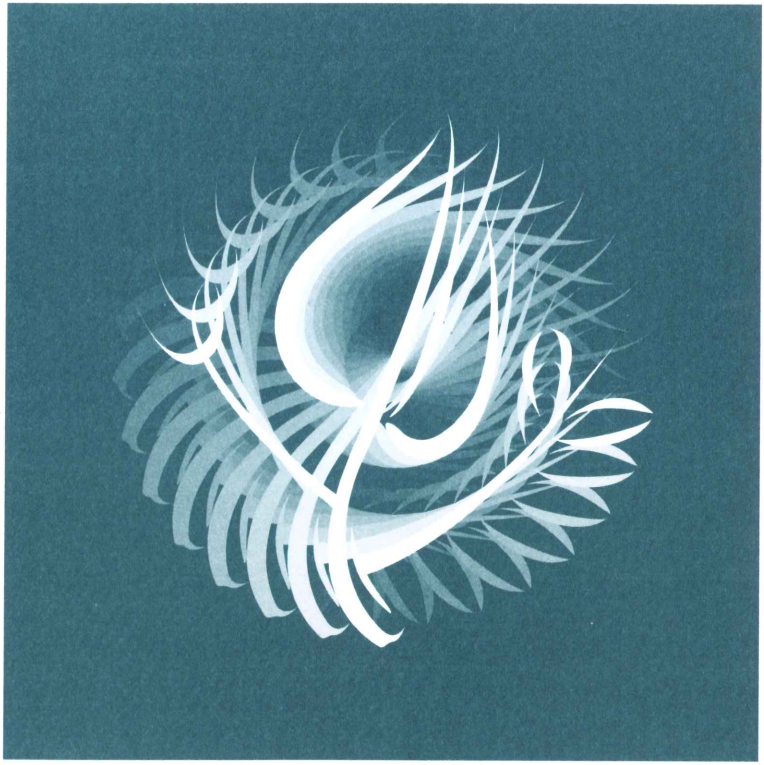
Teabevihik

Tartu 2001

11

EESTI STATISTIKASELTS

STATISTIKAMEETODID KESKKONNAKAITSES JA ÖKOLOOGIAS



11

EESTI STATISTIKASELTS

Eesti Statistikaseltsi Teabevihik

Statistikameetodid keskkonnakaitses ja ökoloogias

Tartu 2001

Toimetus:
Tõnu Kollo, Tõnu Möls, Ene-Margit Tiit, Säde Koskel

Kaanekujundus:
Michael Walsh

ISSN 1406–314X

ISBN 9985–9168–4–0

Tartu Ülikooli Kirjastuse trükikoda
Tiigi 78, Tartu 50410
Tellimus nr. 228

Sisukord

Saateks

Käesolev kogumik sisaldab põhiliselt 16.–17. aprillil 1998 toimunud Eesti Statistikaseltsi 11. konverentsi "Statistikameetodid keskkonnakaitses ja ökoloogias" materjale. Nimetatud konverents, mille korraldamise idee pärines prof. E. Tiidult, äratas laialdast huvi ja kattis mitmeid biomeetria valdkondi. Siiski jäi real põhjustel enamik konverentsi materjale avaldamata. Vahepealsed aastad on aga näidanud, et tookord esitletud biomeetria meetodid ei ole oma väärtust järgnenud aastate jooksul kaotanud. Vastupidi, mitmed konverentsil läbiarutatud ideed on praeguseks leidnud mitmekülgset kasutamist ökoloogilistes ja keskkonnauuringulistes töödes. Nii on Võrtsjärve ökosüsteemi diferentsiaalseostel põhinev mudel olnud mitmete artiklite teemaks ja ka toeks uute uuringute planeerimisel. Eve Aruvee kirjeldatud sesoonsuse modelleerimine β-funktsioonide abil on saanud peaaegu et standardiks Peipsi ja teiste järvede parameetrite muutuste kirjeldamisel suurte statistiliste mudelite abil. Konverentsil esitatud Peipsi hüdrokeemilise seisundi modelleerimise põhimõtted on olnud aluseks eesti- ja ingliskeelsete Peipsi monograafiate vastavate peatükkide ja mitmete artiklite koostamisel ning uute uurimisprojektide planeerimisel.
A. Kiviste metsakasvu uurimise meetodid, rakendatuna ulatuslikule andmebaasile, on rikastanud meie teadmisi ja leidnud kasutamist ka vastavates erikursustes.

Praeguseks on kõiki neid konverentsil esitatud meetodeid edasi arendatud ja täiendatud ning edasine töö jätkub. Mõnede ettekantud ideede areng on olnud sedavõrd dünaamiline, et vastava artikli koos–tamine ei ole tundunud autoritele mõistlik. Näitena võiks tuua T. Eimre ettekande keemiliste parameetrite jaotusest eesti järvedes. Selles ettekandes esitatud ideed keemiliste analüüside lävivääртuste valikust tekitasid konverentsil vastakaid arvamusi, kuid neid on pidevalt edasi arendatud. Siiski ei ole autor siiani temale lõplikult vastuvõetavat lahendust leidnud. Ka H. Timmi tööd põhjaloomastiku kasutamisest vee seisundi hindamiseks on pidevalt jätkunud, kuid valdkonna keerukuse ja diskuteeritavuse tõttu ei ole autor

vastavat eestikeelset artiklit veel koostanud. Seda kõike arvestades võib peagi osutuda vajalikuks korraldada järgmine biomeetria-alane konverents, et tutvustada vahepeal saadud tulemusi ning üheskoos otsida lahendusi raskematele üleskerkinud küsimustele.

Toimetaja T. Kollo on lisanud konverentsi materjalidele ka kaks hiljem kirjutatud tööd ja kroonika. E. Tiidu artiklis statistilisest sõltuvusest on kokku võetud ja näidetega illustreeritud mõned praktilist huvi pakkuvad ja samal ajal teoreetiliselt ilusad tulemused autori pikemaajalisest uurimistööst, mis on äratanud ka suurt rahvusvahelist huvi. Teises artiklis on E. Käärik tutvustanud nn. Gini kordajat, mille kasutamine on perspektiivikas eelkõige sotsioloogilistes ja majanduslikes uuringutes. Kogumiku lõpliku toimetamise ja vormistamise eest on toimetus tänu võlgu Säde Koskelile.

Märts, 2001.

ESS 11. konverentsi organisaator Tõnu Möls.

Eesti Statistikaseltsi 11. Konverents
LUS maja saalis, Struve 2, Tartus

Statistikameetodid keskkonnakaitses ja ökoloogias

PROGRAMM

16. aprill 1998

9.30–10.00. Registreerimine, ESS liikmemaksu tasumine, ESS trükistega tutvumine.

I istung. Taimestiku mudelid.

Juhataja: E.-M. Tiit
10.00. Konverentsi avamine. *E.-M. Tiit*
10.10–10.40. *Juhan Ross, Madis Sulev, Peeter Saarelaid.* Taimkatte kiirgusrežiimi statistiline käsitlus
10.50–11.20. *Vello Ross, Juhan Ross.* Taimkatte arhitektuuri statistilised parameetrid
11.30–12.00. Kohvipaus
12.00–12.30. *Andres Kiviste*. Muutus Eesti metsade kasvukäigus ja selle hindamisest
12.30–13.00. Diskussioon
13.00–14.00. Lõuna

II istung. Veekogude ökosüsteemide modelleerimine

Juhataja: J. Ross
14.00–15.30. *Märt Möls, Tiina Nõges, Peeter Nõges.* Veekogu ökosüsteemi mudel Võrtsjärve näitel
15.30–16.00. Kohvipaus.
16.00. Diskussioon.

17. aprill 1998
III istung. Veekogude seisundi modelleerimine

Juhataja: K. Pärna
10.00–11.40. *Tõnu Möls.* Peipsi järve hüdrokeemilise seisundi modelleerimine
11.40–12.00. Kohvipaus
12.00–12.30. *Henn Timm.* Põhjaloomastik kui keskkonna seisundi indikaator
12.30–13.00. *Eve Aruvee.* Sesoonsuse modelleerimine
13.00–13.30. *Tiina Eimre* Keemilise seisundi parameetrite jaotused eesti järvedes.
13.30. Diskussioon

Sesoonsuse ja selle muutuste modelleerimine

Eve Aruvee
Eesti Põllumajandusülikool, Matemaatika instituut

Ökoloogiliste parameetrite sesoonsuse ja selle pikaajalise muutumise hindamine mitmesuguste segavate faktorite olemasolul ja arvestamisel eeldab suhteliselt keerukat statistilist metoodikat.

Üheks tavalisemaks ja levinumaks teeks on regressioonanalüüs kovariatsioonanalüüsi raamides, kus sesoonsuse modelleerimiseks kasutatakse polünoome.

Näitaja sesoonsus avaldub tema aastaringses reeglipärases muutumises, näiteks südasuvises suurenemises ja talvises vähenemises. Sesoonsuse hindamine keerulisest ökoloogilisest andmestikust on suhteliselt tülikas. Põhilised raskused on seotud prognoosi tsüklilisuse saavutamisega, paljuparameetriliste mudelite hindamisel esinevate arvutuslike probleemidega, samuti vaatlusandmete ebaühtlase jaotusega aasta sees. Seni on selleks kasutatud tavalisi polünoommudeleid. Näitena esitame ühe võimaliku mudeli keemilise ühendi kontsentratsiooni jaoks.

kontsentratsiooni
logaritm $= a_1 + a_2A + a_3A^2 + a_4A^3 + a_5D + a_6D^2 + a_7D^3 + a_8AD +$
$a_9AD^2 + a_{10}A^2D = a_1 + a_2A + a_3A^2 + a_4A^3 + (a_5 + a_8A + a_{10}A^2)D +$
$(a_6 + a_9A)D^2 + a_7D^3.$

Siin ja edaspidi on kasutatud skaleerimisi:

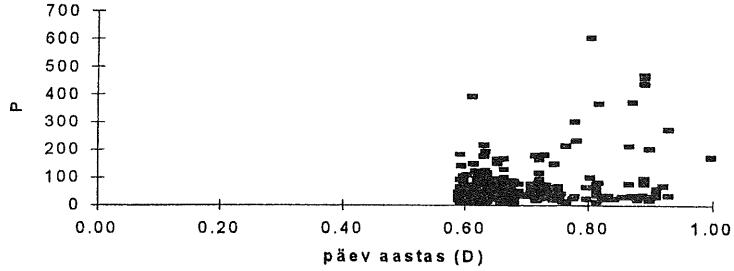
$$A = \frac{\text{aasta} - 1925}{10},$$

ja

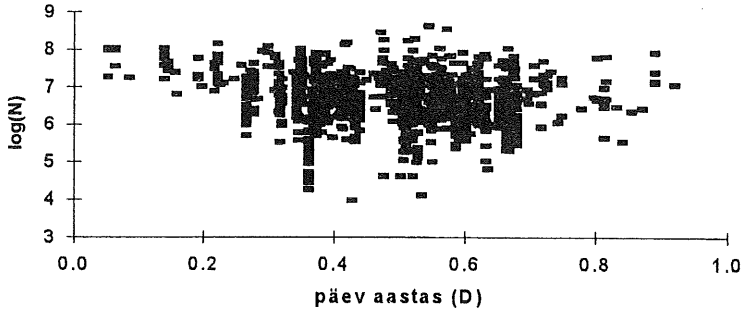
$$D = \frac{\text{päeva järjekorra nr aastas}}{365}.$$

Tunnuse A väärtused on peale skaleerimist lõigus [0,10] ja tunnuse D väärtused on lõigus [0,1]. Fikseerides A, saame konstantsete kordajatega sesoonsuse mudeli. Erinevatele aastatele vastavad kordajad on erinevad, nii saab uurida sesoonsuse sõltuvust aastast.

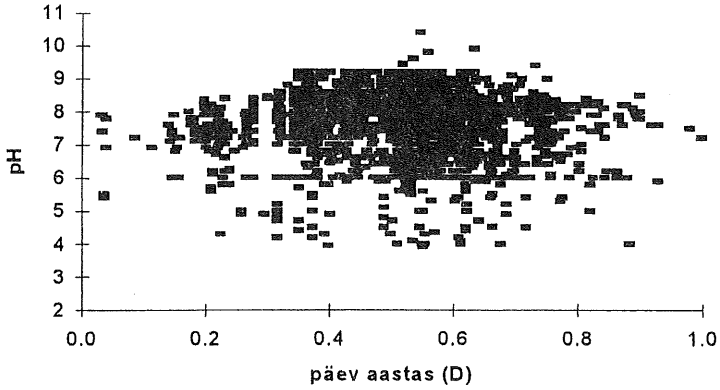
Töö uurimusliku osa aluseks on ZBI Eesti väikejärvede andmebaas, kus vaatlused algavad 1925. aastast. Tänaseks on kogunenud andmeid 350 järve kohta kokku üle 4600 kirje. Kuid see andmestik on lünklik, sest mõõtmisi ei ole teostatud alati mitte kõigis järvedes ja mitte alati ei ole uuritud kõiki aineid. Enamus mõõtmisi on tehtud soojal ajal. Talvisi mõõtmisi on äärmiselt vähe, mistõttu sesoonsuse uurimisel on selle perioodi hinnangud ebatäpsed või vähe usaldatavad.
Paljudest uuritud ainetest on töös analüüsitud üldfosforit, üldlämmastikku ja aluselisust pH. Üldfosforit on uuritud 111 järves ja vaatlusi on kokku 492. Üldlämmastiku andmed on kogutud 189 erinevast järvest, andmekirjeid on kokku 1189. Vesinikioonide kontsentratsiooni pH on mõõdetud kõikidest järvedest ja vaatlusi on 3139. Joonistel 1-3 on esitatud uuritavate ainete hajuvusdiagrammid sõltuvalt vaatlusajast.



Joonis 1. Üldfosfori kontsentratsiooni hajuvus sõltuvalt sesoonist.



Joonis 2. Üldlämmastiku logaritmilise kontsentratsiooni hajuvus sõltuvalt sesoonist.



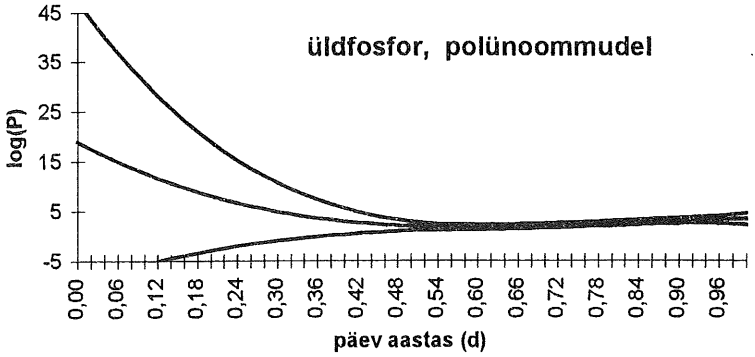
Joonis 3. pH hajuvus sõltuvalt sesoonist.

Eespool kirjeldatud polünoommudelitel on mitmesuguseid puudusi looduslike nähtuste kirjeldamisel. Loetleme neist mõningaid.

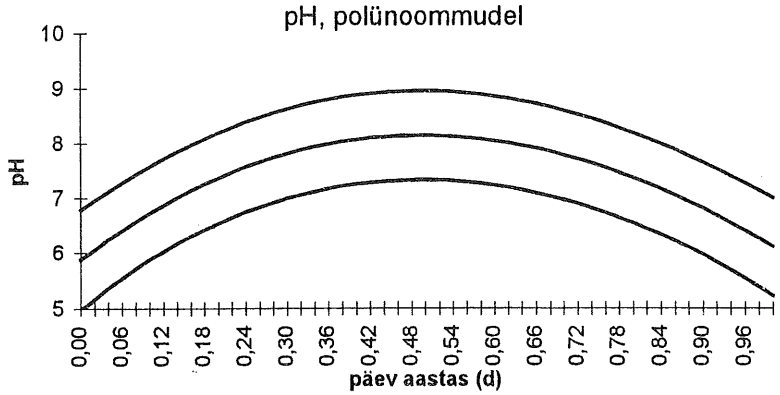
- Mudel ei kajasta hästi üleminekut ühest aastast teise, sest funktsioontunnuse väärtus aasta lõpul ei ühti tema väärtusega uue aasta alguses. Nii näiteks on joonisel 4 prognoositud log(P) tase aasta esimesel päeval 18, aasta viimasel päeval aga 4.

- Mudel prognoosib sageli ebareaalseid muutusi nendes argumentide piirkondades, kus andmeid on vähe. Nii näiteks on üldfosfori mudeliga prognoositud aasta alguse väärtuseks e^{18}, mis kaugelt ületab reaalselt esinevaid väärtusi (vt joon. 4). Sedalaadi mudeli sobimatus on arvatavasti tingitud asjaolust, et astmefunktsioonid kui mudeli komponendid ei ole sobivad looduses esinevate nähtuste kirjeldamiseks.

Esitame nende ainete kontsentratsiooni logaritmi prognoosid koos usalduspiiridega (vt joon. 4-6).

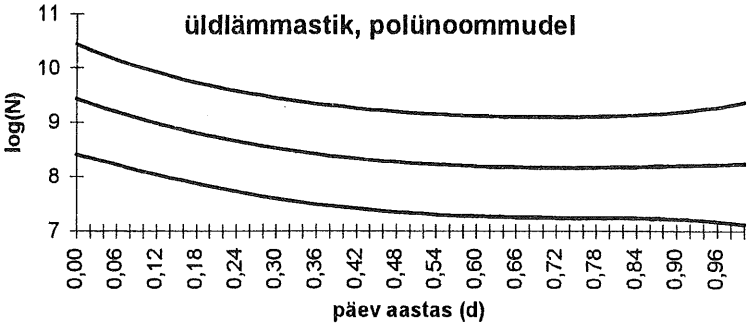


Joonis 4. Üldfosfori kontsentratsiooni sesoonsuse prognoos koos 95-usalduspiiridega.



Joonis 5. pH sesoonsuse prognoos koos 0.95-usalduspiiridega.

Toodud joonistelt on selgesti näha polünoommudelite puudused sesoonsuse modelleerimisel – prognoosi katkemine aasatavahetusel (üldfosfor, üldlämmastik), mittesile üleminek aastavahetusel (pH), kaootiline käitumine andmevaeses ajavahemikus (üldfosfor).

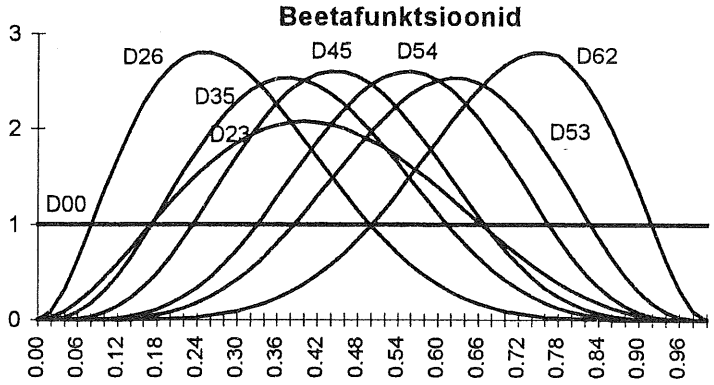


Joonis 6. Üldlämmastiku kontsentratsiooni sesoonsuse prognoos koos 0.95-usalduspiiridega.

