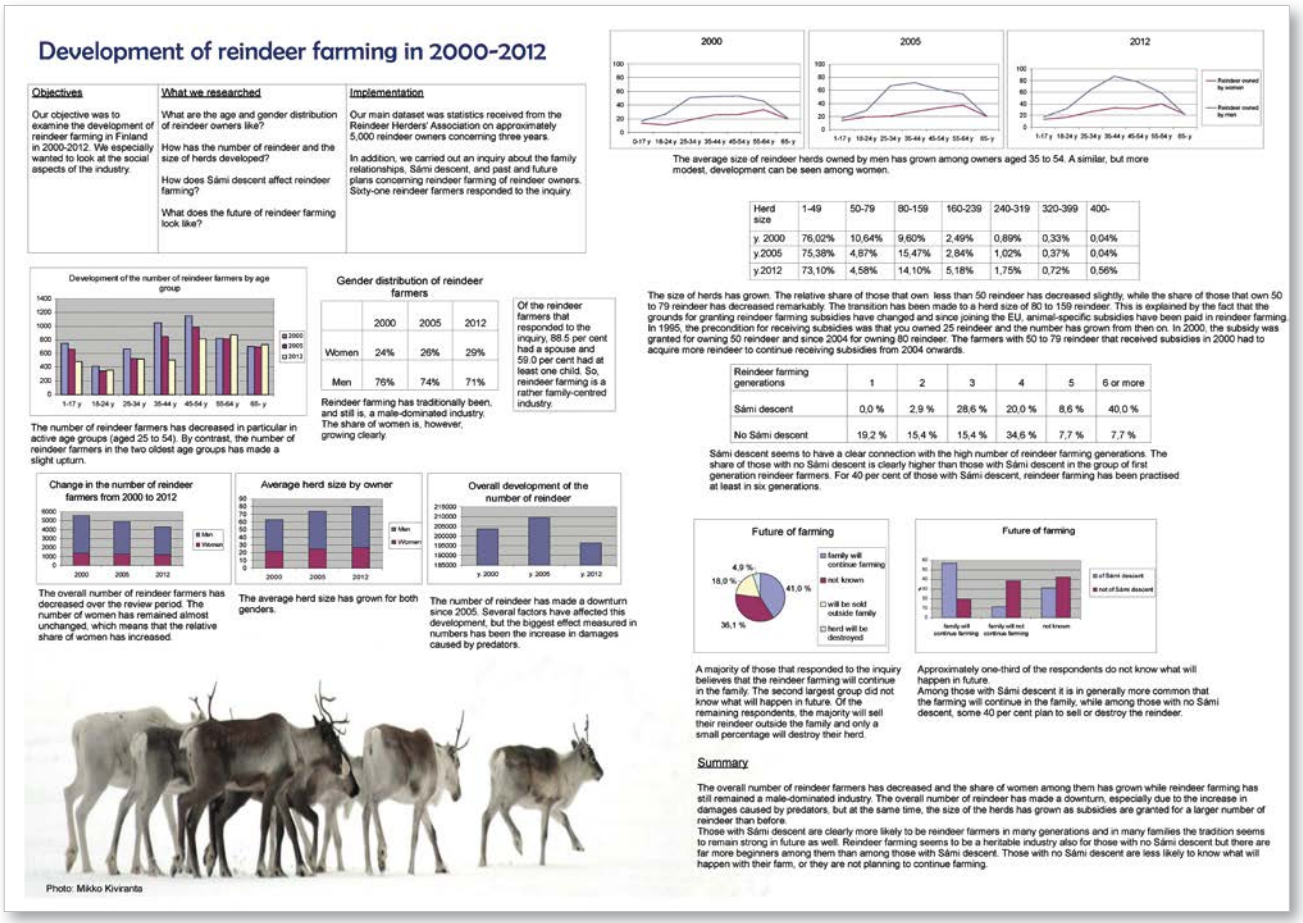
Kasvatustieteen kandidaatti Jenny Ståhlberg
Helsingin yliopisto
Korkeakouluharjoittelija, Tilastokeskus

Miten saada nuoret innostumaan numeroidentäyteisestä tilastojen maailmasta? MAOL ry, Suomen Tilastoseura ry ja Tilastokeskus järjestävät joka toinen vuosi käynnistyvän Suomen kansallisen Tilastokilpailun, jossa yläkoulu- ja lukioikäiset pääsevät joukkueina näyttämään tutkimuksentekotaitonsa. Kilpailun ideana on, että jokainen joukkue tekee pienen tutkimuksen valitsemastaan aiheesta: määrittää tutkimuskysymyksen, kertoo hieman austatietoja, kerää aineiston, analysoi sen ja tiivistää tutkimuksen kulun sekä saamansa tulokset posteriin eli tietotauluun. Jokainen kilpailuun osallistuva koulu valitsee parhaan posterin yläkoulu- ja lukiosarjasta ja lähettää ne Tilastokeskukseen Suomen kansallisen raadin arvioitavaksi.

Suomen sarjojen voittajaposterit jatkavat matkaansa kansainväliseen Tilastojen luku- ja käyttötaitokilpailuun, johon osallistui viimeksi oppilaita kolmestakymmenestä eri maasta. Kansainvälisen kilpailun järjestää the International Statistical Literacy Project ISLP, jonka johtajana on Tilastokeskuksen kehittämispäällikkö Reija Helenius (reija.helenius@tilastokeskus.fi) sekä apulaisjohtajat Pedro Campos (pedro.campos@ine.pt) Portugalista ja Steve MacFeely (steve.macfeely@cso.ie) Irlannista. Projektin ohjausryhmän muodostavat Heleniuksen lisäksi IASEn puheenjohtaja Iddo Gal Israelista, John Harraway Uudesta-Seelannista, Adriana D'Amelio Argentiinasta sekä James Nicholson ja Stephen Penneck Isosta-Britanniasta. ISLP-projekti toimii the International Association for Statistical Education IASEn sekä the International Statistical Institute ISIn alaisuudessa.

