TARTU ÜLIKOOL

Matemaatika-informaatikateaduskond
Matemaatilise statistika instituut

Marina Haldna

RUUMILISED KOVARIATSIOONISTRUKTUURID JA NENDE KASUTAMINE VEE KVALITEEDI-NÄITAJATE MODELLEERIMISEL

Magistritöö

Juhendajad
Märt Möls (TÜ)
Tõnu Möls (EMÜ)

Tartu 2007

SISUKORD

SISSEJUHATUS

Arvutustehnika ja statistikatarkvara areng on teinud võimalikuks segamudelite kasutamise seal, kus see varem praktilistel põhjustel polnud võimalik. Eesti Maaülikooli Limnoloogiakeskuses on Peipsi järve vee andmetel välja töötatud rida lineaarseid mudeleid, mis ei ole seni kasutanud segamudelite poolt pakutavaid lisavõimalusi. Üle kahekümne aasta pidevalt täiendatud Peipsi mudelites on olnud asukoha mõju sees lineaarselt, fikseeritud koordinaatidena (vt Möls, T. jt. [10,11] ja antud töö punkt 3.1.1).

Käesolevas magistritöös uuritakse, kuidas saaks enam infot pakkuvaid, kuid teoreetiliselt keerukaid segamudeleid kasutada ruumiliste kordusmõõtmiste mõjude arvestamiseks. Töö eesmärkideks on selgitada, kuidas proovipunktide sarnasus (aine sisaldus) sõltub nende punktide vahelisest kaugusest; kontrollida, kas selle sarnasuse arvestamine võimaldab saada täpsemaid tulemusi kui seniste kasutatud mudelite abil saadud tulemused; valida välja parim mudel ja kontrollida selle headust; lõpuks, rakendades parimat mudelit, uurida järveteadlastele huvipakkuvaid prognooside kaardistamise võimalusi. Uurimisaluseks materjaliks valiti üldfosfori kui ühe kõige rohkem järve "headuse" seisundit iseloomustava näitaja andmed. Fosfori kontsentratsioonist sõltuvad mitmed teisedki vee bioloogilised näitajad (vetikate biomass, taimede kõrgus jms), samas on fosfori sisaldus küllaltki varieeruv.

Magistritöö esimeses peatükis on toodud lineaarsete segamudelite üldine kuju, esitatud on segamudelite hindamiseks vajalikud mõisted, erilist tähelepanu on pööratud funktsioontunnuse väärtuste prognoosimisele korreleeruvate jääkidega mudeli korral.

Teises peatükis on toodud mudeli valiku põhiprintsiibid korreleeruvate andmete korral; kovariatsioonistruktuuride olemus ja ruumiliste andmete analüüsiks kasutatavate struktuuride kirjeldused; simulatsioon kontrollimaks, kuidas kovariatsioonistruktuuri valikust sõltub parameetrite prognooside täpsus.

Kolmandas peatükis on esitatud kokkuvõte Peipsi järve mudelite ajaloost, seejärel uute, kovariatsioonistruktuuridega mudelite testimine, nende võrdlemine nii omavahel kui ka senikasutatud lineaarse mudeliga, lõpuks tulemuste analüüs ja nende praktilise kasutamise näited.

Kirjandusest on teada mitmed rakendused segamudelite kasutamisest juhuslike faktorite mõjude hindamisel. Näiteks meditsiinis huvitutakse ravimite mõjust juhuslikele patsientide (kordusmõõtmised) või juhuslikult valitud haiglates ravitud patsientidele (mitmetasemeline

mudel) (vt Diggle jt. [3], Littell jt. [7]). Tõuaretuses uuritakse lehmade (juhuslike) vanemate mõju produktiivsusele (Henderson [5], Robinson [13]). Tänu SAS paketi MIXED protseduuri täiendamisele REPEATED käsuga on võimalik hinnata kordusmõõtmistel tekkivaid korrelatsioonistruktuure, mis tekivad juhul, kui mudeli jäägid ei ole sõltumatud nagu tavaliste fikseeritud faktoritega lineaarsete mudelite korral, vaid korreleeruvad mingi ajalise või ruumilise edasiliikumise tõttu. Ruumiliste andmete modelleerimisel on segamudelite teooriat veel vähe kasutatud, magistritöö autoril kirjandusest näidisanalüüsi leida ei õnnestunudki.

Lisatud on kasutatud kirjanduse loetelu ning lisad (SAS protseduuride kirjeldused ning programmi tekst).

Magistritöö on valminud tänu Riikliku sihtfinatseerimisteema 0362483s03 ja ETF grantide nr. 6008 ja 6820 toetusele. Autor tänab juhendajaid, töökaaslasi ja perekonda kannatlikkuse ja mõistva suhtumise eest.