Tavaliste polünoommudelite asemel soovitame kasutada beeta-funktsioonide lineaarkombinatsioone. β-funktsiooniks parameetritega p ja q (p, q ≥ 0) nimetatakse lõigul [0,1] määratud funktsiooni

$$f_{p,q}(x) = x^p (1-x)^q \frac{\Gamma(p+q+2)}{\Gamma(p+1)\Gamma(q+1)},$$

kus Γ on gammafunktsioon.



Joonis 7. Mõned sesoonsuse modelleerimisel kasutatavad β-funktsioonid.

Kui soovime konstrueerida niisuguseid sesoonsuse kõveraid, mille väärtus aasta alguses ühtib väärtusega aasta lõpus, siis sobivate β-funktsioonide valik aheneb. Tegelikult tuleb nõuda ka sujuvat üleminekut ühelt aastalt teisele, mis tingib tuletise võrdsuse D = 0 ja D = 1 korral. Neid tingimusi rahuldavad β-funktsioonid, kus p,q =0 või p,q > 1. Mõnede niisuguste β-funktsioonide graafikud on esitatud joonisel 7, kus tähistuse D järel on näidatud parameetrite p ja q väärtused. Käesolevas töös on p ja q väärtused vahemikus 2...6.

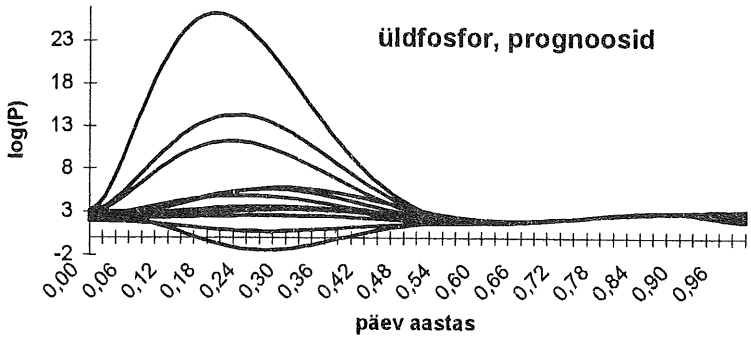
Kirjeldatud metoodika rakendamise näitena vaatame üldfosfori P sesoonsuse modeleerimist, kasutades β-funktsioonil põhinevaid mudeleid. Modelleerime fosfori kontsentratsiooni asemel tema naturaallogaritmi. Näiteks β-funktsioonide D36, D52 ja D62 korral hindame mudelit (edaspidi Mudel 1)

$$\ln(P) = a_1 + a_2A + a_3A^2 + a_4A^3 + a_5AD62 + a_6AD36 + a_7AD52 + a_8A^2D62 + a_9A^2D36 + a_{10}A^2D52 + a_{11}A^3D62 + a_{12}A^3D36 + a_{13}A^3D52.$$

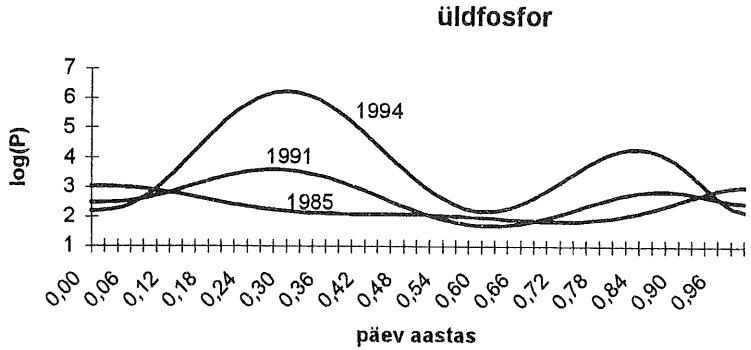
Siin $a_1,\ldots, a_{13}$ on mudeli hinnatavad parameetrid. Mudelis 1 on sesoonsust kirjeldavad liikmed polünomiaalselt sõltuvad teisendatud aastast A. Seega võimaldab antud mudel kirjeldada ka sesoonsuse muutumist aastati. Kui mudelisse valida erinevaid β-funktsioonide komplekte, saame mitmesuguseid alternatiivseid mudeleid.

Üldfosfor on andmestikus aine, mida aastatel 1978 - 1994 on jälgitud ainult juulist kuni novembrini (vt joonis 1). Joonisel 8 on kujutatud erinevate mudelite abil leitud üldfosfori kontsentratsiooni sesoonsuse prognoosid. Graafikult näeme, et toimunud vaatluste piirkonnas ($0.56 < D < 0.92$) prognoosid peaaegu ühtivad, aga piirkonnas, kus vaatlusi ei ole, prognoosivad mudelid erinevalt.

Stabiilses looduslikus keskkonnas ei toimu väga järske muutusi, seega kindlasti ei ole usaldatavad need mudelid, mis prognoosivad kontsentratsiooni väga suurt muutust aasta alguses. Näeme, et kontsentratsioon sügisperioodil $D \in [0.64, 0.88]$ veidi tõuseb. Seda me võime usaldada, sest vastavas ajavahemikus on tehtud reaalseid mõõtmisi ja kõik mudelid prognoosivad selles vahemikus ühesuguse sesoonsuse.

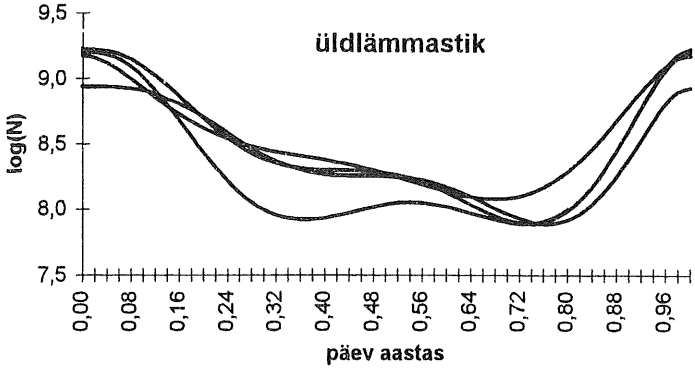


Joonis 8. Erinevate β-funktsioonide komplektidega mudelite abil prognoositud üldfosfori kontsentratsiooni sesoonsus 1991. a. jaoks.

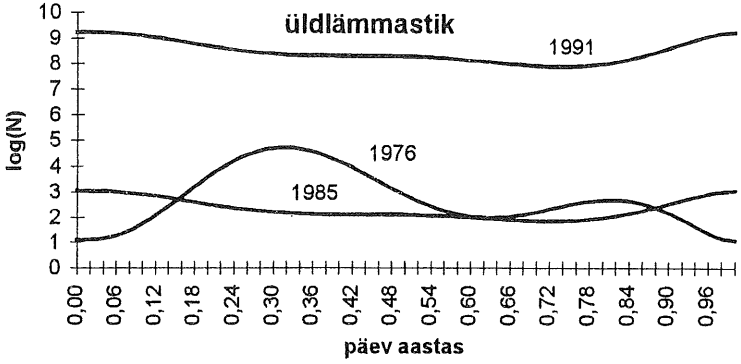


Joonis 9. Üldfosfori kontsentratsiooni sesoonsuse prognoosid kolmel erineval aastal.

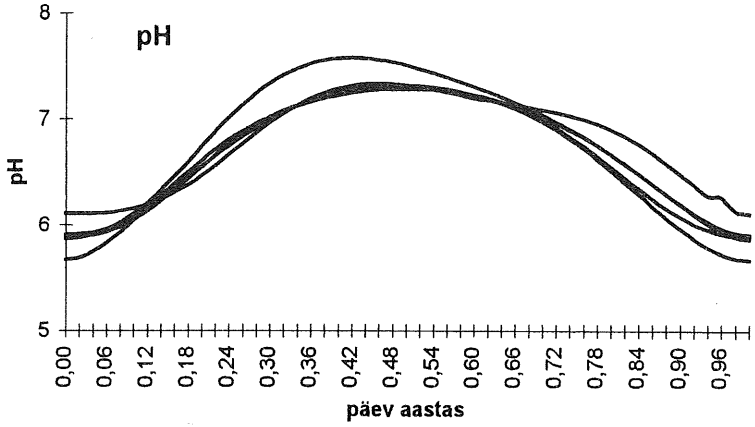
Mudel 1 prognoosib sesoonsust üle kõikide aastate, kasutades SAS GLM protseduuris ESTIMATE-lauset saame leida prognoosid iga konkreetse aasta jaoks (vt joonis 9). Nagu jooniselt näha, on erinevatel aastatel üldfosfori sesoonsus võrdlemisi erinev. Seda asjaolu on võimalik kasutada sesoonsuse pikaajaliste muutuste uurimiseks. Analoogiliselt saame tulemused üldlämmastiku ja pH jaoks (joonised 10 - 13).



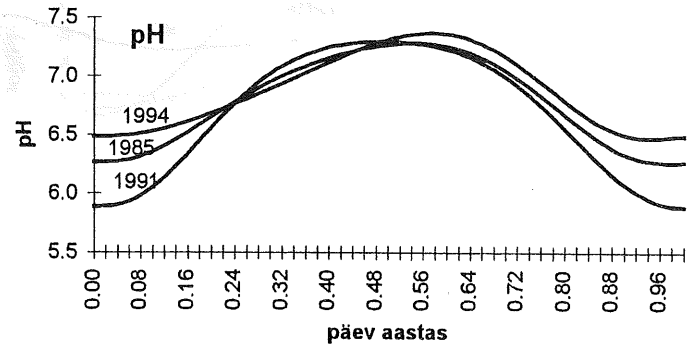
Joonis 10. Erinevate β-funktsioonide komplektidega mudelite abil prognoositud üldlämmastiku kontsentratsiooni sesoonsus 1991. a. jaoks.



Joonis 11. Üldlämmastiku kontsentratsiooni sesoonsuse prognoosid kolmel erineval aastal.

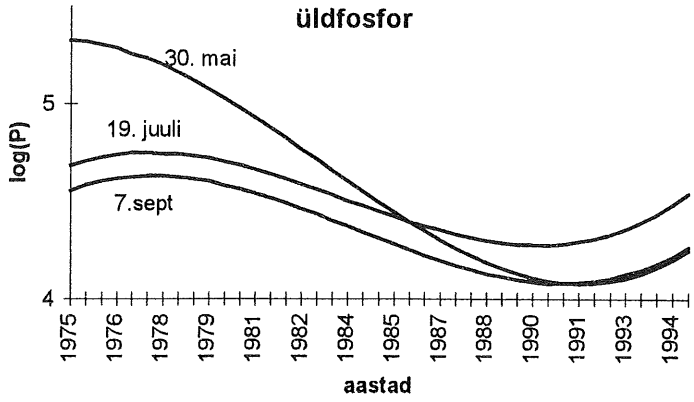


Joonis 12. Erinevate β-funktsioonide komplektidega mudelite abil prognoositud pH sesoonsus 1991. a. jaoks.

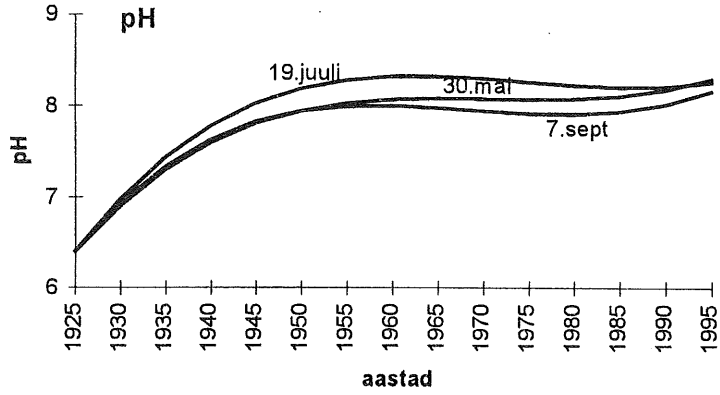


Joonis 13. pH sesoonsuse prognoosid kolmel erineval aastal.

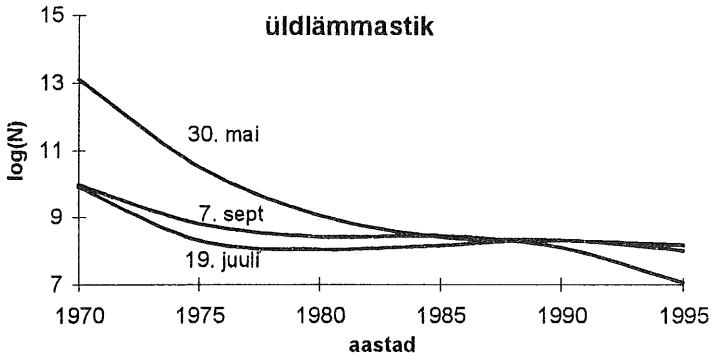
Eelnevat metoodikat kasutades, saame leida ainete kontsentratsiooni prognoosid kindlal päeval aastate jooksul kui muudame päeva konstandiks (vt joonised 14 - 16).



Joonis 14. Üldfosfori kontsentratsiooni ajalised muutused kolme erineva päeva lõikes.



Joonis 15. pH kontsentratsiooni ajalised muutused kolme erineva päeva lõikes.



Joonis 16. Üldlämmastiku kontsentratsiooni ajalised muutused kolme erineva päeva lõikes.

Eespool eeldasime, et punktides 0 ja 1 on prognoosifunktsiooni tuletis null, st graafiku joon nendes punktides on horisontaalne. See tingimus sunnib graafikut vastavalt pöörduma isegi siis, kui see pole andmetega kooskõlas. Tegelikult võib arvata, et ka aastavahetusel on aine kontsentratsioon tõusva või langeva iseloomuga.

Kirjeldatud ebaloomuliku piirangu vältimiseks arvestame, et teataval ajal (kuid üldiselt mitte 1. jaanuaril) on regressioonifunktsiooni graafik horisontaalne. Seega võiks lähendada seda funktsiooni nihutatud β-funktsioonidega $f^{\circ}_{p,q}(x)$, kus

$$f^{\circ}_{p,q}(x) = \begin{cases} f_{p,q}(x - D), \text{ kui } D \le x \le 1 \\ f_{p,q}(x + 1 - D), \text{ kui } x < D \end{cases}$$

ja $0 \le D \le 1$ on täiendav parameeter, nn nihe, mis ei sõltu parameetritest p ja q. Mudeli hindamine sisaldab seega lisaks teiste parameetrite hindamisele ka nihke hindamist. Esitame selleks järgmise algoritmi:

1° Hinnata regressioonifunktsioon eelpool kirjeldatud metoodika järgi, st tuletis on null lõigu [0,1] otspunktides.

2^{o} Leida saadud regressioonifunktsiooni tuletisfunktsiooni nullkoht (ekstreemumpunkt) D lõigu [0,1] sees.

3^{o} Teostada algandmetega nihe $x \rightarrow x + D$ (x rollis D).

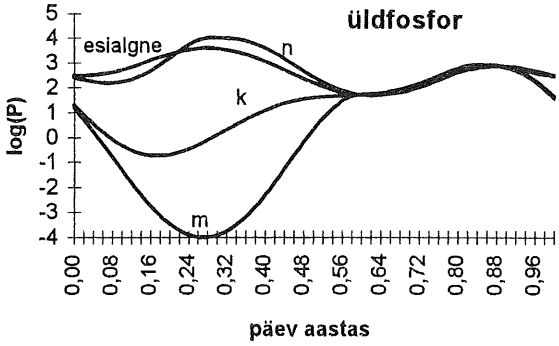
4^{o} Korrata sammu 1.°

5^{o} Nihutada leitud regressioonifunktsioon x-teljel tagasi (D võrra).

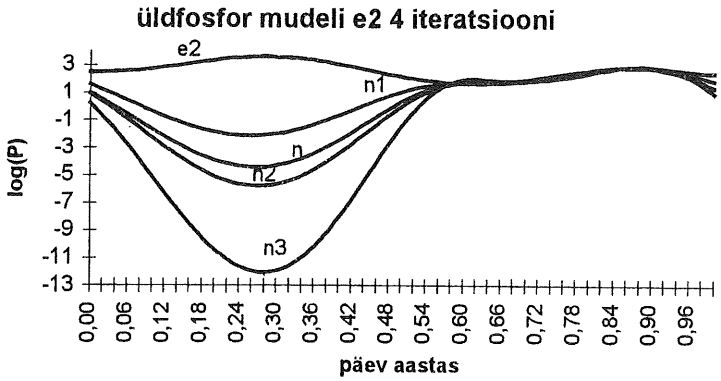
Võttes aluseks Mudeli 1, on prognoosifunktsioonil kolm ekstreemumpunkti, seega saame ka vastavalt kolm prognoosi (vt joon. 17).

Joon n on ekstreemumpunktile 0.29617 vastav regressioonifunktsioon, joon m on ekstreemumpunktile 0.61665 vastav regressioonifunktsioon ja joon k on ekstreemumpunktile 0.8522 vastav regressioonifunktsioon.

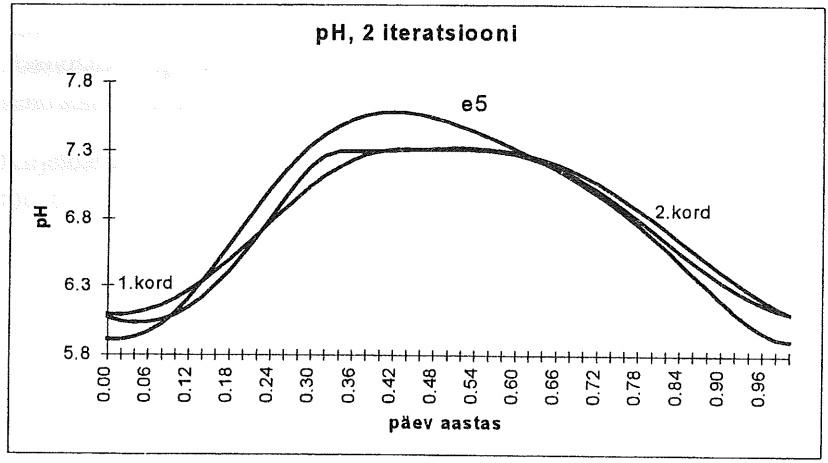
Kirjeldatud protseduuri võib rakendada korduvalt. Kuid mudelite korduval hindamisel kerkib üles ekstreemumite identifitseerimise probleem. Nimelt tuleks regressioonifunktsiooni hindamisel püüda kasutada kogu aeg identset eksreemumit. Siin peab uurija tõenäoliselt hoidma hindamisprotsessi oma kontrolli all.



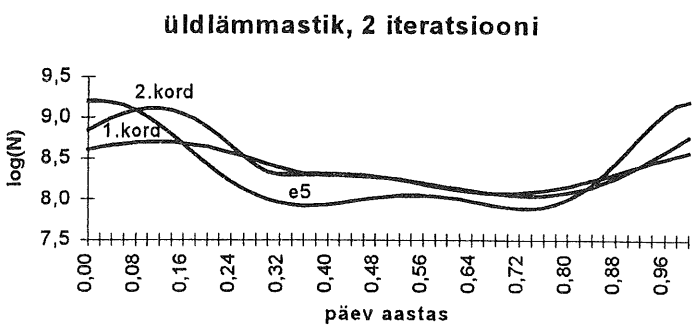
Joonis 17. Prognoosid, kui aluseks on võetud erinevad ekstreemumpunktid.