Projektin tavoitteena on parantaa tilastojen käyttö- ja lukutaitoa ympäri maailman ja lisätä kaikkien kansalaisten tilastoihin liittyvää tietämystä erilaisten aktiviteettien, kuten kilpailujen, avulla. Projekti tekee yhteistyötä eri maiden tilastokeskusten, yliopistojen ja oppilaitosten sekä tilastollisten yhteisöjen ja seurojen, kuten esimerkiksi Suomen Tilastoseuran kanssa.

Kansainvälisen tilastokilpailun molempien sarjojen voittajaposterit esitellään aina ISIn järjestämissä tilastokongresseissa, seuraavan kerran maailman 60. tilastokongressis-



sa Brasiliassa kesällä 2015. Erilaiset kongressit ja konferenssit ovatkin mainioita tilaisuuksia jakaa osaamista ja tietoa tilastotieteen opetuksesta ja hyvistä käytännöistä yli kansallisten rajojen. Suomi on pärjännyt kansainvälisessä kilpailussa loistavasti: lukiosarja on voitettu jo kolme kertaa peräkkäin. Vuoden 2013 lukiosarjan voitti suomalainen joukkue Rovaniemeltä aiheenaan porojen kasvatus.

Posterien teko on Suomen menestymisen myötä muodostumassa uudeksi varteenotettavaksi opetusvälineeksi tilastojen käytön, tilastotieteen ja monen muunkin oppiaineen opetuksessa. Kilpailussa ei siis ole ainoastaan kyse yhteistyötaitojen, tutkimuksenteon ja tilastojen käytön opettelusta, vaan myös laajemmasta uudesta opetustavasta, jonka avulla nuoret saadaan innostumaan, kiinnostumaan ja oppimaan uusista asioista käytännön tekemisen kautta.

Tilastokoulu – ovi tilastojen maailmaan

Tilastokeskus tarjoaa myös toisen erinomaisen opetusvälineen tilastollisen tiedon lisäämiselle: Tilastokoulun, joka löytyy verkosta osoitteesta http://tilastokoulu.stat.fi. Tilastokoulu sisältää Tilastokeskuksen asiantuntijoiden tekemiä kursseja eri aiheista, ja niitä on tällä hetkellä yhteensä viisi. Kursseja ja oppimateriaaleja päivitetään ja lisätään jatkuvasti.

Tilastojen ABC -kurssi tarjoaa perustiedot tilastojen ymmärtämiselle ja käyttämiselle sekä tilastollisen tutkimuksen tekemiselle. Kurssia voivat mainiosti hyödyntää ala- ja yläkouluopettajat omassa opetuksessaan sekä esimerkiksi Tilastokilpailuun osallistuvat joukkueet. **Työmarkkinatilastot -kurssi** opettaa työmarkkinatilastoinnin peruskäsitteet, työmarkkinatilastojen, kuten palkka- ja työvoimakustannustilastojen, muodostamisen sekä työmarkkinoiden analysoinnin niin kotimaisten kuin kansainvälistenkin aineistojen pohjalta. **Indeksit -kurssi** tutustuttaa erilaisiin indekseihin, joita ovat muun muassa hinta-, kustannus- ja määräindeksit, indeksien laskentakaavoihin sekä niiden eroihin. **Väestötieteen perusteet -kurssi** taas kuvaa väestötieteen keskeiset käsitteet, tarkastelee väestömuutoksiin vaikuttavia tekijöitä sekä väestönkehityksen ja yhteiskunnan taloudellisen ja sosiaalisen kehityksen välistä suhdetta. **Kansantalouden tilinpito -kurssilla** käydään läpi kansantalouden tilinpidon käyttöalueet ja sen historia, sen tärkeimmät määritelmät ja käsitteet sekä kansantalouden tilinpidon laskennan yleiset periaatteet.

Tilastokoulu tarjoaa jokaisen kurssin yhteydessä havainnollistavia esimerkkejä ja hyödyllisiä harjoitustehtäviä sekä erikseen eri luokka-asteille suunnattuja harjoitustehtäviä yleiseen tilastojen luku- ja käyttötaitoon liittyen. Tilastokoulun yhteydessä voi myös pelata hauskaa **Tilastovisaa**, joka tutustuttaa yksityiskohtaisiin tilastollisiin tietoihin sekä Tilastokeskuksen toimintaan ja tarjontaan. Tilastokoulusta löytyy myös tiedonlähdevinkkejä opettajille ja opiskelijoille sekä esimerkiksi opas opinnäytetyötä tekevälle ja tietoa Tilastokeskuksen tarjoamista koulutuspalveluista. Tilastokoulua ja sen oppimateriaaleja voivat siis hyödyntää kaikki tilastotiedosta ja tilastoista kiinnostuneet alakoulusta lähtien!