1. LINEAARSED SEGAMUDELID

1.1 Lineaarse segamudeli üldkuju

Maatriksesituses kirja panduna on segamudel järgmise kujuga:

$$Y = X\beta + Z\eta + \varepsilon \quad , \tag{1.1}$$

kus Y on $n\times 1$ vaatluste vektor, X on $n\times p$ fikseeritud faktorite disaini- ehk plaanimaatriks, β on tundmatu $p\times 1$ parameetervektor, mille liikmed on hinnatavad kui fikseeritud kordajad, Z on etteantud juhuslike faktorite $n\times q$ disainimaatriks, ning η on tundmatu $q\times 1$ parameetervektor, mille liikmed modelleeritakse kui juhuslikud suurused, ning ε on $n\times 1$ prognoosijääkide vektor. Eeldatakse, et $E\eta = E\varepsilon = 0$, seega $E(Y) = X\beta$, ning et η ja ε on omavahel sõltumatud. Plaanimaatriksid X ja Z seostavad vastavalt fikseeritud ja juhuslike efektide mõju Y jaoks iga vaatluse korral. Segamudel on lineaarse mudeli edasiarendus. Kui fikseeritud faktoritega lineaarsete mudelite korral eeldatakse, et prognoosivead on omavahel sõltumatud ja sama dispersiooniga, siis segamudelite korral lubatakse vigade korreleeritust ning võimalik on hinnata juhuslike faktorite mõjusid. Toome sisse järgmised tähistused:

juhuslike kordajate kovariatsioonimaatriks

$$G := Var(\eta),$$

mudeli jääkide kovariatsioonimaatriks

$$R := Cov(\varepsilon),$$

juhusliku suuruse Y kovariatsioonimaatriks

$$V := Cov(Y) = ZGZ^T + R.$$

Eeldame, et G ja R on mittekõdunud maatriksid.

Lisades kordajate keskväärtustele tehtud eeldused, saame kokkuvõttes

$$E\begin{pmatrix} Y \\ \eta \\ \varepsilon \end{pmatrix} = \begin{pmatrix} \mathrm{X}\beta \\ 0 \\ 0 \end{pmatrix},\ Var\begin{pmatrix} Y \\ \eta \\ \varepsilon \end{pmatrix} = \begin{pmatrix} V & ZG & R \\ GZ^T & G & 0 \\ R & 0 & R \end{pmatrix}. \tag{1.2}$$

Enamasti pole G ja R täpselt teada, vaid sõltuvad tundmatutest kovariatsiooniparameetritest. Mõnikord eeldatakse juhusliku suuruse Y normaaljaotust, sellisel eeldusel on G ja R hindamise metoodika juba pikaajaliselt väljatöötatud. Nimelt, kui $Y \sim N(X\beta, V)$, kus V sõltub tundmatutest parameetritest, kasutatakse kovariatsiooniparameetrite hindamiseks kõige sagedamini suurima tõepära meetodit (*ML*) või selle modifikatsiooni – kitsendatud suurima tõepära meetodit (*REML*). Matemaatilise statistika ja tarkvara kiire areng võimaldab hinnata kovariatsioonimaatrikseid juba ka teiste jaotustega algtunnuste korral.

1.2 Parameetrite hindamine ja prognoosimine

Üldise lineaarse mudeli $Y = X\beta + \varepsilon$ korral, kus eeldatakse, et $Var(\varepsilon) = \sigma^2 I$ (st prognoosivead on omavahel sõltumatud ja sama dispersiooniga), on fikseeritud parameetervektori β hindamiseks võimalik kasutada vähimruutude meetodit. Olgu y juhusliku suuruse Y realisatsioonide vektor. Vähimruutude hinnang β jaoks on siis kujul $\hat{\beta} = (X^T X)^- X^T y$, $\hat{\beta}$ ei pruugi olla aga üheselt määratud. Juhul kui $X^T X$ on pööratav, siis on $\hat{\beta}$ üheselt määratud ($\hat{\beta} = (X^T X)^{-1} X^T y$) nihketa ($E(\hat{\beta}) = \beta$), lineaarne ja vähima dispersiooniga ($Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$). Suurus $X\hat{\beta} = X(X^T X)^- X^T y$ on aga alati üheselt määratud, parim, nihketa, lineaarne hinnang suurusele $X\beta$ – matemaatilise statistika teoorias tuntud ka kui *BLUE (Best Linear Unbiased Estimator*) (täpsemalt võib lugeda näiteks Isotalo jt. [6], Möls, M. [9] või Searle jt [14]).