Joonis 18. Sesoonsuse graafiku muutumine 4 iteratsiooniga. Ekstreemumid on valitud D = 0.62 ümbruses.

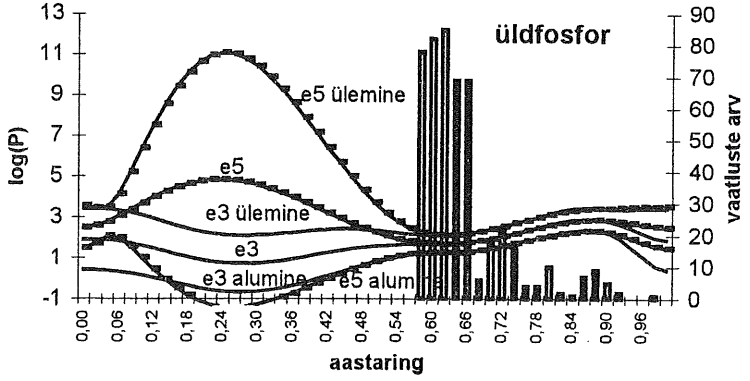


Joonis 19. pH sesoonsuse prognoosid mudeli e5 ja ekstreemumpunktidele 0.61665 ja 0.49305 sobitatud mudelite alusel.



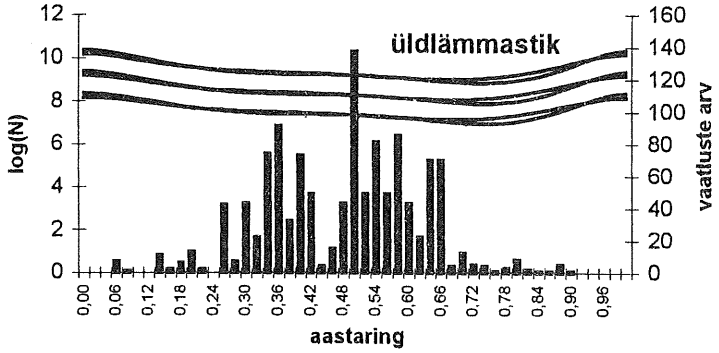
Joonis 20. Üldlämmastiku kontsentratsiooni sesoonsuse prognoosid mudeli e5 ja ekstreemumpunktidele 0.29617 ja 0.3333213 sobitatud mudelite alusel.

Kokkuvõttes võib öelda, et sesoonsuse hindamine nihutatud β-funktsioonide abil võib anda realistlikuma tulemuse kui tavaliste β-funktsioonide kasutamisel. Arvutusprotsessi on aga suhteliselt keeruline kontrollida. Seepärast on põhjust edaspidistes uuringutes otsida mõnda teist meetodit, mis asendaks β-funktsioonide nihutamist.



Joonis 21. Mudelitega e3 ja e5 hinnatud sesoonsuse prognoosid ja usalduspiirid üldfosfori kontsentratsiooni jaoks koos vaatluspäevade empiirilise jaotusega algandmestikus.

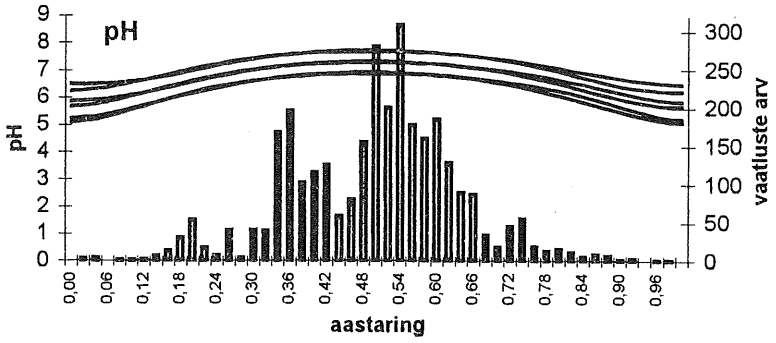
SAS GLM protseduuri ESTIMATE-lause võimaldab arvutada regressioonifunktsiooni argumendi erinevate väärtuste jaoks tunnuse prognoosid ja nendele vastavad usalduspiirid.
Uurimise käigus selgus, et sesoonsuse usalduspiirid sõltuvad vaatluste aastaringsest jaotusest. Selle kujukaks näiteks on üldfosfori andmestik. Kui vaatlusi on mingis piirkonnas vähe (vt joon 21), siis usalduspiirkonna kuju ja laius sõltuvad mudelist ning sellest, missugune käik on regressioonijoonel piirkonnas, kus vaatlusi on piisavalt. Määravaks võib osutuda regressioonijoone käitumine andmevaese piirkonna ja andmerikka piirkonna vahelisel üleminekualal. Joonisel 21 on kahe mudeli - e3 ja e5 - prognoosid koos oma 0.95-usalduspiiridega. Näeme, et mudeli e5 prognoosil on andmevaesel piirkonnal palju laiem usaldusvahemik kui mudeli e3 prognoosil. Piirkonnas, kus on tehtud piisavalt reaalseid mõõtmisi, on mõlema mudeli usalduspiirkonnad kattuvad.



Joonis 22. Kahe erineva mudeliga (e1 ja e2) hinnatud sesoonsuse prognoosid ja usalduspiirid üldlämmastiku kontsentratsiooni jaoks koos vaatluspäevade empiirilise jaotusega algandmestikus.

Üldlämmastiku ja pH korral on vaatlusi kogu aasta ulatuses, talvel küll suhteliselt vähe (vt joon 22 ja 23). Nendel joonistel on ka üldlämmastiku ja pH kahe erineva mudelitega hinnatud sesoonsuse prognoosid ja usalduspiirid 1991. aasta jaoks.

Tulemuste õigsus sõltub mudeli valikust. Kui anname ette mudeli valesti, st antud mudel ei kirjelda protsessi õigesti, siis selle mudeliga saadud tulemused võivad olla vastuolus mõne teise mudeliga saadud tulemustega. Vastuolu võib tähendada seda, et D mingi väärtuse korral usalduspiirkonnad ei lõiku, on ühisosata (ühisosa on tühi). Kui andmed on modelleeritud, siis saame öelda, kas mudel kajastab neid õigesti või ei. Reaalsete andmete korral seda ei saa. Kaks sesoonsuse mudelit on praktiliselt kooskõlas siis, kui nendele vastavate prognooside usalduspiirkonnad on lähedased (nagu joonistel 22 ja 23).



Joonis 23. Kahe erineva mudeliga (e1 ja e2) hinnatud sesoonsuse prognoosid ja usalduspiirid pH jaoks koos vaatluspäevade empiirilise jaotusega algandmestikus.

Kokkuvõtteks.

Üldiselt on β-funktsioonid sobivateks komponentideks sesoonsust kirjeldava regressioonifunktsiooni koostamisel. Nende abil on võimalik saada aastast aastasse sesoonsust kirjeldavat pidevat funktsiooni, seega loomuliku muutumisprotsessi. Protsessi muutumise käiku aastast aastasse on eriti sobiv kirjeldada nihutatud β-funktsioonidega, siis saame ka aastavahetuspunktis tõusva või langeva funktsiooni. Mudelite valikul ja hindamisel kollineaarsus ei mängi olulist rolli.

Kui algandmestikus on andmeid kogu aasta ulatuses, siis mudeli usalduspiirid ei sõltu oluliselt mudeli valikust ja vaid sel juhul on võimalik saavutada mudelite vahelist kooskõla.
Usalduspiirid sõltuvad vaatlusaegade empiirilisest jaotusest. Aastaringi andmevaeses osas võivad sesoonsuse prognoos ja usalduspiirid olla väga erinevad. Sellise piirkonna jaoks leitud prognoose tuleb kasutada ettevaatusega.

Kirjandus

1.SAS Institute Inc., SAS/STAT User's Guide, Version 6, Fouth Edition, Volume 2, Cary, NC: SAS Institute Inc., 1989. 846 pp
2.SAS Institute Inc., SAS/STAT User's Guide, Version 6, Fouth Edition, Volume 1, Cary, NC: SAS Institute Inc., 1989. 943 pp

Muutus Eesti metsade kasvukäigus ja selle hindamine

Andres Kiviste
Eesti Põllumajandusülikool, Metsakorralduse instituut

1. Sissejuhatus

Enamiku puistu kasvukäigu mudelite koostamisel on siiani vaikimisi eeldatud kasvutingimuste (kliima, mullaviljakus, -niiskusrežiim, süsihappegaasi sisaldus õhus jms) muutumatust ajas. Reaalsete puistute kasvutingimused aga ei ole muutumatud, vaid võivad puistute elüea jooksul inimtegevuse (kuivendamine, väetamine, kütuste põletamine, keskkonna saastamine, hooldusraied, metsakultiveerimine jne) ja puhtlooduslike protsesside (kliimamuutused, mullatekkeprotsess, konkureerivate liikide kasv jne) mõjul arvestatavalt muutuda. Vastavalt sellele võib muutuda ka puistute kasv. Esimest korda täheldas Eesti puistute kasvu kiirenemist professor Artur Nilson 1973.-1977. a. metsakorralduse takseerkirjelduste analüüsimisel (Nilson, Kiviste, 1984). Samal aastal ilmus ka rahvusvahelises metsanduslikus kirjanduses esimene viide metsa kasvu kiirenemisele Soome reservaatpuistute näitel (Hari *et at*, 1984). Hüpotees metsa kasvu kiirenemist esitati ka metsamonitoringu konverentsil Kaunases (Нильсон, Кивисте, 1986), kuid tollal ei olnud meil piisavalt empiirilisi andmeid selle tõestamiseks. Viimasel ajal on publitseeritud mitmeid artikleid puistute kasvu kiirenemisest. Näiteks võrreldi Lõuna-Rootsi kuusikute püsiproovitükkidel kahe viimase metsapõlvkonna andmeid, millest selgus, et viimase metsapõlvkonna tootlikkus oli 40% suurem kui samal kohal asunud eelmises metsapõlvkonnas (Eriksson, Johansson, 1993). Rootsi statistilise metsainventeerimise andmete analüüsimisel selgus, et viimase 40 aasta jooksul on nii puistute kõrgus kui ka läbilõikepindalade summa suurenenud 0,4...0,8% aastas (Elfving, Tegnhammar, 1996). Euroopa Metsainstituudi poolt läbiviidud uurimisprojekti tulemustes täheldati metsa kasvu kiirenemist enamuses Euroopa riikides (Spiecker *et al*, 1996).

2. Algandmed

Parim materjal puistu kasvu uurimiseks oleks metsa püsiproovitükkidel pika ajavahemiku jooksul perioodiliselt mõõdetud andmed. Kahjuks on Eestis metsa püsiproovitükke vähe ning nendel vähestelgi subjektiivselt valitud proovialadel on enamasti vaatlusseeriad lühikesed või metoodiliselt puudulikud. Teine võimalus puude kasvu uurimiseks oleks puutüvede aastarõngaste analüüs. Puutüvede analüüse on tehtud Eestis paljude uurijate poolt, kuid suur osa neist on kaotsi läinud ning ka enamus allesjäänud tüveanalüüsidest on metoodiliselt sobimatud puistu kasvukäigu analüüsimiseks.

Kõige usaldusväärsem info Eesti metsade kohta on metsakorralduse takseerkirjeldustes. Eesti riigimetsas tehakse traditsiooniline lausmetsakorraldus kümneaastase perioodi järel, mille käigus hinnatakse (takseeritakse) silmamõõduliselt iga puistu takseertunnused (vanus, koosseis, kõrgus, diameeter, tagavara jne). Metsakorralduses kasutatav silmamõõduline takseerimismetoodika on kogu sõjajärgsel perioodil olnud enam-vähem ühesugune, mistõttu erinevate aastakümnendite andmeis ei tohiks olla sellest tingitud süstemaatilisi erinevusi. Riigimetsade viimase korralduse takseerkirjeldused on salvestatud andmebaasifailidesse. Varasemate metsakorralduste takseerkirjelduste andmed on ainult takseerkirjelduste raamatutes.

Antud uurimuses on valitud 23 metskonna (Järvselja, Vahastu, Peedu, Kambja, Alatskivi, Pühajärve, Otepää, Kärkna, Vesneri, Tähtvere, Erastvere, Kubja, Orava, Põlva, Rõuge, Vastseliina, Käsmu, Kunda, Loobu, Porkuni, Sagadi, Sõmera ja Vihula) 1950-ndate (esimese sõjajärgse) ja 1990-ndate (viimase) metsakorralduse takseerkirjelduste andmed. Nende kahe erineva metsakorralduse andmete sidumiseks koostati puistuplaanide põhjal fail, kus on kõigi puistute kvartali ja eralduse numbrid nii 1950-ndate kui ka 1990-ndate aastate tähistuses. Vaatlusaluse nelja aastakümne jooksul on metsaeralduste piirid paljuski muutunud, mistõttu vajaduse korral jagati 1990-ndate metsakorralduse eraldused väiksemateks osaeraldusteks. Selle tulemusena tekkinud mosaiiksest kaardimaterjalist korrektse andmestiku loomine osutus väga suurt tähelepanu ja täpsust nõudvaks tööks. Kokku ana-

lüüsiti 56958 puistut 72206 hektaril, mis moodustab 3,6% Eesti metsamaa pindalast.

Vaadeldavate 1950-ndate ja 1990-ndate aastate metsakorralduste välitööde tehnoloogia (puistute silmamõõduline takseerimine) ei ole oluliselt muutunud. Antud uurimuse seisukohalt on tähtsamad järgmised erinevused.

1. 1950-ndate aastate takseerkirjeldustes puudub tänapäeva metsanduses väga oluline klassifitseeriv tunnus – puistu kasvukohatüüp.
2. 1990-ndate aastate takseerkirjeldustes on tunduvalt rohkem tunnuseid kui 1950-ndate aastate takseerkirjeldustes ning nad on sisestatud andmebaasifailidesse. 1950-ndate aastate takseerkirjelduste raamatud asuvad Sagadi Metsamuuseumis ning vajaminevad andmed tuli eeldtööde käigus arvutisse sisestada.
3. Mõnede perioodide metsakorraldustes ei ole välistatud teatud süstemaatilised vead puistute takseertunnuste silmamõõdulisel hindamisel. Nii näiteks on täheldatud puistu tagavara ja vanuse mõningast allahindamist 1970.-1980. aastatel. Vaatlusaluste metsakorralduste kohta ei ole teada, et andmeid oleks moonutatud.

3. Metoodika

Puistu kasvukoha headuse näitajana kasutatakse metsanduses nn puistu boniteeti, mis on puistu vanuse ja kõrguse funktsioon. Eesti metsanduse praktikas on siiani kasutatud Orlovi boniteerimistabeleid, mida on tülikas arvutis rakendada. Põhjamaade metsanduslikus praktikas kasutatakse kasvukoha headuse näitajana nn kasvukoha indeksit, mis on puistu kõrguse prognoos teatud baasvanuseks (näit 50 aastat) antud kasvukohale sobiva kasvufunktsiooni järgi. Selles uurimuses on kasutatud puistute boniteerimiseks Orlovi kohmaka tabeli asemel puistu kõrguse kasvufunktsiooni diferentsvõrrandi näol, mis lisaks puistu vanusele ja kõrgusele arvestab argumentidena ka enamuspuuliiki ja kasvukohatüüpi iseloomustavat mulla kõduhorisondi tüsedust (Kiviste, 1997). Selle mudeli järgi saab arvutada puistu kõrgust H_2 suvalises vanuses A_2 valemiga

$$H_2 = \frac{H_1 + d + r}{2 + \frac{4 \cdot \beta \cdot A_2^{-b_2}}{H_1 - d + r}},$$

kus H_1 – puistu teadaolev kõrgus teatud vanuses A_1,

$\beta = b_1 - 493 \cdot \ln(OHOR + 1)$,

$d = \beta / 50^{b_2}$,

$$r = \sqrt{(H_1 - d)^2 + \frac{4 \cdot \beta \cdot H_1}{A_1^{b_2}}},$$

OHOR – puistu kasvukohta iseloomustav mulla kõduhorisondi tüsedus, cm (muutub alates 1 cm parimates kasvukohtades kuni 50 cm rabas),

b_1, b_2 – peapuuliigist sõltuvad mudeli parameetrid (Tabel 1).

Tabel 1

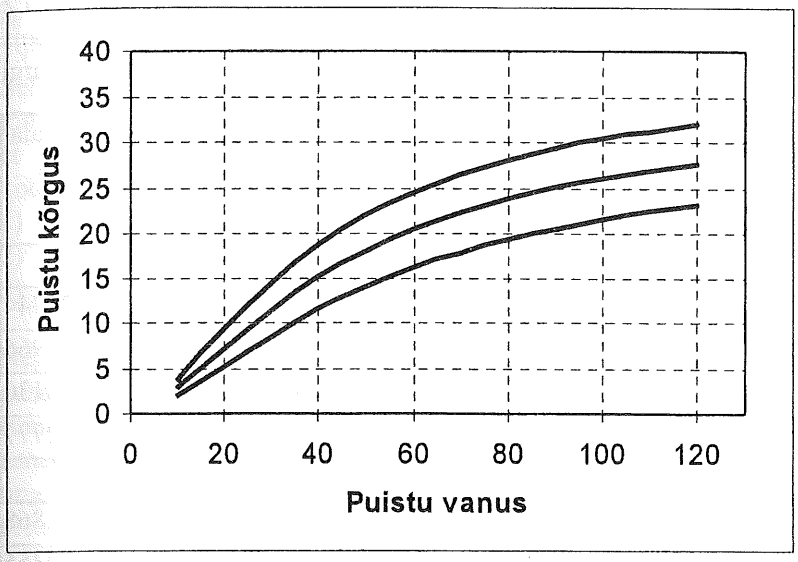
Eesti puistute kõrguse diferentsvõrrandi parameetrite hinnangud

Parameeter	Mänd	Kuusk	Kask	Haab	Sanglepp	Hall lepp	Tamm	Saar
b1	8319	12867	4990	3882	4228	2749	6742	3732
b2	1.58	1.71	1.48	1.3	1.41	1.38	1.61	1.35

Esitatud diferentsvõrrandi parameetrite hinnangud on arvutatud 1984.-1993.a. metsakorralduste kogu Eesti riigimetsa takseerkirjelduste andmeil (kokku üle 500 tuhande puistu). Antud töös kasutatud puistu kõrguse kasvukäigu mudeli illustreerimiseks on joonisel 1 esitatud näide männikute kõrguse kasvukõveratest mustika kasvukohatüübis.