Lisätietoja Tilastojen luku- ja käyttötaitokilpailusta löydät osoitteesta
http://tilastokoulu.stat.fi/verkkokoulu_v2.xql?page_type=opettajalle
ja kansainvälisestä kilpailusta *http://iase-web.org/islp/.*
Tilastokoulu opettaa osoitteessa *http://tilastokoulu.stat.fi.*

ILTAPÄIVÄSEMINAARI 10.6.2014

Helsingin kaupunki
Tietokeskus

SUOMEN TILASTOSEURA

Aihe: Kokemuksia sähköisistä tiedonkeruumenetelmistä
Aika: 10.6.2014 klo 14–17
Paikka: Helsingin kaupunkisuunnitteluviraston info- ja näyttelytila Laituri, vanhan linja-autoaseman rakennus Narinkan laidalla, Kamppi
Järjestäjät: Helsingin kaupungin Tietokeskus ja Suomen Tilastoseura

Ohjelma:

Tilaisuuden avaus
FT Jyrki Möttönen, Suomen tilastoseura

Kokemuksia työvoimatutkimuksen WEB-pilotista
kehittämispäällikkö Kirsti Pohjanpää, Tilastokeskus

Tiedonkeruumenetelmä vaikuttaa tutkimustuloksiin:
web-keruun ja puhelinhaastattelun yhdistäminen
yliaktuaari Tara Junes, Tilastokeskus

Tietoa kulttuurin kuluttamisesta lumipallomenetelmällä
yliaktuaari Sini Askelo ja tutkija Vesa Keskinen, Helsingin kaupungin tietokeskus

Helsinki 2050 – kokemuksia karttakyselyn toteuttamisesta
vuorovaikutussuunnittelija Maija Mattila,
Helsingin kaupunkisuunnitteluvirasto

Kommenttipuheenvuoro
professori Risto Lehtonen, Helsingin yliopisto

Kokemuksia sähköisistä tiedonkeruumenetelmistä

Ari Jaakola
Tilasto- ja tietopalvelupäällikkö, Helsingin kaupungin tietokeskus
Varaesimies, Suomen tilastoseura

Suomen tilastoseura ja Helsingin kaupungin tietokeskus järjestivät 10.6.2014 perinteisen iltapäiväseminaarin otsikolla ”Kokemuksia sähköisistä tiedonkeruumenetelmistä”. Ajankohtainen teema houkutteli Helsingin kaupunkisuunnitteluviraston Laituri-tilaan salintäyteisen, lähes seitsemänkymmenen hengen yleisön. Seminaari koostui neljästä erilaisia sähköisiä tiedonkeruumenetelmiä esitelleestä esityksestä sekä professori Risto Lehtosen kommenttipuheenvuorosta.

Tilastoseuran esimies Jyrki Möttösen avauspuheenvuoron jälkeen estradin otti haltuunsa kehittämispäällikkö Kirsti Pohjanpää Tilastokeskuksesta. Pohjanpää esitteli työvoimatutkimuksen web-pilotista saatuja kokemuksia. Pilotissa verrattiin web-tiedonkeruumenetelmällä saatuja tuloksia perinteiseen puhelinhaastattelulla saatuihin tuloksiin. Hankkeessa tarkasteltiin mm. ”EOS”-vastaamista (”Ei osaa sanoa”) sekä erilaisia ratkaisumalleja työtunteja ja työn etsimistä koskeneisiin kysymyksiin. Hankkeessa hyödynnettiin ns. ”Split plot”-asetelmaa, jossa vastaajajoukko oli jaettu kahteen samansuuruiseen ryhmään. Ryhmille tarjottiin hieman erilaiset tavat vastata kysymyksiin, mikä mahdollisti tarjottujen vastaustapojen vaikutusten vertailun.

Pilotista saatiin monia arvokkaita kokemuksia. Ensinnäkin havaittiin, että ensimmäiset päivät ovat tiedon keruun onnistumisen kannalta ratkaisevia. Suuri osa vastauksista saatiin jo parin ensimmäisen päivän aikana. Viikon jälkeen vastauksia saatiin enää vähän. Muistutusviestit tosin aktivoivat otoksessa mukana olleita ja paransivat vastausprosenttia. Pohjanpää piti prosessin nopeutta yhtenä web-kyselyn selvänä etuna. Toiseksi, naiset vastasivat hieman miehiä paremmin. Vastausasteet olivat vanhemmissa ikäluokissa nuorempia korkeammat. Kolmanneksi vastausasteet vaihtelivat myös vastaajien koulutusasteen mukaan. Koulutetumpien vastausaste oli korkeampi kuin vähemmän koulutetuilla.

Keskeisenä tuloksena Pohjanpää nosti esiin sen, että web-lomakkeella kerätty aineisto antoi samansuuntaisia vastauksia kuin mitä perinteisilläkin tiedonkeruumenetelmillä saadaan. Vaikka eroja voitiin havaita, ne eivät olleet tilastollisesti merkitseviä. Lisäksi web-tiedonkeruun avulla on mahdollista kerätä hyvin monenlaista taustatietoa, kuten vastausajankohta, vastaamiseen käytetty aika jne. Toisaalta web-tiedonkeruu on

haavoittuvaista. Vastaajat menettävät helposti mielenkiintonsa ja jättävät vastaamatta, mikäli esimerkiksi tekniikka ei syystä tai toisesta toimi kunnolla. Myös lomakesuunnittelun merkitys on suuri. Esimerkiksi ”EOS”-vastausvaihtoehdon tarjoamista kannattaa välttää. Lopuksi Pohjanpää totesi, että ns. Mixed mode –tiedonkeruu tekee selvästi tuloaan. Tarvitaan kuitenkin vielä lisätutkimuksia ja lisää kokemuksia, jotta ymmärretään paremmin, miten tämä uusi menetelmä vaikuttaa kuvattavasta ilmiöstä saataviin tietoihin.

Seuraavan esityksen piti Tilastokeskuksen yliaktuaari Tara Junes, joka esitteli web-tiedonkeruun ja puhelinhaastattelun yhdistämisestä saatuja tuloksia. Junesin esitys perustui kuluttajabarometrin taustatietojen keruumenetelmistä saatuihin kokemuksiin. Kuluttajabarometri kuuluu EU:n yhdenmukaistettuun suhdannesurvey-ohjelmaan ja se kuvaa kotitalouksien kulutus- ja säästämisaikeita sekä niihin vaikuttavia mielipiteitä. Kohdejoukkona ovat 15–84 -vuotiaat suomalaiset ja tiedot kerätään kuukausittain puhelinhaastattelujen muodossa.

