Tilastollisia perusmenetelmiä yleistajuisesti

REIMANN, CLEMENS, PETER FILZMOSER, ROBERT GARRETT & RUDOLF DUTTER (2008). *Statistical data analysis explained. Applied environmental statistics with R*. 343 s. Wiley, Chichester.

Maantieteellisessä tutkimuksessa käytetään tilastollisia menetelmiä enenevässä määrin. Menetelmien suosio on lisännyt tieteenalakohtaisesti räätälöityjen käsikirjamaisten teosten tarvetta, ja tähän rakoon iskee myös kahden geokemistin (Clemens Reimann ja Robert Garrett) ja kahden tilastotieteilijän (Peter Filzmoser ja Rudolf Dutter) poikkitieteellinen kirja.

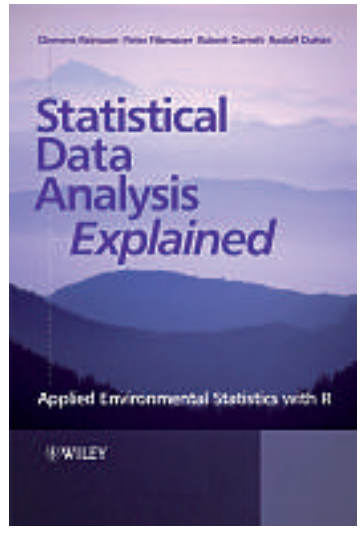
Teos koostuu 19 luvusta. Aluksi lukija johdatetaan esimerkeissä käytettyyn Kuolan ekogeokemia -projektin aineistoon ja taulukkoaineiston muokkaukseen R-ohjelmistoon sopivaksi. Projekti toteutettiin yhteistyönä Geologian tutkimuskeskuksen, Norjan geologisen tutkimuslaitoksen ja Venäjän Keski-Kuolan Expedition kesken. Kotimaisen lukijan mielenkiintoa nostaa varmasti se, että huomattava osa aineistosta on Itä-Lapista. Pääosa kirjasta esittelee yleistajuisesti, mutta paikoin melko yksityiskohtaisesti graafisia ja laskennallisia perusmenetelmiä geokemiallisen esimerkkiaineiston avulla. Teos käsittelee muun muassa tilastollisen jakauman graafisia esitystapoja ja niiden yhdistelmiä, alueellisen aineiston esittämistä kartalla, poikkeavien havaintojen tunnistusmenetelmiä, tilastollista testaamista, aineiston muunnoksia sekä aineiston laadunvalvontaa. Lisäksi kirjoittajat esittelevät monimuuttujamenetelmiä, kuten pääkomponentti-, faktori-, ryhmittely-, regressio- ja erotteluanalyysin. Viimeisessä luvussa käydään lyhyesti läpi R-ohjelmisto ja DAS+R-graafinen käyttöliittymä.

Kirjoittajien mukaan teoksen päätavoitteena on opastaa konkreettisin esimerkein eksploratiivisien data-analyysimenetelmien käytössä ympäristötieteissä. Pääpaino on aineistossa, jolla on alueellinen ulottuvuus. Kirjassa esitellään menetelmien etuja ja haittoja. Erityinen huomio kohdistuu klassisen (gaussilainen) tilastotieteen

Hjort, Jan & Kuusisto-Hjort, Paula (2009). Tilastollisia perusmenetelmiä yleistajuisesti. *Terra* 121:4, 323–324.
© 2021 kirjoittajat. Kirjoitus on lisensoitu Creative Commons Nimeä 4.0 Kansainvälinen (CC BY 4.0) -lisenssillä.

heikkouteen maantieteellisiä ilmiöitä tarkasteltaessa. Tuloksia voivat vääristää muun muassa spatiaalinen autokorrelaatio, epänormaalit jakaumat, monimutkaiset riippuvuussuhteet, poikkeavat havainnot ja aineiston sulkeutuneisuus. Lääkkeeksi ongelmiin tarjotaan aineiston muunnoksia sekä robusteja ja epäparametrisia lähestymistapoja. Puhtaasti geokemiallisista esimerkeistä huolimatta kirjoittajat uskovat teoksen sopivan niin ympäristötieteilijöille, hydrologeille, metsätieteilijöille, ekologeille, maantieteilijöille kuin terveystieteilijöillekin. Lisäksi he mainostavat esipuheessaan kirjaa ainutlaatuiseksi, koska ”se mahdollistaa suoran pääsyn sovelletun ympäristötilastotieteen ohjelmistoratkaisuihin”.