Selleks, et hinnata muutust eesti puistute kasvukäigus, arvutati igale andmestikus olevale puistule kõrguse prognoosid 50 aasta vanuseks (boniteet) nii 1950-ndate kui ka 1990-ndate aastate vanuse ja kõrguse andmeil ($H50_{1950}$ ja $H50_{1990}$). Muutust puistute kasvukäigus viimase 40 aasta jooksul käsitleti kui puistu kõrguse prognoosi 50 aasta vanuseks H50 (boniteedi) muutust 1950-ndatest aastatest 1990-ndate aastateni ($H50_{1990}$ - $H50_{1950}$). Järgneva analüüsi käigus püüti selgusele jõu-

da, kas Eesti metsades on viimase 40 aasta puistute boniteet paranenud ning millised faktorid boniteedi muutumist oluliselt mõjutavad.



Joonis 1. Näide Eesti männikute kõrguse kasvukäigust mustika kasvukohatüübis boniteetide H50 = 14, 18 , 22 m korral.

4. Andmete eelanalüüs

1950-ndate ja 1990-ndate aastate puistuplaanide võrdlemisel selgus, et vahepealse nelja aastakümne jooksul on puistuealduste piirid märgatavalt muutunud. Takseerkirjelduste võrdlemisel osutus, et lisaks metsapõlvkondade vaheldumisele on toimunud suuri muutusi ka sama põlvkonna puistute koosseisus (Tabel 2). Muutused sama metsapõlvkonna kooseisus võivad olla põhjustatud nii liikidevahelisest konkurentsist kui ka hooldusraietest, mille käigus vähemväärtuslikumad puuliigid on välja raiutud. Asjaolu, et maharaiutud männikud on taas kultiveeritud enamasti männiga ja teised puistud enamasti kuusega, peegeldab Eestis rakendatud metsapoliitikat, millega püüti rajada eelkõige okaspuumetsi. Looduslikult uuenenud puistute liigilise jaotuse vaatlemisel (Tabel 2), saab veelkord kinnitust tõsiasi, et lageraielangid

uuenevad looduslikult enamasti kase- või väheväärtusliku lehtpuumetsaga.

Tabel 2
Analüüsitud puistute pindala jaotus
enamuspuuliigi, metsapõlvkonna ja päritolu järgi

Enamus-puuliik 1950-ndail	Enamuspuuliik 1990-ndail				Pindala ha
	Mänd	Kuusk	Kask	Teised	
Sama metsapõlvkond					
Mänd	**91%**	4%	4%	0%	28744
Kuusk	10%	**75%**	11%	4%	8966
Kask	12%	14%	**64%**	10%	10130
Teised	3%	17%	35%	**46%**	2777
Uus metsapõlvkond (kultiveeritud)					
Mänd	**83%**	15%	1%	0%	1868
Kuusk	15%	**72%**	12%	1%	2356
Kask	5%	86%	**8%**	1%	1368
Teised	2%	90%	4%	**4%**	285
Uus metsapõlvkond (loodustekkeline)					
Mänd	**42%**	14%	40%	4%	2065
Kuusk	4%	**29%**	48%	19%	2057
Kask	1%	9%	**68%**	23%	2226
Teised	0%	4%	48%	**48%**	709
Pindala ha	31380	14360	13696	4116	63550

Lisaks muutustele metsaeralduste piiride, põlvkonna ja kooseisu osas leidus andmestikus vasturääkivusi puistu kasvukohatüübi, kuivendatuse, vanuse ja teiste tunnuste osas, mis enamasti võivad olla seletatavad puistute loodusliku arenguga või siis metsamajandusliku tegevuse tulemusega, ehkki täiesti välistada ei saanud ka vigu takseerkirjeldustes või andmesisestuses.

Puistu boniteedi H50 arvutamiseks kasutatavat diferentsmudelit saab rakendada määramispiirkonnas alates 10 aasta vanusest lehtpuu- ja alates 20 aasta vanusest okaspuumetsas. Seetõttu sai boniteedi muutuse hindamiseks arvesse võtta ainult neid metsaeraldusi, mis olid nii 1950-ndail kui ka 1990-ndail mudeli rakendamiseks sobiva vanusega. Selleks, et hinnata (kaaluda) iga metsaeralduse sobivust andmestikku, moodustati kolm heuristilist lisatunnust, mis väljendaksid tõenäosust, et vaadeldava eralduse andmed 1950-ndail ja 1990-ndail on kooskõlas:

1) eralduse pindala osas,
2) puistu vanuse osas,
3) puistu liigilise koosseisu osas.

Nende sobivustunnuste abil rakendati antud andmestikule erineva rangusastmega filtertingimusi mittevõrreldavate metsaeralduste elimineerimiseks. Erinevate valimite põhjal arvutatud tulemused osutusid siiski üsnagi ühesugusteks. Järgnevalt esitatud tulemused on saadud filtertingimustel, kus kõigi kolme heuristilise sobivustunnuse väärtused ületasid 0,5.

5. Tulemused

Pärast sobimatute metsaeralduste filtreerimist allesjäänud 21409 metsaeralduse andmeil koostatud tinglike keskmiste ja standardhälvete tabelist (Tabel 3) selgub, et viimase 40 aasta jooksul on Eesti metsade kasv märgatavalt kiirenenud. Puistute boniteet H50 on viimase nelja aastakümne jooksul suurenenud keskmiselt üle 2 meetri, st ligikaudu 5 cm aastas. Eesti peamiste puuliikide osas oli boniteedi keskmine suurenemine kõige suurem kaasikuis (2,5 m) ja kõige tagasihoidlikum kuusikuis (1,6 m). Männikute boniteedi suurenemine langes kokku üldkeskmisega (2,1 m). Tinglike keskmiste ja standardhälvete tabelis (Tabel 3) selgub ootuspäraselt, et kuivendatud metsaaladel on boniteedi kasv (2,8 m) tunduvalt suurem kui kuivendamata metsaaladel (2,0 m). Samas tuleb tõdeda, et analüüsitav muutuja – boniteedi paranemine viimase 40 aasta jooksul ($H50_{1990}$ - $H50_{1950}$) – on väga suure varieeruvusega (standardhälve 2,9 m).

Tabel 3

Puistu boniteedi H50 suurenemise (viimase nelja aastakümne jooksul) aritmeetilised keskmised, valimi mahud ja standardhälbed kuivendatud (kraavitatud) ja kuivendamata (ülejäänud) metsaaladel

Enamus-puuliik	Kuivendamata metsad	Kuivendatud metsad	Kõik kokku
Keskmine boniteedi suurenemine $H50_{1990}$-$H50_{1950}$ meetrites			
Mänd	2.03	3.11	**2.13**
Kuusk	1.66	1.55	**1.65**
Kask	2.40	2.64	**2.47**
Teised	2.50	2.73	**2.54**
Kõik kokku	**2.02**	**2.75**	**2.11**
Metsaeralduste arv			
Mänd	11641	1206	**12847**
Kuusk	4066	269	**4335**
Kask	2556	1050	**3606**
Teised	535	86	**621**
Kõik kokku	**18798**	**2611**	**21409**
Boniteedi suurenemise $H50_{1990}$-$H50_{1950}$ standardhälve m			
Mänd	2.82	3.26	**2.88**
Kuusk	2.61	2.89	**2.63**
Kask	3.19	2.94	**3.12**
Teised	2.73	2.20	**2.66**
Kõik kokku	**2.84**	**3.10**	**2.88**

Boniteedi suurenemist mõjutavate faktorite uurimiseks valiti metsakorralduse takseerkirjelduste andmetest järgmised tunnused: enamuspuuliik, metsapõlvkond, metskond, kuivendatus ja tekkeviis (diskreetsed tunnused) ning puistu vanus ja mulla kõduhorisondi tüseduse OHOR logaritmiline teisendus (pidevad tunnused). Rakendusstatistika programmi SAS üldistatud lineaarmeetodite protseduuri GLM abil saadi boniteedi H50 suurenemist kirjeldav mudel (Tabel 4).

Tabel 4

Boniteedi H50 suurenemist kirjeldava mudeli dispersioon-analüüsi tabel ja faktorite mõjude dispersioonalüüsi tabel

Valimi maht			21409	
Funktsioontunnus: Boniteedi H50 suurenemine ($H50_{1990}$-$H50_{1950}$)				
Dispersioonanalüüsi tabel				
Varieeruvuse allikas	Vabadus-astmete arv	Hälvete ruutude summa	Keskruut	F
Mudel	33	24763	750.4	105
Viga	21375	153064.4	7.16	
Üldine	21408	177827.05		
R^2	Jääk-standard-hälve	Keskmine		
0.14	2.68	2.11		
Faktorite mõjude dispersioonanalüüsi tabel (tüüp III)				
Faktor	Vabadus-astmete arv	Hälvete ruutude summa	Keskruut	F
Puuliik	3	1090	363	51
Põlvkond	1	1061	1061	148
Metskond	22	15121	687	96
Kuivendus 1950	1	478	478	67
Kuivendus 1990	1	576	576	80
Tekkeviis 1950	1	256	256	36
Tekkeviis 1990	1	114	114	16
Vanus 1950	1	779	779	109
LN(OHOR+1)	1	644	644	90

Tabelis 4 esitatud mudel kirjeldab puistu boniteedi H50 suurenemise varieeruvusest (dispersioonist) 14%, mistõttu mudeli kirjeldusvõime on üsna tagasihoidlik (jääkstandardhälve 2,7 m). Boniteedi H50 suurenemist mõjutab kõige enam puistu vanus (koos tunnusega, kas 1950-ndail ja 1990-ndail on sama või erinev metsapõlvkond). Ka ülejäänud mudelis esindatud faktorid on olulised olulisuse nivool 0,001.

Antud faktorite mõjud boniteedi H50 suurenemisele oli üldiselt ootuspärased. Tunduvalt suurem mõju boniteedi paranemisele peaks olema puistu kuivendatuse faktoril. Arvatavasti on olnud metsatakseerimise käigus antud tunnuse registreerimine ebatäpne. Mõneti üllatavaks võiks esmapilgul pidada metskonna faktori olulisust. Kuna metsakorralduse tehnoloogia kohaselt tavaliselt takseerib ühe metskonna puistusid üks taksaator, kinnitavad saadud tulemused seisukohta, et metsataksaatoritel on individuaalsed süstemaatilised vead, mis tehakse silmamõõdulisel puistu kõrguse takseerimisel.

Selleks, et saada ülevaade erinevate faktorite efektist, telliti protseduurilt GLM boniteedi H50 suurenemise oodatavad väärtused diskreetsete faktorite erinevatel tasemetel (Tabel 5).

Tabel 5

Puistu boniteedi H50 suurenemise oodatavad väärtused diskreetsete faktorite erinevatel tasemetel

Puuliik	Mänd	Kuusk	Kask	Teised
	2.12	1.46	1.80	1.60
	1950-ndail		1990-ndail	
Kuivendus	Ei	Ja	Ei	Ja
	1.49	2.00	1.43	2.06
Tekkeviis	Looduslik	Kultuur	Looduslik	Kultuur
	1.94	1.55	1.86	1.63
Põlvkond	Sama	Erinev		
	2.65	0.84		

6. Kokkuvõte

Käesoleva uurimuse tulemusena on saadud arvestatav kinnitus, et üleüldine metsa kasvu kiirenemise tendents toimub ka Eestis. Puistute boniteet H50 on Eesti metsades viimase nelja aastakümne jooksul paranenud keskmiselt ligikaudu 2 meetrit. Saadud hinnang on samas suurusjärgus sissejuhatuses nimetud naabermaades saadud tulemustega. Saadud boniteedi H50 paranemise hinnang 2 m on saadud konkreetset puistu kasvumudelit (Kiviste, 1997) rakendades. Mõnda teist puistu kasvumudelit kasutades võib boniteedi paranemise hinnang osutuda teistsuguseks.

Tuleb tõdeda, et puistute kasvu kiirenemist põhjustanud faktorite osas on veel palju selgusetut. Antud uurimuses ei olnud võimalik uuritavate faktorite hulka lülitada globaalseid faktoreid nagu atmosfääris süsihappegaasi kontsentratsiooni suurenemine, õhusaaste jms ega ka paljusid otseselt puistu kasvu mõjutavaid faktoreid (näiteks hooldusraieid). Samuti tuleb tunnistada mõnede faktorite (näiteks kuivendatus) registreerimise puudulikkust.

Puistu kasvukäigu modelleerimine on siiani isegi kasvutingimuste muutumatust eeldades olnud küllalt keeruliseks probleemiks. Teades nüüd, et meie metsade kasvutingimused on ajas muutuvad, lähevad metsa kasvu mudelid veelgi komplitseeritumaks isegi siis, kui oskaksime ennustada puistu kasvu mõjutavate faktorite muutumist tulevikus. Antud uurimuses saadud tulemuste valguses on selgunud, olemasolevad Eesti puistute kasvumudelid vajavad korrigeerimist.

Kirjandus

Elfving, B. and Tegnhammar, L. 1996. Trends of tree growth in Swedish forests 1953-1992: An analysis based on sample trees from the National Forest Inventory. Scandinavian Journal of Forest Research. Vol. 11: 26-37.

Eriksson, H. and Johansson, U. 1993. Yields of Norway spruce (Picea abies (L.) Karst.) in two consecutive rotations in southwestern Sweden. Plant and Soil. Vol. 154: 239-247.

Hari, P.; Arovaara, H; Raunemaa, T.; Hautojärvi, A. 1984. Forest growth of energy production: a method of detecting trends in growth potential of trees. Canadian Journal of Forest Research. Vol. 14: 437-440.

Kiviste, A. 1997. Eesti riigimetsa puistute kõrguse, diameetri ja tagavara vanuseridade diferentsmudel 1984.-1993.a. metsakorralduse takseerkirjelduste andmeil. Eesti Põllumajandusülikooli teadustööde kogumik. 189: 63-75.

Nilson, A., Kiviste A. 1984. Männikute kasvukäigu mudel tüpiseerimata kasvukohatingimuste järgi. Eesti Põllumajanduse Akadeemia teaduslike tööde kogumik 151: 50-59.

Spiecker, H., Mielikäinen K., Köhl M., Skovsgaard, J.P.(Eds.) 1996. Growth Trends in European Forests. Springer, 372 pp.

Нильсон А., Кивисте А. 1986. Отражение изменения окружающей среды в моделях хода роста леса, составленных разными методами. Мониторинг лесных экосистем. Каунас-Академия. С. 336-337.

Gini kordajast

Ene Käärik
Tartu Ülikool, Matemaatilise statistika instituut

Gini kordaja on kasutusel kui üldine ebavõrdsuse mõõt ja teatav tulu jaotuse hajuvuse näitaja. Ta väljendab ebavõrdsuse astet, näidates, millised on erinevused ühiskonnas rikkuse jaotumises ühiskonna erinevate sotsiaalsete kihtide vahel.
Gini defineeris 1912 nn keskmise erinevuse (*Gini mean difference*) $\Delta=E(|X-Y|)$ ja 1914 üldise jaotusvaba ebavõrdsuse mõõdu (*Gini ratio*), mis baseerus keskmise erinevusel (Kotz, 1983; Basmann, 1985). Gini kordaja on seotud nn Lorenzi kõveraga (L), mille kujutamisel graafikuna on x-teljel on kumulatiivne tulu saajate osakaal ja y-teljel vastav tulu osakaal. Lorenzi kõver on defineeritud võrdusega:

$$L(y)=\int_0^y \frac{xdF(x)}{E(Y)},$$

kus Y on tulu, mille jaoks eksisteerib keskväärtus $E(Y)$ ja $F(y)$ on tulu saajate jaotusfunktsioon. Kui kõik üldkogumi liikmed saavad võrdset tulu, siis Lorenzi kõver on nurgapoolitajaks. Lorenzi kõvera omadusi vt lähemalt Kotz (1983), lk 157.

1. Gini kordaja arvutamise algoritm*

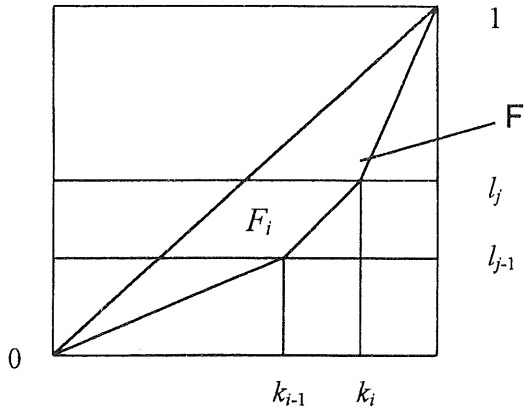
Oletame, et meil on n uuritava tunnuse väärtust esindatud q erineva vaatlusega $a_1 < a_2 < \ldots < a_q$, sagedustega $H_1, H_2, \ldots, H_q$ (suurused a_i võivad olla näiteks klassikeskmised), $\Sigma H_i = n$.
Eeldame, et kehtib seos

$$\sum_{j=1}^{n} x_j = \sum_{i=1}^{q} a_i H_i .$$

Gini kordaja (indeks) on defineeritud kui Lorenzi kontsentrat-

* Algoritmi esitus põhineb Hartung, J jt 1975.

sioonimõõt (*LKM*), mis avaldub kui kahekordne Lorenzi kõvera ja nurgapoolitaja vaheline pindala (*LKM*=2*F*) , ehk pindala *F* suhtena kogu nurgapoolitaja all olevasse pindalasse.



Joonis 1.Lorenzi kõver.

Lorenzi kõveral asuva punkti koordinaadid avalduvad järgmiselt:
a) abstsiss: (*kumuleeritud normeeritud kaal*)

$$k_i = \sum_{j=1}^{i} \frac{H_j}{n},\ i = 1,2,\ldots, q;$$

b) ordinaat: (*kumuleeritud normeeritud kaalutud tulu, a_jH_j –kaalutud tulu*)

$$l_i = \frac{\sum_{j=1}^{i} a_j H_j}{\sum_{j=1}^{q} a_j H_j},\ i = 1,2,\ldots, q.$$

Lorenzi kõvera ja nurgapoolitaja vaheline pindala arvutatakse kui üleval-pool Lorenzi kõverat oleva pinnaosa ja ülemise kolmnurga pindalade vahe st $F = \sum_{i=1}^{q} F_i - 0.5$. Iga üksiku trapetsi pindala F_i avaldub järgmiselt:

$$F_i = \frac{k_{i-1}+k_i}{2}(l_i - l_{i-1}) = \frac{k_{i-1}+k_i}{2}\frac{1}{\sum_{j=1}^{q} a_j H_j}\left(\sum_{j=1}^{i} a_j H_j - \sum_{j=1}^{i-1} a_j H_j\right)$$

$i=1, \ldots, q.$

Et definitsiooni järgi $LKM=2F$, siis saame Lorenzi kontsentratsiooni-mõõdu arvutamiseks järgmise seose:

$$LKM = 2\sum_{i=1}^{q} F_i - 1.$$

Asendame saadud seosesse suuruse F_i avaldise ja rühmitades liidetavad saame tulemuseks

$$LKM = \sum_{i=1}^{q} \frac{k_{i-1}\sum_{j=1}^{i} a_j H_j - k_i \sum_{j=1}^{i-1} a_j H_j}{\sum_{j=1}^{q} a_j H_j} + \sum_{i=1}^{q} \frac{k_i \sum_{j=1}^{i} a_j H_j - k_{i-1}\sum_{j=1}^{i-1} a_j H_j}{\sum_{j=1}^{q} a_j H_j} - 1$$

Siin teise summa leidmisel arvestame seost $\sum_{i=1}^{q}(k_i l_i - k_{i-1} l_{i-1}) = 1.$

Olemegi leidnud Gini kordaja arvutamiseks valemi.