Hankkeessa kokeiltu testitiedonkeruu koostui internet-kyselystä sekä kyselyyn vastaamatta jättäneille suunnatuista puhelinhaastatteluista. Vastauskatoon kohdistettu tarkastelu osoitti, että testitiedonkeruussa kieltäytyneiden vastaajien osuus oli jonkin verran perinteistä kuluttajabarometrin tiedonkeruuta suurempi. Lisäksi nuorempien ikäluokkien tavoitettavuus internetin kautta oli vanhimpia ikäluokkia parempi. Mielenkiintoinen piirre oli se, että työikäisten keskuudessa erot eivät olleet kovin suuria.

Varsinaiset tulokset osoittivat kuitenkin selvän eron testitiedonkeruun ja perinteisen tiedonkeruun välillä. Testitiedonkeruu ja erityisesti internet-tiedonkeruu tuotti huomattavasti puhelinhaastatteluja negatiivisemman barometriarvon. Internet-tiedonkeruun avulla muodostettu kuluttajabarometri antoi perinteiseen tiedonkeruuseen perustuvaa indikaattoria synkemmän kuvan suomalaisten luottamuksesta talouden kehitykseen. Toisaalta internet-kysely tuotti suuremman määrän positiivisia vastauksia, kun tutkittiin ostoaikeita. Lisäksi ”En osaa sanoa” -vastausten määrä oli testitiedonkeruussa suurempi kuin puhelinhaastattelussa.

Iltapäivän kolmas esitys pidettiin Helsingin kaupungin tietokeskuksen yliaktuaari Sini Askelon ja tutkija Vesa Keskisen toimesta. Esityksen aiheena oli Helsingin kulttuurikysely ja siihen ns. lumipallomenetelmällä kerätty kyselyaineisto. Lumipallomenetelmällä tarkoitetaan menetelmää, jossa verkkokyselyyn johtava linkki leviää vapaasti potentiaalisten vastaajien keskuudessa sähköpostin, sosiaalisen median tms. välineen kautta. Tässä tapauksessa linkki lähetettiin aluksi sadalle potentiaaliselle vastaajalle. Vastauspyynnön saajia puolestaan pyydettiin välittämään viestiä eteenpäin heidän omissa verkostoissaan. Menetelmä tuotti kaikkiaan yli 800 vastausta.

Lomakkeen lyhyydestä huolimatta saatu aineisto osoittautui varsin laajaksi. Esimerkiksi joka viides vastaaja kirjoitti mielipiteitään ”Sana on vapaa” -osioon. Askelon ja Keskisen mukaan menetelmä tavoitti perinteistä kyselytutkimuksen otospohjaista tiedonkeruuta paremmin erityisryhmiä, kuten graffitiporukoita sekä junioriurheilijoita. Myös kulttuurin ammattilaiset tavoitettiin lumipallomenetelmällä paremmin. Menetelmä osoittautui edulliseksi tavaksi kerätä nopeasti tuoretta tietoa rajatusta ja neutraalista aiheesta. Arkaluonteisten tai jollakin tapaa intohimoja herättävien teemojen kohdalla menetelmää ei tekijöiden mielestä kannata hyödyntää. Sen sijaan menetelmää kannattaisi kokeilla esimerkiksi vieraskielisten tavoittamiseen. Kysely leviäisi tällöin omakielisten lähettämänä. Tekijät korostivat myös sitä, että kyselyssä ei ollut otosta eikä kerätty aineisto edusta helsinkiläisiä pienoiskoossa. Näin ollen raportoinnissa ei voida puhua helsinkiläisistä vaan kyselyyn vastanneista. Myöskään vastausten karhuaminen ei ollut mahdollista.

Seminaarin neljäs ja samalla viimeinen varsinainen esitys käsitteli Helsinki 2050-kyselyä ja siitä saatuja kokemuksia. Esityksen piti vuorovaikutussuunnittelija Maija Mattila Helsingin kaupunkisuunnitteluvirastosta. Helsinki 20150-kysely liittyy Helsingin yleiskaavatyöhön. Sen avulla haluttiin kerätä kartalle kaupunkilaisten näkemyksiä tulevan yleiskaavan pohjaksi. Kaikkiaan kyselyyn vastasi noin 4700 henkilöä ja erilaisia näkemyksiä ja ehdotuksia kartalle kertyi yli 33 000. Helsingin väestön ikäjakaumaan verrattuna vastaajissa yliedustettuna olivat 20–50 -vuotiaat, erityisesti 30–39 -vuotiaat, ja aliedustettuina toisaalta alle kaksikymmenvuotiaat ja yli kuusikymmenvuotiaat. Vastaajien määrä vaihteli myös alueittain: Eniten vastaajia oli Kalliosta, Herttoniemi-Roihuvuoresta, Lauttasaaresta ja Oulukylästä.

Mattilan mukaan kerätty aineisto muodostaa rikkaan kokonaisuuden, josta riittää analysoitavaa. Vastausten perusteella vastaajat on voitu mm. luokitella urbaanien ja toisaalta tiivistyskriittisten ryhmiin. Aineisto tuokin hyvin esiin kaupungin moniäänisyyden; kaupunkilaisten mielipiteet eroavat toisistaan monissa keskeisissä kysymyksissä. Lisäksi aineisto pitää sisällään runsaasti kehittämisehdotuksia. Aineiston esittäminen kartalla helpottaakin kansalaisten kanssa käytävää keskustelua ja auttaa asioiden jäsentämisessä. Aineisto on avattu kaikkien käytettäväksi avoimena datana Helsinki Region Infoshare (HRI-) -palvelussa (*www.HRI.FI*).