Teosta onkin hyvä arvioida näiden tavoitteiden ja ”mainospuheiden” pohjalta. Kirja etenee johdonmukaisesti yksikertaisista aineiston kuvaamisen tavoista kohti haastavampia monimuuttujamenetelmiä, mutta sen lukeminen kannesta kanteen on hieman puuduttavaa. Teoksen onkin tarkoitus olla paitsi oppikirja myös käsikirjamainen apuväline, jonka tiedoilla kokenutkin käyttäjä voi virkistää muistiaan. Kirjan alkupuoli, jossa esitellään aineiston kuvaamista tunnuslukujen ja kuvien avulla, on tilastotieteen perusteet hallitseville suureksi osaksi tuttua asiaa. Kiinnostavuutta lisäävät kuitenkin perusasioiden lomaan sijoitetut robustien vaihtoehtojen esittelyt.



Aiheiden käsittely ei sellaisenaan tarjoa eväitä menetelmien syvälliseen ymmärtämiseen. Tämä on ollut tekijöiltä tietoinen ratkaisu. Kirja on kohdennettu muille kuin tilastotieteilijöille ja matemaattisten kaavojen viidakkoa on pyritty välttämään. Kirjoittajat painottavat onnistuneesti peruskäyttäjän kannalta oleellisia seikkoja, kuten sitä miten aineiston ominaispiirteet ja menetelmien valinta vaikuttavat analyysien tuloksiin ja tulkintaan. Näitä seikkoja käsitellään usein valitettavan suppeasti tilastotieteen oppikirjoissa.

Teoksen kiinnostavuutta on pyritty lisäämään tukeutumalla todelliseen aineistoon. Tämä on suuri etu, mutta vain, jos lukijalla on geokemistin tausta. Teoksen visuaalinen anti ei hivele maantieteilijän silmää, vaikka ulkoasu ja kuvat ovat yksittäisiä poikkeuksia ja virheitä lukuun ottamatta selkeitä. Syykin on selvä: kuvat ovat pääosin perinteisiä mustavalkoisia tilastollisia kuvaajia ja ne on laadittu R-ohjelmistolla, joka on heikko alusta visuaalisesti palkitsevien esitysten laadintaan. Ohjelmiston puutteet näkyvät varsinkin teoksen kartoissa.

Kirjan vahvuutena pidämme ongelmallisten geokemiallisten ja alueellisten aineistojen tilastollisen analysoinnin esittelyä. Kirjoittajien mielestä klassisen tilastotieteen menetelmät sopivat huonosti ympäristötieteiden aineistojen käsittelyyn. Robusteja vaihtoehtoja esitelläänkin muun muassa poikkeavien havaintojen tulkintaan, tilastolliseen testaukseen, korrelaatioiden laskuun, mutta myös erilaisien monimuuttujamenetelmien, kuten pääkomponenttianalyysin käyttöön. Robustien menetelmien esittely herättää väistämättä lukijan kiinnostuksen, mutta samalla herää kysymyksiä, joihin teos ei vastaa. Aihetta olisi kannattanut käsitellä laajemmin, koska näiden menetelmien tarpeellisuus on yksi kirjan tärkeimmistä viesteistä.

Todennäköinen lukijakunta on otettu hyvin huomioon tarkasteltavia asioita valittaessa. Useimmat ympäristötutkijat joutuvat ennen varsinaisten tilastollisten menetelmien käyttöä tutustumaan kriittisesti analyysitulosten luotettavuuteen ja varmistamaan, että aineisto soveltuu suunniteltuihin tilastollisiin menetelmiin. Teos perehdyttää melko syvällisesti näihin käytännön työssä tärkeisiin esityövaiheisiin, vaikka ei tarjoakaan vakioratkaisuja esille tuleviin ongelmiin. Käsittely antaa melko hyvät valmiudet omien ratkaisujen tekemiseen ja näiden perustelemiseen.

Loppuun kirjoitettua johdantoa R-ohjelmistoon ja DAS+R-graafiseen käyttöliittymään pidämme erittäin hyvänä lisänä – joillekin jopa korvaamattomana apuna. Esimerkiksi kaikkien kirjassa esiintyvien kuvaajien laadintaan ja niiden taustalla olevien analyysien suorittamiseen löytyy R-koodi kirjassa esiteltävästä internetosoitteesta. Tämä mahdollistaa menetelmin tarkastelun lukijan omalla aineistolla.

Teos paneutuu yleistajuisesti tilastotieteen perusasioihin, joten se soveltuu melko hyvin maantieteilijöiden oppi- ja käsikirjaksi. Kirjassa tarjotaan hyvät eväät eksploratiivisten tilastollisten analyysien tekemiseen aineiston ominaispiirteet huomioon ottaen. Erityisesti suosittelemme teosta (geo-)kemiallisten aineistojen kanssa työskenteleville. Varsinaisen hyödyn saaminen edellyttää kuitenkin tilastotieteen perusteiden hallintaa. Kaiken kaikkiaan maantieteellisten aineistojen ongelmallisuus on hyvänä muistutuksena kaikille, jotka tarkastelevat alueellisia ilmiötä tilastollisin menetelmin.

JAN HJORT & PAULA KUUSISTO-HJORT
Maantieteen laitos,
Helsingin yliopisto