Gini kordaja arvutamise valem:

$$LKM = \sum_{i=1}^{q} \frac{k_{i-1}\sum_{j=1}^{i} a_j H_j - k_i \sum_{j=1}^{i-1} a_j H_j}{\sum_{j=1}^{q} a_j H_j}$$

Esitatud algoritm on kasutusel ka Eesti Statistikaametis (vt metoodika Braber, R., Gijsberts, I.,1995)

Selliselt arvutatud Gini kordaja rahuldab võrratust $0 \le LKM \le \frac{n-1}{n}$.

Et kordaja maksimaalne väärtus oleks 1, siis tihti ta normeeritakse, st

$$LKM_{nor} = \frac{n}{n-1} LKM .$$

Gini kordaja esitatakse tihiti ka protsentides.

Mida suurem on Gini kordaja, seda suurem on ebavõrdsus. Täieliku võrdsuse korral (kõik ühiskonna liikmed saavad ühesugust tulu) langeb Lorenzi kõver kokku nurgapoolitajaga ja Gini kordaja on null. Teine äärmuslik olukord on siis, kui üks isik saab kogu riigi tulu ja teised ei saa midagi. Lorenzi kõver on siis kõdunud ruudu alumiseks ja parempoolseks küljeks. Sel juhul saame Lorenzi kõvera ja nurgapoolitaja vaheliseks pindalaks pool ruudu pindala ja Gini kordaja (kui selle kahekordne) on võrdne ühega.

2. Gini kordaja arvutamisest

Toome näite Gini kordaja algoritmi rakendamise kohta. Selleks, et arvutustehniliselt oleks lahenduskäik ülevaatlik, võtame aluseks mitte kogu valimi, vaid ainult 5 punkti, st kvintiilid*

- **Arvutamine kvintiilide järgi**

Võtame aluseks Eesti leibkondade sissetulekute ja kulutuste uuringu (*LSKU*) 1997 a andmed, leiame netotulu pereliikme kohta .

1. VALIMI KVINTIILID

Arvutame kvintiiliklasside mediaanid (10-, 30-, 50-, 70-, 90-protsentiilpunktid) pereliikme netotulule. Kasutame eeltoodud algoritmi ja saame Gini kordaja 27.24%. Tulemused on alljärgnevas tabelis:

Tabel 1.

Gini kordaja arvutamine valimi kvintiilide järgi*

NETOTULU	KAAL	KTULU	KTNORM	KKAAL	CUMKT	CUMK	ZZ
741	20	14820	0.0893	0.2	0.0893	0.2	
1105	20	22100	0.1332	0.2	0.2246	0.4	0.0088
1392	20	27840	0.1678	0.2	0.3902	0.6	0.0226
1883	20	37660	0.2269	0.2	0.6171	0.8	0.0581
3177	20	63540	0.3829	0.2	1	1	0.1829

* Kui me jagame sissetulekute järgi järjestatud indiviidid viieks võrdseks osaks, siis iga viiendiku sissetulek on sissetuleku kvintiil. Jagades kümneks osaks saame detsiilid ja sajaks – protsentiilid.

2. ÜLDKOGUMI KVINTIILID

Laiendame valimit kaaludega, saame pereliikme netotulu väärtused üldkogumi jaoks. Arvutame kvintiiliklasside mediaanid (10-, 30-, 50-, 70-, 90-protsentiilpunktid). Kasutame eeltoodud algoritmi ja saame Gini kordaja 28.52%. Tulemused on alljärgnevas tabelis:

Tabel 2.

Gini kordaja arvutamine üldkogumi kvintiilide järgi*

NETOTULU	KAAL	KTULU	KTNORM	KKAAL	CUMKT	CUMK	ZZ
808	20	16160	0.0897	0.2	0.0897	0.2	
1147	20	22940	0.1274	0.2	0.2171	0.4	0.0075
1149	20	28980	0.1609	0.2	0.3780	0.6	0.0209
2018	20	40360	0.2241	0.2	0.6021	0.8	0.0589
3583	20	7166	0.3979	0.2	1	1	0.1979

Analüüsides eelnevaid tulemusi, näeme, et valimi kvantiilid annavad madalama Gini kordaja, mis viitab väiksemale ebavõrdsusele. See on kooskõlas ka andmetega. Tõepoolest

a) viimase ja esimese kvantiili suhe valimis on 3177/741=4.29 ja üldkogumis 3583/808=4.43, mis viitab suuremale ebavõrdsusele üldkogumis;

b) valimis saab 40% elanikest 22.46% sissetulekutest, 60% elanikest 39.02% sissetulekutest ja 80% elanikest 61.71% sissetulekutest. Üldkogumis on vastavad sissetulekute osakaalud pisut madalamad, vastavalt 21.71%, 37.8%, 60.2%, mis viitab suuremale ebavõrdsusele (tabelites vastavalt veerud CUMKT ja CUMK).

---

* Tabelis kasutatavad tähistused:
NETOKULU - netokulu pereliikme kohta
KTULU - kaalutud tulu (kaal*netokulu)
KTNORM - kaalutud normeeritud tulu (kaal*netotulu/kaalutud tulude summa)
KKAAL - normeeritud kaal (kaal/kaalude summa)
CUMKT - kumuleeritud kaalutud normeeritud tulu (y-koordinaat Lorenzi kõveral)
CUMK - kumuleeritud normeeritud kaal (x-koordinaat Lorenzi kõveral)
ZZ - ristkorrutiste vahed valemis

Üldiselt Gini kordaja arvutamisel kvintiilide või ka detsiilide järgi saame mõnevõrra alahinnatud tulemuse (punktide arv on liiga väike pindala täpseks arvutamiseks).

3. Märkused

1. Mida peame järjestama?

   Lähtudes algoritmist tuleb *järjestada valimi sissetulekud.* Kuigi meie andmete korral on suurtel peredel valimisse sattumise tõenäosus suurem ning seega on valimis suured pered üleesindatud ja valimis on rohkem vaeseid, ei põhjusta lähtumine valimi järjestamisest nihet Gini kordaja suuruses. Kaalude arvestamisega laiendatakse valimi tulu väärtused selliselt, et arvestatakse perede tegelikku struktuuri üldkogumis, samas aga nende järjekord ei muutu. *Valimi tulude järjestus vastab üldkogumi tulude järjestusele.*

2. Rääkides kaalutud tulust tuleb eristada kaaluga laiendatud tulu ja kaaluga korrutatud tulu.

   - Järjestades kaaluga korrutatud tulud (kaal*tulu), me tegelikult järjestame pered nende sissetulekute järgi ja hindame Gini kordajat arvestades peret kui ühikut, arvestamata pereliikmete arvu. Sellise järjestuse põhjal arvutatud Gini kordaja ei vasta objektiivsele reaalsusele. Kui kaks peret saavad näiteks 1000 krooni tulu, siis pere sissetuleku mõttes on nad võrdsed, aga kui üks peredest on üksiku pere ja teine 5-liikmeline, siis tegelikkuses on teine pere vaesem.
   - Järjestades eelnevalt laiendatud tulud st laiendades tulu kaalu kui sagedusega, seejärel võttes arvesse, et nüüd on iga indiviid valimis kaaluga 1 st tegelikult, et pered on nüüd võrdselt esindatud (meil on üldkogum), saame samad tulemused kui järjestades valimi tulud ja arvestades Gini kordaja algoritmis pere kaalu.

3. Gini kordaja arvutamisel on kasutatud ka teatavat andmete silumist alt ja ülalt (vt *Luxemburgi sissetulekute uuring*). Alt silumise korral need sissetuleku väärtused, mis on väiksemad kui 1% mediaanist, asendatakse 1% mediaaniga ja ülalt silumise korral need väärtused, mis on suuremad kui 10-kordne mediaan, asendatakse 10-kordse mediaaniga. Selliselt silutud 1997 a leibkonnauuringute andmete järgi on Gini kordaja 37.29% (ainult alt silutud) ja 36.33% (nii alt kui ülalt silutud). Silumata andmete järgi on Gini kordaja 37.34%.

4. Luksemburgi sissetulekute uuringus pakutakse ekvivalentsus-skaalaks ruutjuurt pereliikmete arvust st sissetulek pereliikme kohta arvutatakse kui kogu leibkonna sissetuleku ja pereliikmete arvu ruutjuure suhe. Selliselt arvutatud keskmise pereliikme sissetulek 1997 a leibkonnauuringutes annab Gini kordajaks silumata andmete korral 38.48%, alt silumise korral 38.4% ja mõlemalt poolt silumise korral 37.37%.

Kasutatud kirjandus

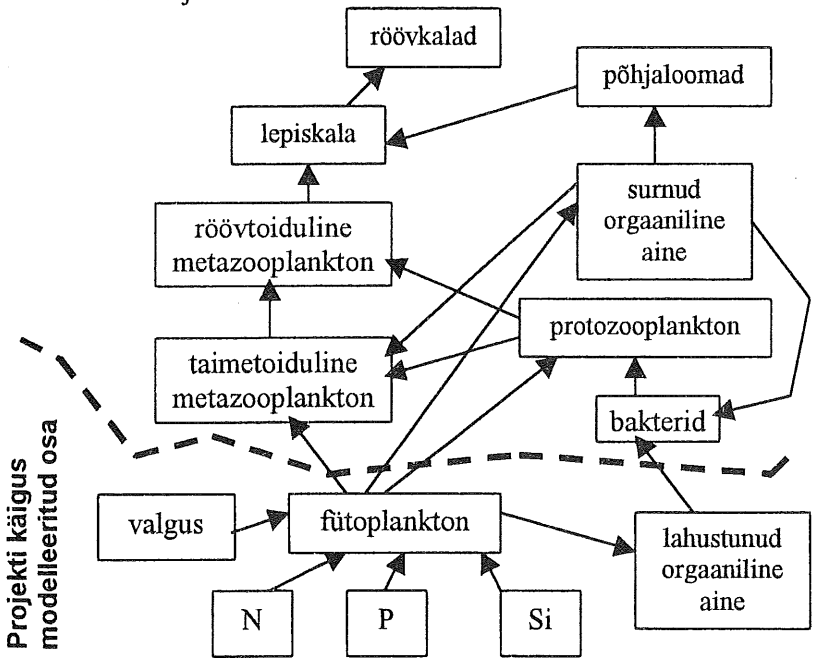
1. Kotz, S., N.L. Johnson (1983) Encyclopedia of Statistics Sciences. Vol 4, 5, John Wiley.
2. Kendall, M., Stuart, A.(1958) The advanced theory of statistics . Vol 1.Distribution theory. pp 46-51, 241, Criffin&Co Ltd, London.
3. Dunn, W.N. (1994) Public policy analysis: an introduction, pp 364-368, Prentice Hall.
4. Hartung, J., Elpelt, B., Klösener, K.-H. (1975?) Statistik. Lehr- und Handbuch der angewandten Statistik, S. 50-53, Oldenbourg Verlag, München,Wien.
5. Braber, R., Gijsberts, I. (1995) Practical Approaches to Poverty. Monitoring and Analysis. REsearch Report No 95/4
6. Advances in Econometrics, Vol 4 (1985). Econimic Inequality: survey, methods and measurements. Ed Basmann, R.L., Rhodes, G.F., JAI press Inc, London.
7. LIS Inequality Indices (1998). Luxembourg Income Study Database.

Võrtsjärve ökoloogia modelleerimisest

Märt Möls, Tiina Nõges, Peeter Nõges
Tartu Ülikool, Võrtsjärve Limnoloogiajaam

Baltimaade suuruselt teine järv Võrtsjärv on madal, keskmise sügavusega 2,8m ja alates 1960. aastatest kiirelt eutrofeeruv järv. Seoses järve seisukorra halvenemisega uuriti aastail 1993-1997 Soome-Eesti ühisprojekti raames intensiivselt Võrtsjärve olukorda ja inimkäitumise mõju sellele. Projekti raames koguti andmeid nii järve kui ka tema vesila kohta ja alustati neid kasutades järve mudeli koostamist. Võrtsjärve mudeli eesmärgiks seati võimalus mängida läbi erinevaid klimaatilisi ja inimkäitumisest tingitud situatsioone ja jälgida erinevate stsenaariumite mõju järvele. Esimese etapina on valminud fütoplanktoni, lahustunud orgaanilise aine ja hapniku kogust kirjeldav mudel (vaata joonis 1).

Joonis 1. Võrtsjärve ökosüsteemi lihtsustatud skeem



Siinkohal on ära toodud näitena leitud mudelist ränivetikate kontsentratsiooni arvutamist käsitlev osa (täielikku kirjeldust vaata Nõges *et al*, 1998):

$$\frac{dC_{ränivetikad}}{dt} = \left(\mu - \rho - \frac{Q_{sissevool}}{V}\right) \cdot C_{ränivetikad},$$

kus

ρ - vetikate kadu suremise, ärasöömise või settimise kaudu. Funktsioon, sisaldab mitmeid parameetreid ja mõjutegureid;

$\frac{Q_{sissevool}}{V}$ - järve sissevoolava vee mõju vetikate kontsentratsioonile;

$\mu = \mu_{max\ kiirus} \cdot f_{temp} \cdot f_{valgus} \cdot f_P \cdot f_N \cdot f_{Si}$ - vetikate kasvukiirus, milles

$\mu_{max\ kiirus}$ - on maksimaalset võimalikku kasvukiirust iseloomustav konstant;

f_{temp} – ebasobivast temperatuurist tingitud kasvu kidumine;
f_{valgus} – valguse vähesusest tingitud kasvu pidurdumine;
f_P – fosforinappuse mõju;
f_N – lämmastikunappuse mõju ja
f_{Si} – räninappuse mõju.

Enamiku kasvukiirust pärssivate tegurite L mõju on piisavalt hästi esitatav kujul $f_L = \frac{L}{L+C}$, milles C on järvespetsiifiline konstant.

Ehkki mõned konstandid on hinnatud laboratoorselt, sõltub enamik neist siiski uuritavast järvest ja nad muutuvad tingitult erinevatest teguritest, näiteks bioloogilisest kooslusest.

Tundmatute konstantide hinnanguks valitakse harilikult väärtused, mille puhul erinevus mõõtmistulemuse ja mudeli põhjal arvutatud väärtuste vahel on väikseim (tavaliselt minimiseeritakse vigade ruutude summa). Kuna hüdroloogias kutsutakse taolist parameetri hinnangu valikuprotseduuri harilikult parameetri kalibreerimiseks, siis viitame edaspidi sellele lähenemisviisile kui klassikalisele kalibreerimisele.

Võrtsjärve ökoloogia modelleerimise käigus kerkisid esile mõned probleemid, mis muutsid küsitavaks traditsioonilise hüdroloogias kasutatava metoodika efektiivsuse juhul, kui mudelis ei arvestata

kõiki, ka vähetähtsaid, mõjufaktoreid või kui kasutatakse mõõtmis-aparatuuri piiratud täpsusest tingitult ebatäpseid algandmeid, see on juhul, kui ei saavutata absoluutset kooskõla algandmete ja mudeli prognooside vahel.

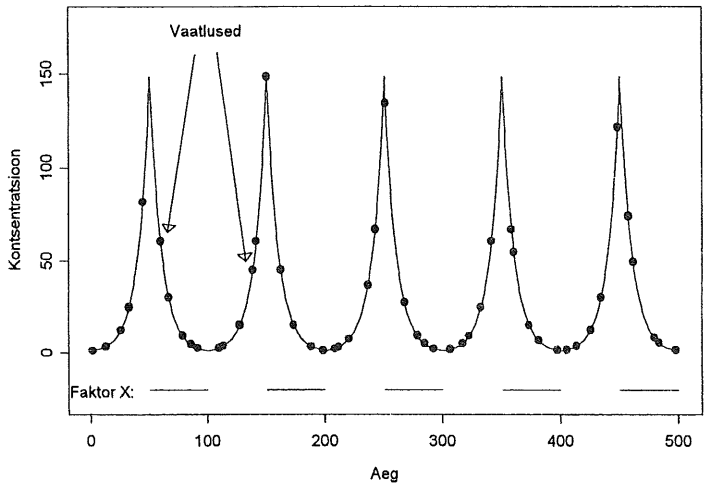
Üks võimalus kasutatava metoodika kontrollimiseks on jälgida tema tulemuslikkust teadaolevate süsteemiparameetrite hindamisel. Kuna praktikas ettetulevate andmete puhul on harva täpselt teada neid andmeid tekitanud mehhanism, võib reaalsete andmete asemel kasutada kunstlikku süsteemi, milles andmed genereeritakse vastavalt etteantud mudelile.

Antud juhul võeti katsemudeliks lihtne kasvu-kadumise mudel kontsentratsiooni C kirjeldamiseks:

$$\frac{dC}{dt} = (c_1 + c_2 X) \cdot C .$$

Arvutuste hõlbustamiseks eeldati, et mõõdetud mõjufaktor X on binaarne, võimalike väärtustega 0 ja 1. Süsteemi illustreerib joonis 2.

Joonis 2. Mudelit $dC/dt = (0{,}1 - 0{,}2X)\ C$ kasutades modelleeritud andmed

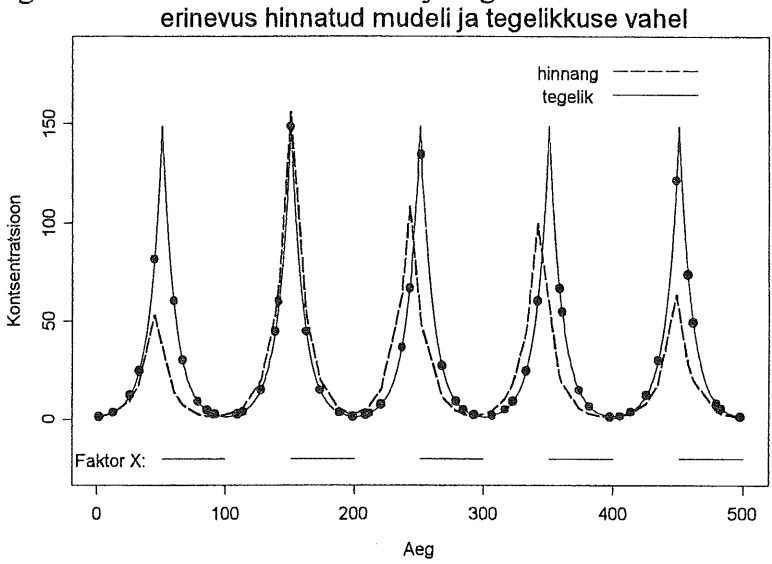


Juhul, kui on täpselt teada faktori X muutumine ajas, saab parameetrite väärtused ülaltoodud kontsentratsiooniväärtusi kasutades hinnata

absoluutse täpsusega ja hinnatud (kalibreeritud) mudelist saadud tulemused ühtivad täielikult protsessi tegeliku käitumisega.