Esitysten jälkeen kuultiin professori Risto Lehtosen kommenttipuheenvuoro päivän esityksiin. Professori Lehtonen kiinnitti huomiota ensinnäkin siihen, että esitykset jakautuivat kahteen eri ryhmään. Kahdessa ensimmäisessä esityksessä kuvattiin otospohjaisia tiedonkeruita, joilla pyritään hankkimaan perusjoukkoon yleistettävissä olevaa tietoa. Kahdessa jälkimmäisessä puolestaan kuvattiin tiedonkeruuta, jolla pyritään saamaan ketterästi paljon vastauksia, mutta jossa vastaajajoukko on itsevalikoituva. Vastaajien valikoitumismekanismia ei tällöin tunneta eivätkä vastaukset ole näin ollen

yleistettävissä mihinkään perusjoukkoon. Toiseksi, otosperustaisiin tiedonkeruumenetelmiin liittyen Lehtonen kiinnitti huomiota siihen, että työvoimatutkimuksen kohdalla tiedonkeruumenetelmä ei näyttänyt vaikuttavan merkittävästi tuloksiin kun taas kuluttajabarometrin kohdalla tiedonkeruumenetelmän vaikutukset näyttivät olevan merkittäviä. Tarvitaan lisäselvityksiä, jotta ymmärretään mistä erot johtuvat ja miten tiedonkeruun uusia menetelmiä voidaan jatkossa hyödyntää enemmänkin. Kommenttipuheenvuoronsa lopuksi Lehtonen kiitti seminaarin esittäjiä todeten, että sähköiset tiedonkeruumenetelmät tarjoavat koko joukon uusia mahdollisuuksia tiedonkeruuseen ja myös kerättyjen tietoaineistojen hyödyntämiseen. Mielenkiintoisia näkymiä näyttää avautuneen esimerkiksi avoimen datan ja joukkoistamisen muodoissa. Näihin kiitoksiin ja näkemyksiin myös Suomen tilastoseuran ja Helsingin kaupungin tietokeskuksen on helppo yhtyä.

ILTAPÄIVÄSEMINAARI 18.11.2014

Aihe: Suomen Syöpärekisteri – tilastointia ja tutkimusta
Aika: 18.11.2014 klo 14–17
Paikka: Unioninkatu 22, Helsinki
Järjestäjät: Suomen Syöpärekisteri ja Suomen Tilastoseura

Ohjelma:

Tilaisuuden avaus
FT Jyrki Möttönen, Suomen tilastoseura

Syöpärekisterin yleisesittely
johtaja Nea Malila

Tilastotutkimusta ja tuotantoa – Syöpärekisterin tilastotoimintojen esittely
johtava tilastotieteilijä Janne Pitkäniemi, Suomen Syöpärekisteri

Syöpäseulontojen tilastollinen tutkimus
vanhempi tutkija Sirpa Heinävaara

Syöpäpotilaiden eloonjäämisen arviointi
tutkija Karri Seppä, Suomen Syöpärekisteri

Gunnar Modeen -minnesmedaljen

Jukka Hoffrén

Statistiska Samfundet i Finland r.f. har i samband med de nordiska statistikdagarna traditionsenligt delat ut Gunnar Modeen -minnesmedaljen till särskilt meriterade statistiker. Praxisen har varit att dela ut medaljen till en representant för det land där statistikdagarna hålls.

Gunnar Modeen -minnesmedaljen beviljas för en betydande livsgärning inom statistikbranschen. Meningen är att den person som belönas är en framstående senior expert inom statistikbranschen, som uttryckligen utmärkt sig i det praktiska statistikarbetet och som uppskattas av sina kolleger.

Styrelsen för Statistiska Samfundet väljer den person som får medaljen och medaljen överlåts i samband med ett nordiskt statistikermöte. Enligt fondens stadga överlåts medaljen till en betydande nordisk statistiker från det land som respektive år arrangerar mötet. Den första medaljen överläts vid det nordiska statistikermöte som hölls i Finland år 1989.

Bakgrunden till och kriterier för GM-minnesmedaljen

Efter Gunnar Modeens bortgång år 1988 grundades en medaljfond till hans minne. Medaljen utarbetades på basis av den medaljong som Gunnar Modeens familj gett konstnären Matti Haupt i uppdrag att utforma till Modeens 70-årsdag år 1965. Mottagaren av medaljen väljs av styrelsen för Statistiska Samfundet i Finland och medaljen överlåts i samband med ett nordiskt statistikermöte. Enligt fondens stadga överlåts medaljen till en betydande nordisk statistiker från det land som respektive år arrangerar mötet. Den första medaljen överläts vid Nordiska Statistikermötet i Finland år 1989. Priset utdelas vart tredje år till en meriterad statistiker från det land där Nordiska Statistikermötet anordnas.

Allmänna kriterier för Gunnar Modeen -minnesmedaljen:
- priset beviljas för en betydande livsgärning inom statistikbranschen.

Den person som tilldelas medaljen:
- är en expert inom statistikbranschen, som uttryckligen utmärkt sig i det praktiska statistikarbetet
- är en nordisk, framstående senior expert som uppskattas av sina kolleger,
- har akademisk examen (magister, licentiat eller doktor) och
- är villig att ta emot GM-medaljen

Mottagare av GM-minnesmedaljen

Den första medaljen tilldelades Mauno Koivisto, Finlands dåvarande president, som en särskild hedersbetygelse. År 1989 var han beskyddare av Nordiska Statistikermötet i Finland som firade 100-årsjubileum för nordisk statistik. Ytterligare en medalj delades ut på mötet och mottagare var professor Eino H. Laurila. Övriga mottagare av medaljen:

År 1992 tilldelades medaljen inte.
År 1995 direktör Poul Jensen, Danmarks Statistik.
År 1998 professor Sven Nordbotten, Universitetet i Bergen.
År 2001 professor Emeritus Gunnar Kulldorf, Umeå universitet.
År 2004 direktör Asta Manninen, Helsingfors stads faktacentral.
År 2007 generaldirektör Hallgrímur Snorrason, Hagstofa, Island.
År 2010 direktör Lars Thygesen, Danmarks Statistik.
År 2013 Liv Hobbelstad Simpson, pensionerad från Statistisk sentralbyrå (SSB) som Head of National accounts och past chair of IARIW

Scandinavian Journal of Statistics

Recognised as a leading journal in its field, the Scandinavian Journal of Statistics is an international publication devoted to reporting significant and innovative original contributions to statistical methodology, both theory and applications. The journal specializes in statistical modelling showing particular appreciation of the underlying substantive research problems. Scandinavian Journal of Statistics is published on behalf of the Danish Society for Theoretical Statistics, the Finnish Statistical Society, the Norwegian Statistical Society and the Swedish Statistical Society. Journal is currently edited by professors Holger Rootzén and Mats Rudemo. National editors for Finland are Jukka Corander and Risto Lehtonen (University of Helsinki).

Members of the Finnish Statistical Society entitled to discount prices when ordering the Scandinavian Journal of Statistics. For further information please see webpage:

http://www.wiley.com/bw/subs.asp?ref=0303-6898&site=1

ISI Journal Citation Reports® Ranking: 2013: 47/119 Statistics & Probability
Impact Factor: 1.063

Suomen Tilastoseuran hallitus vuonna 2013

Board members of the Finnish Statistical Society 2013

Puheenjohtaja Chair	Kimmo Vehkalahti	Valtiotiet. toht. D.Soc.Sc.
Varapuheenjohtaja Vice Chair	Jyrki Möttönen	Filosofian toht. PhD
Rahastonhoitaja Treasurer	Leena Kalliovirta	Valtiotiet. toht. D.Soc.Sc.
Sihteeri Secretary	Kaisa Mäntysaari	Filosofian maist. M.Sc.
Jäsen Member	Ville Hyvönen	Yhteiskuntatiet. maist. M.Soc.Sc.
Jäsen Member	Ari Jaakola	Filosofian maist. M.Sc.
Jäsen Member	Annu Nissinen	Valtiotiet. maist. M.Soc.Sc.
Jäsen Member	Maiju Pankakoski	Valtiotiet. maist. M.Soc.Sc.
Jäsen Member	Samuli Ripatti	PhD
Jäsen Member	Kristiina Tyrkkö	Yhteiskuntatiet. maist. M.Soc.Sc.

Suomen Tilastoseuran hallitus vuonna 2014

Board members of the Finnish Statistical Society 2014

Puheenjohtaja Chair	Jyrki Möttönen	Filosofian toht. PhD
Varapuheenjohtaja Vice Chair	Ari Jaakola	Filosofian maist. M.Sc.
Rahastonhoitaja Treasurer	Leena Kalliovirta	Valtiotiet. toht. D.Soc.Sc.
Sihteeri Secretary	Kaisa Mäntysaari	Filosofian maist. M.Sc.
Jäsen Member	Tara Junes	Valtiotiet. maist. M.Soc.Sc.
Jäsen Member	Tuomo Nieminen	Valtiotiet. Yo. Stud.Soc.Sc.
Jäsen Member	Maiju Pankakoski	Valtiotiet. maist. M.Soc.Sc.
Jäsen Member	Pekka Pere	Doctor of Philosophy DPhil
Jäsen Member	Marjo Pyy-Martikainen	Filosofian toht. PhD
Jäsen Member	Kristiina Tyrkkö	Yhteiskuntatiet. maist .M.Soc.Sc.

Suomen Tilastoseuran julkaisuja

Publikationer utgivna av Statistiska Sammanfundet

Publications issued by the Finnish Statistical Society

1. Monikielinen väestötieteen sanakirja, suomenkielinen laitos, Helsinki 1962.
   Multilangual Demographic Dictionary, Finnish section, Helsinki 1962.

2. Suomen Tilastoseura – Statistiska Sammanfundet i Finland 1920-1970, Porvoo – Borgå 1970.

3. Pohjoismainen tilastosanasto, toinen tarkistettu laitos.
   Nordisk statistik nomenklatur, andra reviderade upplagan.
   Nordic statistical nomenclature, 2nd revised edition. Jyväskylä 1975 (loppuunmyyty)

4. Aikasarja-analyysin menetelmiä, Helsinki 1977.

5. Pekka Tavaila: Leo Törnqvist Posti- ja lennätinhallituksen liiketaloudellisen tutkimuslaitoksen esimiehenä 1949–1977, Helsinki 1982.

6. Otanta teoriassa ja käytännössä. Vesa Kuusela ja Leif Nordberg (toim.). Helsinki 1986.

7. Suomen Tilastoseura 70 vuotta. Statistiska Sammanfundet i Finland 70 år.
   The Finnish Statistical Society 70 years. Helsinki 1991.

Tilastotieteellisiä tutkimuksia

Statistiska undersökningar

Statistical Research Reports

ISSN 0356–3499

1. Pentti Manninen: Puolueiden kannatusosuuksien estimoinnin tarkkuus Demingin vyöhykepoiminnassa. [The Accuracy of Party Support Estimation in Deming Zone Selection.] In Finnish with English Summary. Helsinki 1976.

2. Timo Hakulinen: On Competing Risks of Death. Helsinki 1977.

3. Lars-Erik Öller: Time Series Analysis of Finnish Foreign Trade. Helsinki 1978.

4. Pekka Laippala: The Empirical Bayes Two-Action Rules with Floating Optimal Sample Size and Exponential Conditional Distributions. Helsinki 1980.