Kui faktori X väärtuste muutumine ajas on teada vaid vaatlusandmete põhjal (nagu see enamasti on), tekivad hindamisel ebatäpsused, mis on tingitud teadmatusest, kuidas muutuvad kirjeldava faktori väärtused vaatlustevahelisel perioodil. Paljud veekogude modelleerimiseks kirjutatud spetsiaalprogrammid lubavad lähendada sõltumatute faktorite X muutumist vaatlustevahelisel perioodil, kasutades splaine või muid silujaid. Paraku ei õnnestu nende käitumist siiski täiesti täpselt kirjeldada. Taolise infokao tõttu tekkivat ebakõla vaatluste ja mudeli prognooside vahel iseloomustab joonis 3.

Joonis 3. Faktori X muutusest vaatlustevahelisel perioodil tingitud erinevus hinnatud mudeli ja tegelikkuse vahel



Ülaltoodud joonisel esitatud mudeli parameetreid hinnates siluti faktori X väärtusi lineaarselt, st. mõõtmishetkede t_1 ja t_2 vahelisel ajal loeti X väärtuseks suurus

$$X(t, t_1 < t < t_2) = \frac{(t - t_1)X_{t_2} + (t_2 - t)X_{t_1}}{t_2 - t_1} .$$

Hindamisprotseduur töötab ja parameetrite c_1 ja c_2 hinnangud satuvad tegelike väärtuste lähedale.

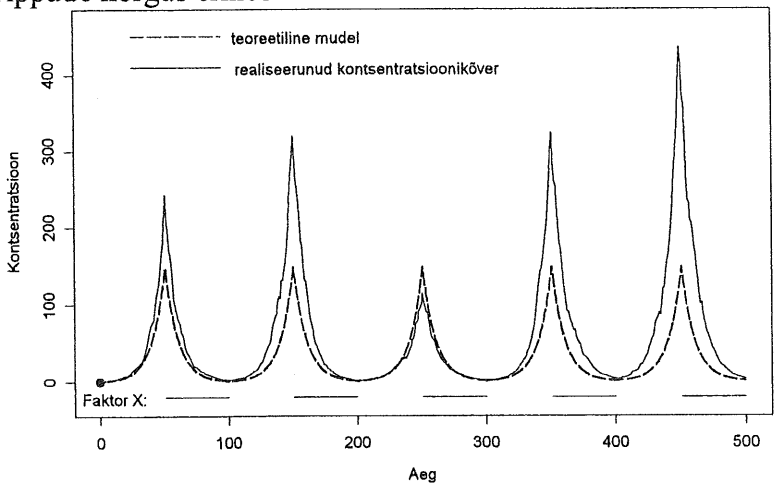
Paraku mõjutavad kontsentratsiooni paljud faktorid, mida on raske arvesse võtta. Tühine lokaalne reostus või pilvevari järve ühe lahesopi kohal võivad kontsentratsiooni muutumise kiirust muuta. Kontsentratsiooni muutumise kiiruse juhuslikud fluktuatsioonid (vead) põhjustavad suuremaid või väiksemaid erinevusi teoreetiliselt õige mudeli ja tegelikult realiseerunud kontsentratsioonikõvera vahel.

Hindamismeetodi töökvaliteet sõltub vigade tekkimise viisist. Järgnevates näidetes on eeldatud, et kasvukiiruse viga on konstantne lühikese ajaperioodi Δt jooksul (Δt on tunduvalt väiksem kui vaatlustevaheline intervall), järjestikuste Δt pikkuste ajalõikude vältel esinevad vead on aga omavahel sõltumatud:

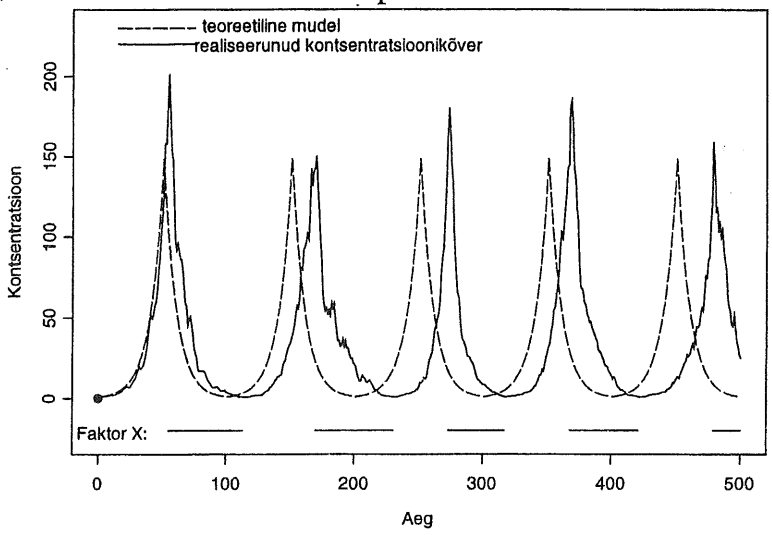
$$\frac{dC}{dt} = (c_1 + c_2 X + \varepsilon_t) \cdot C \,, \quad \varepsilon_t \perp \varepsilon_{t+dt} \,.$$

Joonistel 4 ja 5 on toodud kaks näidet võimalikest kontsentratsioonikõveratest (kõrvuti kontsentratsiooni muutustega häireteta keskkonnas).

Joonis 4. Faktor X ei sõltu kontsentratsioonist C. Tippude kõrgus erineb



Joonis 5. Faktor *X* sõltub kontsentratsioonist *C*. Võimalikud on kõrvalekalded perioodilisusest



Kalibreerimisprotseduuri hindamiseks tekitati tuhat kontsentratsioonikõverat ja nende pealt võetud mõõtmistulemusi kasutades hinnati tuhat korda tundmatuid parameetreid. Võrdluseks teostati lineaarne regressioon kasutades teisendatud suuruseid y,

$$y = \ln\left(\frac{C_{t_i}}{C_{t_{i-1}}}\right)/(t_i - t_{i-1}) .$$

Teisendus on valitud selliselt, et vigade omavaheline sõltuvus oleks väike ja tundmatud parameetrid avalduksid otse regressioonikordajatena. Saadud tulemused kajastuvad tabelis 1 ja joonisel 6.

Tabelis 1 on iga lahtri arvutamiseks genereeritud 1000 korda "realiseerunud" kontsentratsioonikõver, mida kasutades on leitud valim.

Tabel 1.

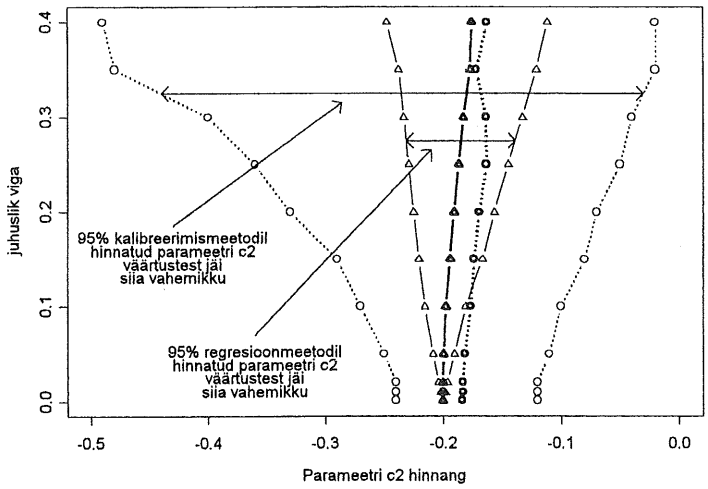
Hindamismeetodite käitumine erinevates situatsioonides.

Meetodi nimetus (sulgudes andmete genereerimisel lisatud häirete standardhälve)			
c_1 hinnangute keskmine ± hinnangute standardhälve	c_2 hinnangute keskmine ± hinnangute standardhälve	uute andmete prognoosimisel tehtavate vigade ruutude summa (mediaan)	
		lühike prognoos (10 vaatlust)	pikk prognoos (70 vaatlust)
õige väärtus			
0,1	-0,2		
klassikaline kalibreerimine (0,01; 60 mõõtmist)			
0,0911±0,0169	-0,1830±0,0370	14816	157135
klassikaline kalibreerimine (0,05; 60 mõõtmist)			
0,0904±0,0194	-0,1814±0,0420	13455	167413
klassikaline kalibreerimine (0,5; 60 mõõtmist)			
0,0710±0,0676	-0,1412±0,1483	218690	$2,78*10^{8}$
klassikaline kalibreerimine (0,5; 20 mõõtmist)			
0,1191±0,0892	-0,2407±0,2034	671936	$7,16*10^{11}$
klassikaline kalibreerimine (0,5; 10 mõõtmist)			
0,1407±0,0988	-0,3119±0,2198	2535428	$3,38*10^{15}$
regressioon teisendatud tunnusega (0,01; 60 mõõtmist)			
0,0998±0,0032	-0,2002±0,0012	14085	153972
regressioon teisendatud tunnusega (0,05; 60 mõõtmist)			
0,0995±0,0046	-0,1994±0,0054	12322	165836
regressioon teisendatud tunnusega (0,5; 60 mõõtmist)			
0,0845±0,0326	-0,1682±0,0476	304071	$1,26*10^{9}$
regressioon teisendatud tunnusega (0,5; 20 mõõtmist)			
0,0946±0,0569	-0,1887±0,0926	398057	$1,24*10^{9}$
regressioon teisendatud tunnusega (0,5; 10 mõõtmist)			
0,1086±0,0891	-0,2146±0,1818	1733486	$6,79*10^{12}$
C vaatluste keskmine (0,05; 60 mõõtmist)			
		16189	123971

Juhuslike vigade suurenedes on loomulik, et ka parameetrite hinnang muutub ebatäpsemaks. Kui aga hindamisel esineb süstemaatiline viga, siis see vihjab hindamismetoodika ebatäiuslikkusele. Tabelit 1 vaadates võib oletada, et juhusliku vea osakaalu suurenedes kiputakse parameetreid alahindama, vaatluste arvu vähenedes (kui vaadeldud kontsentratsioonilainetusi jääb väheks) aga ülehindama. Joonis 6 illustreerib seda tendentsi veelgi selgemalt. Ühtlasi võib märgata, et kalibreerimismeetodil saadud hinnangud kõiguvad tunduvalt, meetod ise näib muutuvat üsna ebatäpseks kui vaatlusandmetele lisandub vähegi arvestatav juhuslik viga. Lisaks osutub mainitud lähenemisviis olevat väga tundlik kirjeldava faktori käitumise suhtes vaatlustevahelisel perioodil (suur hinnangute hajuvus ka praktiliselt ilma juhusliku veata vaatlusandmete korral). Ka vaadeldud regressioonimudel annab nihkega hinnangu, ehkki märksa väiksema, sest antud juhul ei võta regressioonmudel arvesse vea konstantsust ajavahemiku Δt jooksul.

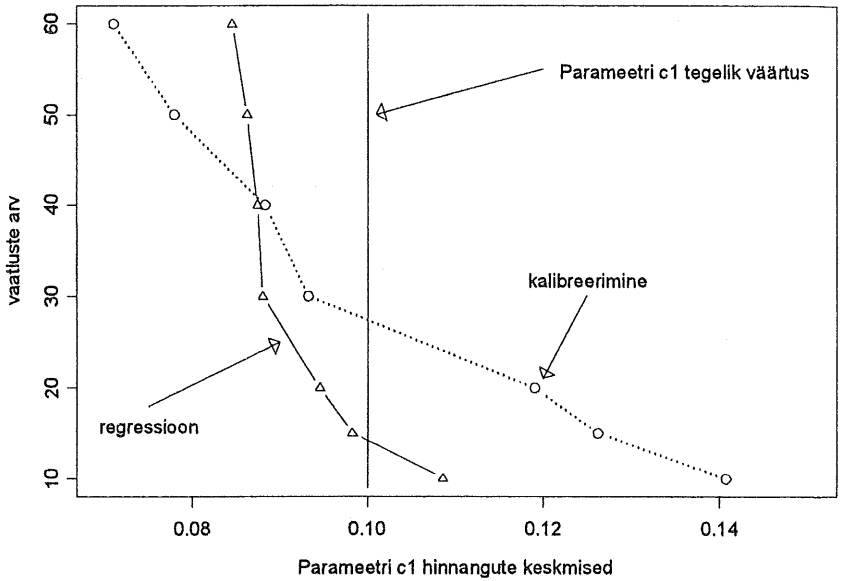
Kontsentratsiooni käitumist prognoosides pikemaks perioodiks osutub vaatlustulemuste keskmine paremakski näitajaks kui teised väljatoodud meetodid (vt. tabel 1). Tõsi küll, antud näite puhul on kasutatud stabiilset kontsentratsioonikõverat.

Joonis 6. Parameetri hinnangu täpsuse sõltuvus juhuslikust veast



Kui kasutatava informatsiooni hulk (valimi maht) suureneb, siis peaks hea hindamismetoodika andma hinnanguteks väärtuseid üha lähemal parameetri tegelikule väärtusele. Ehkki hinnangute hajuvus tõepoolest väheneb, osutub, et hinnangute keskmine võib vaatluste lisandumisel hoopis eemalduda tegelikust parameetri väärtusest (hindamismetoodika annab nihkega hinnangu).

Joonis 7. Parameetri hinnangute keskmine sõltuvalt vaatluste arvust



Vaatamata nõrgale tulemusele parameetrite väärtuste hindamisel ja tuleviku prognoosimisel võib klassikalist kalibreerimisprotseduuri kasutav mudel siiski saavutada väga hea kooskõla kalibreerimiseks kasutatud andmete ja mudeli järgi arvutatud väärtuste vahel, näides tihti välja märgatavalt parem kui täpsemate parameetri väärtustega ja parema prognoosivõimega mudelid.

Kokkuvõte

Hüdroloogias kasutatav veekogude modelleerimise metoodika viib teatavatel juhtudel mudeliteni, mille prognoosiv väärtus on madal. Kasutatav tundmatute parameetrite kalibreerimistehnika võib aga ka

üsna heades tingimustes anda kehvi hinnanguid mudeli parameetrite väärtustele. Vigade olemasolu arvestav statistiline lähenemine võib anda paremaid tulemusi nii prognoosimisel kui ka veekogus toimuvate protsesside kirjeldamisel.

Kirjandus

1. Mathematical Modeling of Water Quality: Streams, Lakes and Reservoirs. Internatonal Series on Applied Systems Analysis; 12. Edited by G. T. Orlob. 1982.
2. Present State and future fate of Lake Võrtsjärv. The Finnish Environment, 209. Edited by Timo Huttula and Tiina Nõges. Tampere, 1998.

Statistilisest sõltuvusest

Ene-Margit Tiit
Tartu Ülikool, Matemaatilise statistika instituut

1. Sõltuvus ja statistiline sõltuvus

Sõltuvus on tavaelus tuntud mõiste, ja tihti negatiivse kõlavarjundiga (poliitikas, majandusteaduses, arstiteaduses). *Funktsionaalne sõltuvus* on üks matemaatika põhimõisteid. *Statistiline sõltuvus* on seotud mõlema eelpool nimetatud sõltuvuse mõistega:

- statistilise sõltuvuse kaudu on tihti võimalik teistes eluvaldkondades ilmnevaid sõltuvusi modelleerida;
- statistiline sõltuvus on teatavas mõttes üldistuseks funktsionaalse sõltuvuse mõistele.

Mis tähtsus on statistilisel sõltuvusel? – Statistiline sõltuvus on nii matemaatilise kui ka rakendusstatistika põhimõiste.

1. Kui me vaatleme statistilisi andmeid, siis me küsime, kas vaatlustulemused on *statistiliselt sõltumatud*.
2. Kui meie eesmärgiks on *konstrueerida mudel*, siis me küsime, kas uuritav juhuslik suurus *sõltub statistiliselt* mudeli argumentidest – vastasel korral on mudeli konstrueerimine mõttetu.
3. Mida *tugevam* on kahe juhusliku suuruse vaheline statistiline sõltuvus, seda rohkem teavet sisaldab üks juhuslik suurus teise kohta ja seda paremini (täpsemalt) on üht juhuslikku suurust kasutades *võimalik teise juhusliku suuruse väärtusi prognoosida.*

2. Statistiline sõltuvus kahe juhusliku suuruse vahel

Olgu X ja Y juhuslikud suurused (statistikas öeldakse ka: *tunnused*) ja (X, Y) juhuslik vektor, mille komponentideks on need juhuslikud suurused. Olgu juhuslike suuruste X ja Y jaotusfunktsioonid vastavalt

$$F_X(x),\ F_Y(y)$$

ning olgu juhusliku vektori (X, Y) jaotusfunktsioon (kahe juhusliku suuruse *ühine jaotusfunktsioon*)

$$F_{XY}(x, y).$$

Definitsioon 1. Öeldakes, et juhuslikud suurused X ja Y on *statistiliselt sõltumatud*, kui iga x ja y puhul kehtib võrdus

$$F_{XY}(x,y) = F_X(x)F_Y(y) \tag{1}$$

ja juhuslikud suurused X ja Y on *statistiliselt sõltuvad*, kui võrdus (1) ei kehti kõigi argumentide korral, st kui leidub mingi selliste väärtuste paar (x, y), et võrdus (1) ei kehti. ●

Edaspidi me ütleme lihtsuse mõttes statistilise sõltuvuse asemel *sõltuvus*, sest kõneleme ainult statistilisest sõltuvusest ja selle erijuhtudest. Statistilise sõltuvuse definitsioonist järeldub, et kui kaks juhuslikku suurust on sõltuvad, siis ühe juhusliku suuruse *tinglik jaotus,* mis vastab teise juhusliku suuruse erinevatele väärtustele, üldiselt muutub. Sõltumatute juhuslike suuruste korral on aga kõik tinglikud jaotused ühesugused.

3. Regressioonsõltuvus ja selle tugevus

Definitsioon 2. Kui ühe juhusliku suuruse Y *tinglik keskväärtus* tingimusel $X = x$ muutub sõltuvalt x väärtusest , siis öeldakse, et juhuslike suuruste X ja Y vahel on *regressioonsõltuvus*. Juhusliku suuruse Y tinglik keskväärtus

$$E(Y / X = x) = Y_x \tag{2}$$

ehk Y *regressioon* X *järgi* on juhusliku suuruse Y *parim võimalik prognoos* X järgi. ●

Regressioonsõltuvus on üks statistilise sõltuvuse olulisi erijuhte. Erinevalt statistilisest sõltuvusest ei ole regressioonsõltuvus vastastikune: võib juhtuda, et Y sõltub X-st regressioonsõltuvuse mõttes (kuitahes tugevasti), kuid vastupidine väide pole õige. Parima prognoosi *headust* mõõdetakse prognoosivea

$$Q = E(Y - Y_X)^2 \tag{3}$$

ja juhusliku suuruse dispersiooni DY suhte abil, kasutades seosekordajat η, mida nimetatakse *regressioonisuhteks*

$$\eta = \sqrt{1 - \frac{Q}{DY}}. \tag{4}$$

Prognoos on seda parem, mida suurem (lähedasem ühele) on seosekordaja η. Vastavalt sellele, kui hea on prognoos, räägitakse ka statistilise sõltuvuse *tugevusest*.