5. Markku Nurminen: Some Developments in Quantitative Methods of Epidemiology. Helsinki 1982.

6. Pentti Saikkonen: Comparing Asymptotic Properties of Some Tests Used in the Specification of Time Series Models. Helsinki 1985.

7. Lauri Tarkkonen: On Reliability of Composite Scales. Helsinki 1987.

8. Juni Palmgren: Models for Categorical Data with Errors of Observation. Helsinki 1987.

9. Ari Veijanen: On Estimation of Parameters of Partially Observed Random Fields and Mixing Processes. Helsinki 1989.

10. Ritva Luukkonen: On Linearity Testing and Model Estimation in Non-Linear Time Series Analysis. Helsinki 1990.

11. Hely Salomaa: Factor Analysis of Dichotomous Data. Helsinki 1990.

12. Kenneth Nordström: Contributions to the Comparison of Linear Models and to the Löwner-Ordering Antitonicity of Generalized Inverses. Helsinki 1990.

13. Seppo Laaksonen: Handling Household Survey Nonresponce Data. Helsinki 1992.

14. Mervi Eerola: On Predictive Causality in the Statistical Analysis of a Series of Events. Helsinki 1993.

15. Mikael Linden: Studies in Integrated and Co-Integrated Economic Time Series. Helsinki 1995.

16. Tadeusz Dyba: Precision of Cancer Incidence Predictions Based on Poisson Distributed Observations. Helsinki 2000.

17. Kimmo Vehkalahti: Reliability of Measurement Scales. Helsinki 2000.

18. Sirpa Heinävaara: Modelling survival of patients with multiple cancers. Helsinki 2003.

Suomen Tilastoseuran vuosikirja

Årsbok för Statistiska Sammanfundet i Finland

The Yearbook of the Finnish Statistical Society

ISBN 035–5941

1975, Helsinki 1976
1976, Helsinki 1977
1977, Helsinki 1978
1978, Helsinki 1979
1979, Helsinki 1980
1980, Helsinki 1981
1981, Helsinki 1982
1982, Helsinki 1983
1983, Helsinki 1984
1984, Helsinki 1985
1985, Helsinki 1986
1986, Helsinki 1987
1987, Helsinki 1988
1988–1989, Helsinki 1990
1990, Helsinki 1991
1991, Helsinki 1992
1992, Helsinki 1993
1993, Helsinki 1994
1994, Helsinki 1995
1995, Helsinki 1996
1996, Helsinki 1997
1997, Helsinki 1998
1998, Helsinki 1999
1999–2000, Helsinki 2000
2001, Helsinki 2002
2002, Helsinki 2003
2003, Helsinki 2004
2004, Helsinki 2005
2005, Helsinki 2006
2006, Helsinki 2007
2007, Helsinki 2008
2008, Helsinki 2009
2009, Helsinki 2010
2010, Helsinki 2011
2011–2012, Helsinki 2012

Kaikkia Tilastoseuran julkaisuja voi tilata seuran sihteeriltä sähköpostiosoitteesta sihteeri@tilastoseura. Julkaisun hinta on 12 € kappale + toimituskulut. Joidenkin julkaisujen painokset ovat tosin jo loppuneet.

Muita julkaisuja

Andra publikationer

Other publications

Suomen tilastoseura 1920–1945, Helsinki 1946
Statistiska Sammanfundet i Finland 1920–1945, Helsingfors 1946

Pohjoismainen tilastosanasto – Nordisk statistisk nomenklatur, Kööpenhamina 1954

13:e Nordiska statistikermötet i Helsingfors 14–16 juni 1973, Jyväskylä 1974
The 13th Joint Meeting of the Nordic Statistical Societies in Helsinki June 1973, Jyväskylä 1974

Det 18:e nordiska statistikmötet i Esbo, Hundraårsjubileum, Helsingfors 1990

The Joint Conference of the Nordic Statisticians in Espoo, Finland 1989, Helsinki 1990

Scandinavian Journal of Statistics
Theory and Applications

The Scandinavian Journal of Statistics (SJS) is an international statistical journal which welcomes contributions from all countries. The language is English.

The Main purpose of the journal is to publish research papers in theoretical and applied statistics. It also welcomes statistically motivated papers on relevant aspects of probability and other fields, as well as papers on innovative applications of statistical methodology.

Scandinavian Journal of Statistics is published under the auspices of
the Danish Society for Theoretical Statistics
the Finnish Statistical Society
the Norwegian Statistical Society
the Swedish Statistical Association

Scandinavian Journal of Statistics is published quarterly in March, June, September and December by Blackwell Publishers, 108, Cowlwy Road, Oxford OX4, 1JF, UK or 238 Main Street, Cambridge, MA 02142, USA.

Myönnetyt Leo Törnqvist –palkinnot

1978 **Rene Tigerstedt, Helsingin yliopisto.** En modell för valbeteende i trafiken.

1979 **Pirkko Kirjavainen, Turun kauppakorkeakoulu.** Mallin rakentaminen ja ennusteen laatiminen Suomen sähkön kulutukselle kahta aikasarja-analyysimenetelmää käyttäen.

1980 **Esa Läärä, Helsingin yliopisto.** Ikä-, aika- ja kohorttitekijöiden vaikutukset Suomen miesten keuhkosyöpäsairastavuudessa vuosina 1953–76.

1981 **Arvi Suvanto, Tampereen yliopisto.** Kausivaihtelu aikasarjamalleissa.

1982 **Maija Salo, Helsingin yliopisto.** Yritys prioritiedon käytöstä alkoholijuomien kulutusta selittävän kysyntämallin tukena.Jamel Boucelham, Jyväskylän yliopisto: Tunnustuspalkinto.