- Kui seosekordaja η omandab maksimaalse väärtuse 1, siis on *regressioonseos täielik* (maksimaalse tugevusega).
- Kui seosekordaja η omandab minimaalse võimaliku väärtuse 0, siis *regressioonseos puudub*. Nii on alati, kui juhuslikud suurused X ja Y on sõltumatud, kuid vastupidine väide ei kehti üldiselt.
- Seega regressioonseose tugevust mõõdab regressioonisuhe η skaalal [0, 1].

Paneme tähele, et juhul kui $\eta = 1$, st regressioonsõltuvus on täielik, on Y tinglik keskväärtus (peaaegu kindlasti) *mittejuhuslik*, st *et* peaaegu kindlasti *Y sõltub X-st funktsionaalselt*. Enamasti aga ei ole see funktsionaalne sõltuvus lihtsalt esitatav meile tuntud elementaarfunktsioonide kaudu.

4. Lineaarne mudel ja lineaarne regressioon

Regressioonsõltuvuse puhul on juhusliku suuruse Y tinglik keskväärtus (2) juhusliku suuruse X mingi konstandist erinev mõõtuv funktsioon. Selleks, et konstrueerida *mudelit*, mis võimaldaks prognoosida Y väärtusi X järgi, oleks tarvis aga tinglikku keskväärtust lähendada mingi sobiva funktsiooniga, kusjuures saadud avaldis kindlasti sisaldab ka mingeid konstante, nn *mudeli parameetreid*. Kõige lihtsam funktsioon on teatavasti lineaarne. Sellest lähtudes lähendame Y tinglikku keskväärtust *lineaarfunktsiooniga*. Lähendamiseks võib kasutada mitmesuguseid lähendamise eeskirju, kõige levinum on aga lähendamine *vähimruutude meetodil*. See tähendab, et mudeli parameetrid määratakse nii, et prognoositava muutuja ja prognoosi erinevuse ruut oleks keskmiselt võimalikult väike,

$$Y_x \approx a + bx . \qquad (5)$$

Saime lineaarse mudeli juhusliku suuruse Y prognoosimiseks X järgi, kusjuures konstandid a ja b on mudeli parameetrid.

Definitsioon 3. Funktsioon $a + bx$ on juhusliku suuruse Y *parim lineaane prognoos* ehk *lineaarne regressioon* X järgi. •

5. Korrelatiivne sõltuvus

Arvutame prognoosivea Q taas valemi (3) abil, võttes selles parima prognoosi (2) asemele on parima lineaarse prognoosi (5), saame seosekordaja avaldisest (4) kordaja, mida nimetatakse *korrelatsioonikordajaks* ja mida tähistatakse tähega r. Lineaarset regressioonsõltuvust nimetatakse *korrelatiivseks sõltuvuseks*.

Definitsioon 4. Korrelatsioonikordaja on kordaja, mis mõõdab juhuslike suuruste vahelise *korrelatiivse seose tugevust*. •

- Korrelatsioonikordajale omistatakse kordaja b märk. Vastavalt sellele kõneldakse *positiivsest* ja *negatiivsest korrelatsioonist* kahe juhusliku suuruse vahel.
- Korrelatiivne sõltuvus on vastastikune ja sümmeetriline.
- Absoluutväärtuselt maksimaalse korrelatiivse sõltuvuse korral on juhuslike suuruste X ja Y vahel lineaarne funktsionaalne sõltuvus, st et kui korrelatsioonikordaja on absoluutväärtuselt võrdne ühega, siis leiduvad kordajad a ja b nii, et kehtib üks alljärgnevatest seostest.

$$R(X,Y) = 1 \leftrightarrow Y = a+bX; \quad R(X,Y) = -1 \leftrightarrow Y = -a+bX,\ a>0.$$

- Kui korrelatsioonikordaja väärtus on 0, siis öeldakse, et juhuslikud suurused on *mittekorreleeritud*. Sellest ei järeldu regressioonsõltuvuse puudumine ja ka mitte statistiline sõltumatus.
- Statistiliselt sõltumatute juhuslike suuruste korrelatsioonikordaja on null, st et statistilise sõltuvuse puudumisest järeldub regressioonsõltuvuse puudumine ja samuti ka korrelatiivse sõltuvuse puudumine.

6. Teised statistilise sõltuvuse liigid ja seosekordajad

Statistilise sõltuvuse erijuhte on kirjeldatud palju. Nimetame neist kaht olulisemat.

1. *Monotoonne sõltuvus* kajastab marginaaljaotuste *järjestuse* kooskõla. Monotoonse sõltuvuse puhul märgitakse (samuti kui korrelatiivse sõltuvuse korralgi) ära ka sõltuvuse suund, mis võib olla kas positiivne või negatiivne.
2. Diskreetsete marginaaljaotuste korral leiab maksimaalse tugevusega statistiline sõltuvus aset siis, kui kummagi tunnuse *väärtuste vahel on üks-ühene vastavus*, mille puhul üldiselt väärtuste järjestus ei ole oluline. Niisugust sõltuvust nimetatakse *assotsiatiivsuseks*, ning seda kasutatakse enamasti mittearvuliste tunnuste vahelise seose mõõtmisel.

Erinevat liiki statistilise sõltuvuse mõõtmiseks on defineeritud mitukümmend erinevat sõltuvuskordajat, kusjuures ka sama liiki sõltuvust, näiteks monotoonset sõltuvust ja assotsiatiivsust saab mõõta mitme erineva kordaja abil. Enamasti kasutatakse aga seosekordakate puhul ühesugust skaalat. skaalat:

Sõltuvuskordajad, mille puhul *ei arvestata sõltuvuse suunda*, muutuvad (enamasti) lõigul [0, 1]. Väärtus 1 tähistab *maksimaalse tugevusega sõltuvust* (antud sõltuvuse mõttes) ja väärtus 0 vaadeldava sõltuvuse puudumist. Peaaegu alati järeldub statistilisest sõltumatusest, et sõltuvuskordaja väärtus on 0. Nende seosekordajate hulka kuulub regressioonikordaja η, kuid samuti ka enamus assotsiatiivsuse kordajaid.

Sõltuvuskordajad, mille puhul *arvestatakse sõltuvuse suunda*, muutuvad lõigul [–1, 1]. Nende puhul näitab sõltuvuskordaja absoluutväärtus sõltuvuse tugevust, sõltuvuskordaja märk aga seda, kas juhuslike suuruste muutumine on samasuunaline (ühe suurenedes keskmiselt teine samuti suureneb/ teine suureneb suurema tõenäosusega) või vastassuunaline (ühe suurenedes teine keskmiselt väheneb/ väheneb suurema tõenäosusega). Niisuguste seosekordajate hulka kuuluvad kõik monotoonse seose kordajad. Kõige tuntum ja sagedamini rakendatav sõltuvuskordaja – *korrelatsioonikordaja* kuulub samuti sellesse liiki.

7. Kahemõõtmeline jaotus ja statistiline sõltuvus

Eelöeldust järeldub, et juhusliku vektori jaotus ehk kahemõõtmeline jaotus, mida esitab jaotusfunktsioon $F_{XY}(x,y)$, määrab täielikult vek-

tori (X,Y) ühemõõtmelised marginaaljaotused (vastavalt jaotusfunktsioonidega $F_x(x)$ ja $F_y(y)$) ning samuti ka *vektori komponentide vahelise statistilise sõltuvuse*, sh ka kõigi seosekordajate väärtused.
Huvi pakub aga küsimus: kas ja millal määravad marginaaljaotused ja/või statistiline sõltuvus nende vahel kahemõõtmelise jaotuse? Mõnedel juhtudel on selle küsimuse vastused juba ammu hästi teada, kuid on ka seni vähe tuntud tõiku. Loetleme olulisemad nendest.

1. Marginaaljaotused määravad *üheselt sõltumatute komponentidega* kahemõõtmelise jaotuse. Selline kahemõõtmeline jaotus eksisteerib antud marginaaljaotuste korral alati, ning selle jaotusfunktsioon $F(x,y)$ on marginaalsete jaotusfunktsioonide $F(x)$ ja $F(y)$ kaudu määratud seosega (1).
2. Üldjuhul ei määra marginaaljaotused kahemõõtmelist jaotust üheselt. Antud ühemõõtmeliste jaotuste korral leidub *lõpmata palju erinevaid kahemõõtmelisi jaotusi*, mille marginaaljaotusteks need on. Need kahemõõtmelised jaotused erinevad üksteisest statistilise sõltuvuse poolest.
3. Antud marginaaljaotuste paari korral üldjuhul ei eksisteeri suvalise statistilise sõltuvusega kahemõõtmelist jaotust.
4. Kui on teada kahemõõtmeliste jaotuste klass ja selles klassis on *kahemõõtmelist jaotust identifitseerivaks parameetriks sõltuvuskordaja*, siis määravad marginaaljaotused ja sõltuvuskordaja kahemõõtmelise jaotuse üheselt. Tuntuim näide: kahemõõtmeline normaaljaotus, mille määravad marginaaljaotused ja korrelatsioonikordaja.
5. Kahemõõtmelise jaotuse määramiseks on piisav, kui on teada ühe juhusliku suuruse jaotus ja teise juhusliku suuruse *tinglikud jaotused* kõigi esimese juhusliku suuruse väärtuste korral.

8. Kahemõõtmeline maksimaal- ja minimaaljaotus

Nagu öeldud, vastab antud marginaaljaotuste paarile palju erinevaid kahemõõtmelisi jaotus. Nende hulgas on ka kaks äärmist, millele vastavad maksimaalne ja minimaalne võimalik korrelatsioonikordaja (vt Tiit, 91). Neid jaotusi nimetatakse vastavalt kahemõõtmeliseks maksimaal- ja minimaaljaotuseks, ning nad defineeritakse jaotusfunktsioonide järgi.

Definitsioon 5. Olgu $F_x(x)$ ja $F_y(y)$ antud jaotusfunktsioonid. Siis neile vastava *kahemõõtmelise maksimaaljaotuse* määrab jaotusfunktsioon

$$F_{xy}^{+}(x, y) = \min(F_x(x), F_y(y)).$$

●

Definitsioon 6. Olgu $F_x(x)$ ja $F_y(y)$ antud jaotusfunktsioonid. Siis neile vastava *kahemõõtmelise minimaaljaotuse* määrab jaotusfunktsioon

$$F_{xy}^{-}(x, y) = \max(0, F_x(x) - F_y(y)).$$

●

Tähistame maksimaal- ja minimaaljaotuse korrelatsioonikordajad vastavalt sümbolitega r^+ ja r^-. Alati kehtivad järgmised võrratused.

$$-1 \le r^- < 0 < r^+ \le 1,$$

$$r^- \le r \le r^+,$$

kus r tähistab suvalist korrelatsioonikordajat antud marginaaljaotuste korral. Seega näeme, et kaugeltki kõigi antud ühemõõtmeliste jaotuste korral pole võimalik maksimaalse tugevusega korrelatiivne sõltuvus – tõsiasi, mis pole rakendusstatistikute seas sugugi alati teadvustatud.

Kokkuvõttes saime järgmised tulemused.

1. Iga marginaaljaotuste paari jaoks saab määrata *maksimaaljaotuse* ja *minimaaljaotuse*.
2. Minimaaljaotuse korrelatsioonikordaja on väikseim ja maksimaaljaotuse korrelatsioonikordaja suurim võimalik (antud marginaaljaotuste puhul).
3. Kui on antud marginaaljaotused ja korrelatsioonikordaja r vähima ja suurima korrelatsioonikordaja vahel, siis leidub lõpmata palju erinevaid kahemõõtmelisi jaotusi antud marginaaljaotuste ja antud korrelatsioonikordajaga.

Näide 1. Olgu vaadeldavas rühmas 9 tütarlast vanustega 8–16 aastat, kusjuures nende tütarlaste pikkused on 125, 127, 136, 149, 152, 157, 160, 169 ja 170 cm. Vanuse ja pikkuse ühine maksimaaljaotus on esitatud tabelis 1.

Tabel 1.
Diagonaalne maksimaaljaotus

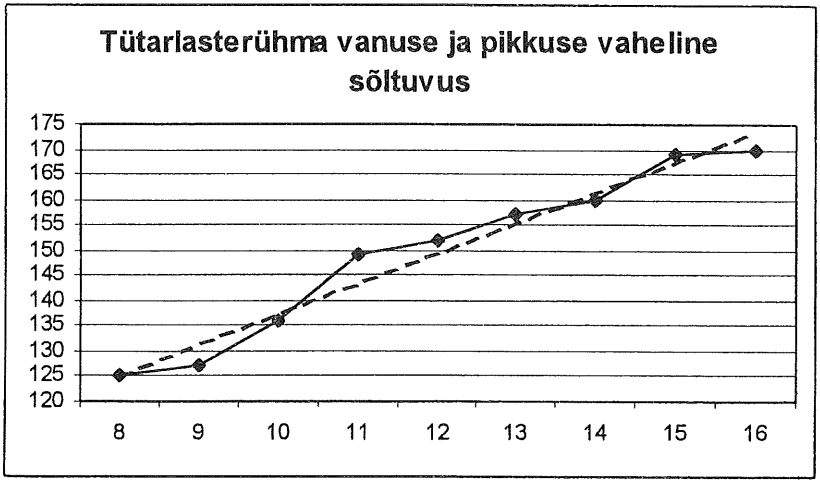
Vanus	*Pikkus*								
	125	127	136	149	152	157	160	169	170
8	**1**	0	0	0	0	0	0	0	0
9	0	**1**	0	0	0	0	0	0	0
10	0	0	**1**	0	0	0	0	0	0
11	0	0	0	**1**	0	0	0	0	0
12	0	0	0	0	**1**	0	0	0	0
13	0	0	0	0	0	**1**	0	0	0
14	0	0	0	0	0	0	**1**	0	0
15	0	0	0	0	0	0	0	**1**	0
16	0	0	0	0	0	0	0	0	**1**

Leiame maksimaaljaotuse põhjal lineaarse mudeli tütarlapse pikkuse arvutamiseks tema vanuse järgi,

$$Pikkus = 77 + 6 \times vanus.$$

See mudel on väga hea (seos on tugev), sest lihtne arvutus näitab, et seda iseloomustav korrelatsioonikordaja on 0,98. Selle põhjuseks on asjaolu, et me vaatleme maksimaaljaotust, seega on tegemist suurima korrelatsiooniga, mis nende andmete korral saab üldse esineda. Siiski pole selle korrelatsioonikordaja väärtus 1. Ühest erineb see sellepärast, et isegi maksimaaljaotuse korral ei avaldu pikkus täpselt vanuse lineaarfunktsioonina, nagu näha joonisel 1, kus lineaarne mudel on esitatud katkendliku joonega.

Seevastu on meie andmete põhjal pikkus vanuse *monotoonselt kasvav* funktsioon, st et vanem tütarlaps on kindlasti pikem nooremast. Selletõttu on ka maksimaaljaotuse korral vanuse ja pikkuse vahelist monotoonset seost mõõtvate seosekordajate (näiteks Spearmani ρ) väärtus 1.



Joonis 1.

Näide 2. Olgu nendest tütarlastest kolm pärit üheliikmelisest, neli – kaheliikmelisest ja kaks kolmeliikmelisest perest. Leiame laste arvu ja vanuse minimaaljaotuse, see on esitatud tabelis 2.

Tabel 2.
Minimaaljaotus

Laste arv peres	*Vanus*								
	8	9	10	11	12	13	14	15	16
1	0	0	0	0	0	0	1	1	1
2	0	0	1	1	1	1	0	0	0
3	1	1	0	0	0	0	0	0	0

Selle minimaaljaotuse põhjal arvutatud minimaalse korrelatsioonikordaja väärtuseks on –0,93. Negatiivne on see sellepärast, et minimaaljaotuse korral on suuremad lapsed on pärit väiksemast perest ja vastupidi. Siiski pole minimaalse korrelatsioonikordaja väärtus –1. Näeme taas, et reaalse andmetiku korral ei tarvitse maksimaalne ja minimaalne korrelatsioonikordaja saavutada vastavalt teoreetiliselt suurimat ja väiksemat väärtust 1 ja –1.

9. Mitmemõõtmeline juhuslik vektor.

Olgu $k > 2$ ja olgu meil tegemist k-mõõtmelise ($k > 2$) juhusliku vektoriga

$$X = (X_1, X_2, \ldots, X_k).$$

Definitsioon 7. Juhuslik vektor X on siis statistiliselt sõltumatute komponentidega, kui on täidetud tingimus

$$F(x_1, \ldots, x_k) = F(x_1) \ldots F(x_k),$$

kus F_i (.) tähistab i-nda komponendi jaotusfunktsiooni. Alati, kui see tingimus ei ole täidetud, on tegemist *sõltuvate komponentidega* juhusliku vektoriga. ●

Ühtekokku on k-mõõtmelisel juhuslikul vektoril k ühemõõtmelist, $k(k-1)/2$ kahemõõtmelist, ...,k $k-1$-mõõtmelist marginaaljaotust, seega kokku $2^k - k - 2$ erinevat marginaaljaotust, millest kõigi puhul on võimalik statistilise sõltuvuse olemasolu või puudumine.

Definitsioon 8. Kui sõltumatute komponentidega juhusliku vektori kõik marginaaljaotused on sõltumatute komponentidega, öeldakse et vektor on *täielikult sõltumatute komponentidega*. ●

Üldjuhul ei järeldu vektori ühe marginaaljaotuse sõltuvusest või sõltumatusest teise marginaaljaotuse sõltuvus või sõltumatus. Ometi on marginaaljaotuste sõltuvused omavahel teatava määrani seotud, seega mõjutavad vektori ühe marginaaljaotuse statistilist sõltuvust üldiselt selle vektori teised marginaaljaotused ja nende sõltuvuse iseloom.

10. Regressioon- ja korrelatiivne sõltuvus mitmemõõtmelise juhuliku vektori puhul

Regressioonsõltuvus on igasuguste statistiliste mudelite aluseks ka mitmemõõtmeliste juhuslike vektorite puhul. Nimetame juhusliku vektori ühe komponendi *funktsioontunnuseks* (olgu see näiteks $Y = X_i$) *ja* ülejäänud $k - 1 = p$ komponenti moodustavad *argumenttunnusvektori*, mille tähis on $\vec{X} = (X_1, \ldots, X_p)$. Siis me võime leida sarnaselt valemile (2) juhusliku suuruse Y *parima prognoosi* argumenttunnusvektori $\vec{X}$ järgi,

$$E(Y / \vec{X} = \vec{x}) = Y_x . \qquad (6)$$

Siin on prognoosiks Y tinglik keskväärtus, kus tingimus on määratud igas argumenttunnusvektori väärtuste ruumi punktis, seega parim prognoos on p argumendi funktsioon. Regressioonsõltuvuse tugevust mõõdab *mitmene regressioonisuhe*, mis määratakse samuti valemiga (3).

Praktiliselt kasutatavate mudelite konstrueerimiseks tuleb seosega (6) määratud funktsioon lähendada mingi funktsiooniga sobivalt valitud küllalt lihtsast *parameetriliste funktsioonide klassist*. Mudeli argumendiks pole tihti tarvis valida argumenttunnusvektorit täies ulatuses, vaid piisab selle mingist alamvektorist (mida kirjeldab üks marginaaljaotustest). Niisuguseid võimalikke marginaaljaotusi on kokku $2^{p-1}-2$, ja kui arvestada ka konstantset mudelit, mis vastab regressioonsõltuvuse puudumisele, siis on neid ühe võrra rohkem.

Mida rohkem on regressioonimudelis argumente, seda parem on üldiselt mudel – vähemalt teoreetiliselt, sest argumendi lisamisel mudelisse saab mitmene regressioonsuhe ainult suureneda. Praktiliselt aga parandavad lisatud argumendid sageli mudelit nii vähe, et nende lisamine ei ole otstarbekas, sest alati on lihtsam, vähem argumente sisaldav mudel parem kui keerukam mudel, milles on rohkem argumente ja vastavalt ka hinnatavaid parameetreid.

Definitsioon 9. Lähendades valemiga (6) esitatud parimat prognoosi Y_k argumenttunnuste lineaarfunktsiooniga

$$Y \approx \beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r$$

saame mudeli, mida nimetatakse *mitmeseks lineaarseks regressioonimudeliks*. ●

Oma lihtsuse tõttu on mitmene lineaarne regressioonimudel üks praktikas enimkasutatavaid mudeleid. Mitmese lineaarse regressioonsõltuvuse tugevust mõõdab *mitmene korrelatsioonikordaja* R, mis arvutatakse samuti valemist (3). Erinevalt korrelatiivsest sõltuvusest kahe juhusliku suuruse vahel on korrelatiivne sõltuvus juhusliku vektori ja juhusliku suuruse vahel alati mittenegatiivne.

11. Mitmemõõtmelise sõltuvusstruktuuri esitus marginaalsõltuvuste kaudu. Korrelatsioonimaatriks

Mitmene korrelatsioonikordaja on statistilise sõltuvuse mõõtmiseks kasutatav sel juhul, kui juhusliku vektori komponentide hulgast üks on funktsioontunnus, mille prognoosimine ülejäänute kaudu pakub huvi. Kui aga niisugust üht komponenti ei ole, tuleb vaadelda kõigi komponentide sõltuvusi kõigist, mida kajastab kokku $k \cdot 2^{k-1}$ regressioonisuhet. See arv kasvab juhusliku vektori komponentide arvu kasvades kiiresti. Suur, raskesti süstematiseeritav kordajate hulk ei sobi hästi ülevaate saamiseks juhusliku vektori komponentidevahelise sõltuvuse struktuurist. Seetõttu püütakse saada ülevaadet juhusliku vektori komponentide vahelistest sõltuvustest, kasutades üksnes madalamat järku marginaaljaotuste sõltuvusi.

Kõige sagedamini võetakse aluseks *jaotuse kahemõõtmelised marginaaljaotused*, mille arv on $k(k-1)/2$. Kahemõõtmelisi jaotusi iseloomustavate erinevate regressioonisõltuvuste arv on $k(k-1)$, korrelatiivsete sõltuvuste arv aga $k(k-1)/2$. Tuntuimaks mitmemõõtmelise sõltuvusstruktuuri karakteristikuks on korrelatsioonimaatriks.

Definitsioon 10. Maatriksit, mille i-nda rea ja j-nda veeru elemendiks on vastavad korrelatsioonikordajad r_{ij} ($i, j = 1, \ldots, k$), st korrelatsioonikordajad juhusliku vektori i-nda ja j-nda komponendi vahel, nimetatakse *korrelatsioonimaatriksiks*. ●

Korrelatsioonimaatriksi kõige tavalisemaks sümboliks on R ja tal on järgmised omadused:

$$R = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1k} \\ r_{21} & 1 & \ldots & r_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ r_{k1} & r_{k2} & \ldots & 1 \end{pmatrix}$$

1. Korrelatsioonimaatriks on sümmeetriline $k \times k$ maatriks (kus k on juhusliku vektori järk).
2. Korrelatsioonimaatriksi elementide väärtused paiknevad lõigul $[-1, 1]$ ja tema peadiagonaalil on ühed.
3. Korrelatsioonimaatriks on mittenegatiivselt määratud.

4. Kui korrelatsioonimaatriks on singulaarne, siis avaldub (vähemalt) üks juhusliku vektori komponentidest ülejäänute lineaarkombinatsioonina.

Seoses korrelatsioonimaatriksiga pakuvad huvi järgmised küsimused:

1. Kui on antud marginaaljaotused ja korrelatsioonimaatriks, kas siis on võimalik konstrueerida nende marginaaljaotuste ja korrelatsioonimaatriksiga juhuslikku vektorit?
2. Kuidas tunda ära, kas mingi sümmeetriline, ühest väiksemate elementidega maatriks saab olla korrelatsioonimaatriks või mitte?

12. Märkida muster ja piirkorrelatsioon

Püüame viimasele küsimusele vastuse leida lihtsamal juhul. Vaatleme korrelatsioonimaatriksit, milles kõik korrelatsioonikordajad on absoluutväärtuselt võrdsed, olgu selleks ühiseks absoluutväärtuseks ρ. Sel juhul määrab korrelatsioonimaatriksi tema *märkide muster* S, mille elementideks on ühed ja miinus ühed,

$$s_{ij} = \operatorname{sgn}(r_{ij}),$$

Sel juhul on korrelatsioonimaatriks R määratud eeskirjaga

$$R = I + (S - I)\,\rho, \tag{8}$$

seega

$$R = \begin{pmatrix} 1 & \pm\rho & \dots & \pm\rho \\ \pm\rho & 1 & & \pm\rho \\ \dots & \dots & \dots & \\ \pm\rho & \pm\rho & \dots & 1 \end{pmatrix}$$

Definitsioon 11. Suurimat võimalikku ρ väärtust $\rho+$, mille korral R on mittenegatiivselt määratud, nimetatakse märkide mustrile S vastavaks *piirkorrelatsiooniks*. •

Osutub, et juhul kui kõik korrelatsioonikordajad erinevad nullist, on erinevate märgimustrite arv $2^{k(k-1)/2}$, vt Helemäe, Tiit, 1997. Piirkorrelatsioonide arv on aga märksa väiksem. On tähelepanuväärne, et

märgimustrite ja piirkorrelatsioonide korral kehtivad järgmised seaduspärasused (vt. H.-L. Viirsalu, 1998).

1. Iga maatriksi järgu k korral leidub 2^{k-1} märkide mustrit, mille puhul piirkorrelatsiooni väärtuseks on 1, nende seas ka maksimaalne korrelatsioonimaatriks, mille kõik elemendid võrduvad ühega.
2. Kõigi ülejäänud märgimustrite puhul on piirkaorrelatsioonid märksa väiksemad. Võib oletada, et nende piirkorrelatsioonide väärtused ei ületa arvu 0.5.
3. Märgimustrid moodustavad ekvivalentsiklassid, millesse kuuluvatele märgimustritele vastab sama piirkorrelatsioon. Igasse eksivalentsiklassi kuulub vähemalt 2^{k-1} erinevat märgimustrit.
4. Märgimustrile, mis sisaldab ainult positiivseid elemente, vastav väikseim võimalik korrelatsioonikordaja on $-1/(k-1)$, kus k tähistab maatriksi järku. Sellele vastavat korrelatsioonimaatriksit nimetatakse *minimaalseks korrelatsioonimaatriksiks.*
5. Kui ρ rahuldab tingimust

$$|\rho| \le \frac{1}{k-1},$$

siis on iga märkide mustri S korral maatriks (8) korrelatsioonimaatriks.

Näide 3. Esitame näitena kaks märkide mustrit. Osutub, et vasakpoolsele (S_1) vastab piirkorrelatsioon 0,5, parempoolsele (S_2) aga 1.

$$S_1 = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad S_2 = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix}.$$

13. Mitmemõõtmeline maksimaaljaotus

Tekib küsimus, kas maksimaal- ja minimaaljaotuse mõiste on üldistatav ka rohkem kui kahemõõtmelise vektori jaoks. Osutub, et kõrgema kui kahemõõtmelise jaotuse puhul ei ole olukord maksimaal- ja minimaaljaotuse suhtes enam sümmeetriline. Maksimaaljaotuse mõiste on lihtsalt üldistatav, minimaaljaotuse oma aga üldjuhul mitte.

Definitsioon 11. Olgu $F_1(x_1),\ldots,F_k(x_k)$ antud jaotusfunktsioonid. Siis neile vastava *k-mõõtmelise maksimaaljaotuse* määrab jaotusfunktsioon

$$F^+(x_1,\ldots,x_k) = \min\left[F_1(x_1),\ldots,F_k(x_k)\right].$$

•

Maksimaaljaotusel on järgmised omadused (vt Helemäe, Tiit, 1997):

- Maksimaaljaotuse puhul kehtib võrratus

$$F^+(x_1,\ldots,x_k) \geq F(x_1,\ldots,x_k)$$

  igas ruumi R^k punktis, kui $F(.)$ on mingi samade marginaaljaotustega k-mõõtmeline jaotusfunktsioon.
- Maksimaaljaotuse kõik marginaaljaotused on maksimaaljaotused.
- Maksimaaljaotuse korrelatsioonimaatriksi elementideks on kahemõõtmeliste marginaaljaotuste maksimaalsed korrelatsioonikordajad.
- Erijuhul, kui kõik ühemõõtmelised marginaaljaotused on võrdsed ja sümmeetrilised, on maksimaaljaotuse korrelatsioonimaatriksiks maksimaalne korrelatsioonimaatriks (milles on kõik elemendid võrdsed ühega).

Regressioonsõltuvuse puhul tõstsime juhusliku suuruse komponentide seast esile ühe – funktsioontunnuse. Mõne ülesande puhul tekib küsimus, kas on võimalik määratleda *maksimaalse tugevusega statistilist sõltuvust,* mis seoks kõiki juhusliku suuruse komponente ilma ühtki neist esile tõstmata. Niisuguse seose määratleb maksimaaljaotus.

Näide 4. Vaatleme veel näites 1 tutvustatud tütarlasterühma. Oletame, et meil on teada, et neist 1 käib esimeses ja üks kolmandas klassis. Niihästi kuuendas, seitsmendas kui ka üheksandas klassis käib samuti üks tüdruk, aga neljandas ja kümnendas klassis käib kaks tüdrukut. Leiame nüüd kolme tunnuse ühise maksimaaljaotuse. See täidab nüüd kolmemõõtmelise tabeli (risttahuka) diagonaali ning selles on nullist erinevaid lahtreid üheksa, need on antud alljärgnevas tabelis 3.

Tabel 3.
Kolme tunnuse maksimaaljaotuse diagonaalielemendid:

Vanus	*Pikkus*	*Klass*	*arv*
8	125	1	1
9	127	3	1
10	136	4	1
11	149	4	1
12	152	6	1
13	157	7	1
14	160	9	1
15	169	10	1
16	170	10	1

On näha, et sellise jaotuse korral on kõigi tunnuste vahelised korrelatiivsed seosed väga tugevad, ent ei saavuta siiski taset 1.

14. Mitmemõõtmeline minimaaljaotus

Kahemõõtmelise juhusliku vektori korral eksisteerib alati niihästi maksimaalne kui ka minimaalne jaotus. Üldjuhul see aga nii ei ole, vt Helemäe, Tiit, 1996. Kolme tunnuse X, Y ja Z puhul on võimalik küll leida niisugune jaotus, kus X ja Y ühisjaotus on minimaalne ning X ja Z ühisjaotus on samuti minimaalne, kuid sel juhul pole võimalik üldjuhul Y ja Z jaotust muuta minimaalseks.

Veelgi enam, on selgunud, et kõrgemamõõtmeliste jaotuste klassis üldiselt minimaaljaotust ei eksisteeri. Minimaaljaotuste kohta saab tuua üksnes kaunis erandlikke näiteid. Ka minimaaljaotuste omadused erinevad maksimaaljaotuste omadustest, vt Helemäe, Tiit, 1997.

- Minimaaljaotuse marginaaljaotused ei ole üldiselt minimaal-jaotused.
- Jaotus, millel on minimaalne korrelatsioonimaatriks, ei tarvitse olla minimaaljaotus.

Näide 5. Vaatleme taas oma tütarlaste rühma näidet ja veendume selles, et üldjuhul pole võimalik konstrueerida jaotust, kus kolm kahemõõtmelist marginaaljaotust oleksid kõik minimaalsed.
Võime seda katsetada kasvõi oma tüdrukute näitel, kasutades lisaks vanusele ja laste arvule peres veel näiteks isa palka, mille suurus olgu ühel tüdrukul 4000, kahel 5000, kolmel 7000, kahel 9000 ja ühel 10000 krooni kuus. Leiame tütarlapse vanuse ning isa palga minimaaljaotuse, see on esitatud alljärgnevas tabelis.

Tabel 4.
Kahe tunnuse minimaaljaotus

Isa palk	*Vanus*								
	8	9	10	11	12	13	14	15	16
4000	0	0	0	0	0	0	0	0	1
5000	0	0	0	0	0	0	1	1	0
7000	0	0	0	1	1	1	0	0	0
9000	0	1	1	0	0	0	0	0	0
10000	1	0	0	0	0	0	0	0	0

Hakkame nüüd koostama laste arvu ja isa palga kahemõõtmelist jaotust (tabelit 5) ning vaatame, kui palju meil selleks mänguruumi on, arvestades, et jaotus peab olema kooskõlas niihästi tabeliga 2 kui ka tabeliga 4.

Tabel 5.
Laste arvu ja isa palga kahemõõtmeline jaotus

Isa palk	*Laste arv peres*		
	1	2	3
4000	1		
5000	2		
7000		3	
9000		1	1
10000			1

Näeme, tabelist 4, et 4000 kroonise palgaga on ainult ühe 16-aastase tütarlapse isa. Tabelist 2 aga on näha, et selle lapse peres on 1 laps. Nii saamegi täita tabelis 5 ülemise vasakpoolse lahtri. 5000-kroonise palgaga on kahe tütarlapse isad, nende mõlema peres on aga tabeli 2 andmetel 1 laps. Saame täita järgmise lahtri, sinna tuleb number 2. 7000 kroonise palgaga on kolme lapse isa, kusjuures tabelist 2 selgub, et need lapsed on kahelapselisest perest. Samal viisil jätkates saame tabeli 5 lõpuni täita. Kui me nüüd arvutame kahe tunnuse – laste arvu ja isa palga vahelise korrelatsiooni, leiame, et selle väärtus on positiivne ja pealegi üpris kõrge: 0,93. Sellest näitest saame teha mõned olulised järeldused:

- Kui mingi kolmemõõtmelise juhusliku vektori kaks marginaaljaotust on antud, siis ei saa üldjuhul kolmas marginaaljaotus muutuda täiesti suvaliselt.
- Rohkem kui kahe juhusliku tunnuse korral pole võimalik defineerida jaotust, kus kõik tunnused oleksid omavahel maksimaalse tugevusega negatiivselt korreleeritud.

Viimane tõik pole küll eriti sageli kommenteeritud matemaatilises statistikas, kuid on seevastu üsna tuttav igapäevaelus.

Viitekirjandus

1. E.-M. Tiit. Jaotuste segude kasutamine mitmemõõtmeliste jaotuste konstrueerimisel ja genereerimisel. EMS Aastaraamat, 1991, 22–39.
2. E.-M. Tiit, H.-L. Helemäe. Multivariate minimal distributions. Eesti TA Toimetised. Füüsika, Matemaatika, **45**, 1996, 317–322.
3. E.-M. Tiit, H.-L. Helemäe. Boundary distributions with fixed marginals. – Distributions with Given Marginals and Moment Problems. Kluwer AP, 1997, 99–106.
4. Viirsalu, H.-L. Mitmemõõtmeliste sõltuvusstruktuuride esitus korrelatsioonimaatriksite abil. Magistritöö, Tartu, 1998.

Avalik loeng Matemaatikateaduskonnas 30. okt. 1998.

Tartu Ülikooli lõpetajad matemaatilise statistika erialal

1998. aasta kevadel lõpetasid Tartu Ülikooli põhiõppe matemaatilise statistika erialal järgmised noored:

1. **Tarmo Kiivit**. Aktsiahindade juhuslikkuse kontroll Eesti Hoiupanga näitel. Juhendaja lektor Martin Viil.
2. **Kristi Kuljus**. Geomeetriliste kvantiilide arvutamine. Juhendaja dotsent Tõnu Kollo.
3. **Meelis Käärik**. Jaotuste k-tsentrite koondumine ja karakteristlikud funktsioonid. Juhendaja professor Kalev Pärna.
4. **Andres Kütt**. Eesti aktsiaturu statistiline analüüs: börsikrahhi mõjud. Juhendaja professor Kalev Pärna.
5. **Külli Laanet**. Regressioonimudeli valiku meetodid. Juhendaja teadur Krista Fischer.
6. **Kandela Puss**. Loote ja vastsündinu kaalu prognoosimine ultrahelimõõtmiste põhjal. Juhendajad professor Ene-Margit Tiit ja professor Helje Kaarma.
7. **Kaire Ruul**. Korrelatsioonimaatriksi jaotuse lähendamine. Juhendaja dotsent Tõnu Kollo
8. **Agne Soome**. Rahvastiku muutumise mudelid. Juhendaja professor Ene-Margit Tiit.
9. **Gabriel Villers**. Optimaalsed tunnuskomplektid lünkliku andmestiku korral. Juhendaja dotsent Tõnu Möls.

Magistriväitekirjade kaitsmine TÜ Matemaatilise statistika instituudis

1998. aastal kaitsti Tartu Ülikooli matemaatilise statistika instituudis järgmised magistriväitekirjad:

1. **Säde Koskel**. *Empiirilise andmestiku statistilise analüüsi strateegia antropomeetrilise materjali näitel.* MSc (matemaatiline statistika), juhendaja prof Ene-Margit Tiit.
2. **Anneli Kukk**. *Hinnangu dispersioon samade marginaaljaotustega valikudisainide klassil.* MSc (matemaatiline statistika), juhendaja dots Imbi Traat.
3. **Anne Pirn**. *Elukindlustusmatemaatika I.* MSc (matemaatiline statistika), juhendaja dots Tõnu Kollo.
4. **Hele-Liis Viirsalu**. *Mitmemõõtmeliste sõltuvusstruktuuride esitusest C-korrelatsiooni-maatriksite kaudu.* MSc (matemaatiline statistika), juhendaja prof Ene-Margit Tiit.

STATISTIKAMEETODID KESKKONNAKAITSES JA ÖKOLOOGIAS

Kuidas modelleerida sesoonsust beetajaotusega?
Kas Eesti metsad on hakanud kiiremini kasvama?
Milline on Võrtsjärve ökoloogiline seisund?
Mida mõõdab Gini kordaja?
Kuidas kirjeldada tunnuste statistilist sõltuvust?

Kõigile neile ja paljudele teistele küsimustele leiate vastuse
Eesti Statistikaseltsi Teabevihikust nr 11

Lisaks sellele
Kroonika: TÜ bakalaureuse- ja magistritööd
matemaatilise statistika erialal 1998. aastal