1983 **Vesa Vihriälä, Helsingin yliopisto.** Aikasarjojen välisen riippuvuuden mittaus ja testaus: sovellus suomalaisiin rahatalouden sarjoihin. Pirkko Welin, Tampereen yliopisto: Tunnustuspalkinto.

1984 **Jari Palsio, Turun kauppakorkeakoulu.** Skenaarioiden rakentaminen ristivaikutusanalyysimallia käyttäen.

1985 **Kenneth Nordström, Helsingin yliopisto.** Gauss-Markov-mallien erikoisongelmista.

1986 **Tapio Nummi, Tampereen yliopisto.** APL-pohjainen ohjelmisto GMANOVA-mallille.

1987 **Ari Veijanen, Helsingin yliopisto.** Pickardin kentän soveltamisesta kuva-analyysissä.Kari Nissinen, Jyväskylän yliopisto: Tunnustuspalkinto.

1988 **Jaason Haapakoski, Helsingin yliopisto.** Binomijakautuneiden muuttujien muutospisteongelma.

1989 **Pasi Korhonen, Helsingin yliopisto.** Kemometrian tilastollisista menetelmistä.

1990 **Päivi Partanen, Jyväskylän yliopisto.** Suljetun populaation koon estimointi merkintä-takaisinpyynti-menetelmällä: log-lineaarinen lähestymistapa. Markku Nurhonen, Tampereen yliopisto: Tunnustuspalkinto.

1991 **Elina Järvinen, Helsingin yliopisto.** Rajoitettujen, stokastisten ja konveksien estimaattoreiden käytöstä polynomisen viipymämallin parametrien estimoinnisssa simulointikokeiden valossa.

1992 **Jouni Kuha, Helsingin yliopisto.** Binääristen regressiomallien selittäjien mittausvirheet ja parametriestimaattien mittausvirhekorjaukset. Juha Heikkinen, Jyväskylän yliopisto: Tunnustuspalkinto.

1993 Palkintoa ei jaettu (yhtään ehdotusta ei saatu).

1994 **Ilkka Taskinen, Jyväskylän yliopisto.** Äärelliset Markovin ketjut ja annelointi.

1995 **Mika Rautakorpi, Teknillinen korkeakoulu.** Application of Markov chain techniques in certification of software.Tuija Jäppilä, Jyväskylän yliopisto: Tunnustuspalkinto.

1996 **Veli-Matti Suppola, Jyväskylän yliopisto.** Robustit menetelmät. Jakaumien vinouden vaikutuksesta korrelaatiomatriisin estimointiin.

1997 **Albert Höglund, Teknillinen korkeakoulu.** An Anomaly Detection System for Computer Networks.

1998 **Samuli Visuri, Oulun yliopisto.** Robustista kovarianssimatriisin estimoinnista ja sen sovelluksista signaalinkäsittelyssä.

1999 **Jani Raitanen, Tampereen yliopisto.** Jalkapallo-ottelun lopputuloksen tilastollinen mallintaminen.

2000 **Reijo Sund, Helsingin yliopisto.** Tilastollisia menetelmiä dynaamisten potilaspopulaatioiden mallintamiseen. Tapahtumahistoria-analyysia hoitoilmoitusrekisterin skitsofreenikoille.

2001 **Samu Mäntyniemi, Oulun yliopisto.** A Hierarchial Bayes Model for Assessing Salmon (Salmo salar L.) Parr and Smolt Populations.

2002 **Ilmari Juutilainen, Oulun yliopisto.** Teräslevyjen lujuuden ennustaminen regressio- ja neuroverkkomalleilla.

2003 **Leena Kalliovirta, Helsingin yliopisto.** Mar-malli.

2004 **Mikko Myrskylä, Jyväskylän yliopisto.** Estimation of Class Frequencies with Micro Level Auxiliary Information.

2005 **Antti Liski, Tampereen yliopisto.** Lonkkamurtumapotilaiden hoitokustannusten vertailu vastaavuuspistemäärään perustuvalla menetelmällä.

2006 **Karri Seppä, Oulun yliopisto.** Suomalaisten paksusuolisyöpäpotilaiden ennusteen analyysi suhteellisen elossapysymisen ja syykohtaisen kuolleisuuden malleilla käyttämällä suurimman uskottavuuden ja Bayesin menetelmiä.

ja

Jukka Siren, Helsingin yliopisto. Populaatioiden geneettisen rakenteen spatiaalinen mallintaminen.

2007 **Outi Ahti-Miettinen, Helsingin yliopisto.** Kaksivaiheisen potenssikiintiöin-nin käyttö otoksen tehostamisessa - Esimerkkinä otoksen suunnittelu työvoimakustannusindeksin tietojen keruulle.

2008 **Paul Catani, Svenska handelshögskolan.** Enhetsrottest och initialvärdet Tillämpning på arbetslösheten i Finland

2009 **Elina Ahola, Jyväskylän yliopisto.** Eksponenttisen perheen tila-avaruusmallien sovellus alkoholikuolleisuusaineistoon Matias Leppisaari, Aalto yliopiston teknillinen korkeakoulu: Tunnustuspalkinto.

2010 **Sanna Peltomäki, Tampereen yliopisto.** Estimation of Below Threshold Intra-EU Trade.

2011–2012 **Tytti Pasanen,** Tampereen yliopisto. Two-Level Structural Equation Modeling with Non-Normal Observed Variables for Assessing Poverty in Laos.

Myönnetyt väitöskirjapalkinnot

2009–2012 **Jukka Sirén,** Helsingin yliopisto. Statistical models for inferring the structure and history of populations from genetic data.

ISSN 0355 – 5941