Üldise lineaarse segamudeli (1.1) korral, kasutades dispersioonimaatriksi tähistusi (1.2), on parim lineaarne nihketa hinnang *BLUE* $X\beta$ jaoks kirja pandav kujul

$$X\hat{\beta} = X(X^T V^{-1} X)^- X^T V^{-1} y .$$

Antud hinnang langeb kokku üldistatud vähimruutude meetodil saadud hinnanguga. Lineaarsete segamudelite üks eeliseid tavaliste lineaarsete mudelite ees on võimalus hinnata kovariatsioonimaatriksite parameetreid. Kovariatsioonimaatriksid on kasutusel nii juhuslike faktorite mõju leidmisel, jääkide struktuuri hindamisel kui ka funktsioontunnuse mistahes väärtuste prognoosimisel. Sageli on mudelite otstarve prognoosida uuritava tunnuse väärtusi mingites seni fikseerimata tingimustes, kas siis tulevikus, olemasolevate

mõõtmiste vahelisel ajal või kohas jms (nimetame neid edaspidi uuteks väärtusteks). Uuritava segamudeli (1.1) korral oleme huvitatud uue juhusliku suuruse

$$Y_{uus} = X_{uus}\beta + Z_{uus}\eta + \varepsilon_{uus} \tag{1.3}$$

prognoosi $\hat{Y}_{uus}$ leidmisest. Uue väärtuse prognoosimisel on segamudelite puhul tarvis täpsemalt kirjeldada, kuidas seostuvad vaadeldud väärtused (valim) ja uued prognoositavad suurused. Toome sisse järgnevad tähistused:

$$\begin{aligned}
R_{uus} &:= Cov(\varepsilon_{uus}); \\
K &:= Cov(\varepsilon_{uus}, \varepsilon); \\
V_{uus} &:= Cov(Y_{uus}) = Z_{uus} G Z_{uus}^{\ T} + R_{uus}; \\
C &:= Cov(Y_{uus}, Y) = Cov(Z_{uus}\eta + \varepsilon_{uus}, Z\eta + \varepsilon) \\
&\quad = Z_{uus} G Z^T + K; \\
C^T &:= Cov(Y, Y_{uus}) = Cov(Z\eta + \varepsilon, Z_{uus}\eta + \varepsilon_{uus}) \\
&\quad = Z G Z_{uus}^{\ T} + K^T.
\end{aligned}$$

Prognoosimisel peame ka täpsustama otsitava prognoosi omadusi. Juhuslike muutujate puhul öeldakse, et parim prognoos $\hat{Y}_{uus}$ juhuslikule suurusele Y_{uus} on selline, mille korral kehtib $E(\hat{Y}_{uus} - Y_{uus})^T(\hat{Y}_{uus} - Y_{uus}) \le E(\tilde{Y}_{uus} - Y_{uus})^T(\tilde{Y}_{uus} - Y_{uus})$, kus $\tilde{Y}_{uus}$ on mistahes teine juhusliku suuruse Y_{uus} prognoos. Juhuslike muutujate puhul ei ole seega prognoosi headuse näitaja enam mudeli jääkide minimaalne dispersioon, nagu fikseeritud parameetrite hinnangute korral, vaid keskmise ruutvea minimaalsus.

Kasutades kirjanduses tõestatud tulemusi (vt näiteks Searly [14]) võib väita, et juhusliku suuruse Y mistahes mõõtmata väärtuse parim prognoos etteantud valimi y korral on tinglik keskväärtus, antud juhul seega

$$\hat{y}_{uus} = E(Y_{uus} \mid y). \tag{1.4}$$

Eeldades uuritava tunnuse Y normaaljaotust ning seda, et sellisel juhul on keskväärtuste vektor teada, saame ühisjaotuseks mitmemõõtmelise normaaljaotuse

$$\begin{pmatrix} Y_{uus} \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} X_{uus}\beta \\ X\beta \end{pmatrix}, \begin{pmatrix} V_{uus} & C \\ C^T & V \end{pmatrix}\right),$$

kust avaldame komponendi $f(Y_{uus} \mid Y)$ jaotuse parameetrid järgmiselt:

keskväärtus (1.4) põhjal

$$\hat{Y}_{uus} = X_{uus}\beta + CV^{-1}(Y - X\beta); \tag{1.5}$$

dispersioon

$$Var(Y_{uus}) = V_{uus} + C^T V^{-1} C\,. \tag{1.6}$$

Samas ei ole meil alati kovariatsioonimaatriksid ja β täpselt teada ning me ei saa parimat prognoosi (1.5) iga kord leida.

Vaatleme järgnevalt mudelis (1.3) fikseeritud ja juhuslike faktorite lineaarset kombinatsiooni $Y_{uus} = X_{uus}\beta + Z_{uus}\eta + \varepsilon_{uus}$ ja toome prognoosi[1] $\tilde{Y}_{uus}$ jaoks sisse järgmised lisanõuded:

1. lineaarsus (*Linearity*): $\tilde{Y}_{uus}$ on olemasolevate vaatlustulemuste *y* lineaarne funktsioon, $\tilde{Y}_{uus} = a + BY$, kus *a* on mingi konstantne vektor ja *B* on konstatne maatriks;
2. nihketus (*Unbiasedness*): $E(\tilde{Y}_{uus}) = E(Y_{uus})$;
3. parim (*Best*) selles tähenduses, et $E(\tilde{Y}_{uus} - Y_{uus})^T A(\tilde{Y}_{uus} - Y_{uus})$ on minimaalne (*A* on mistahes positiivselt määratud maatriks).

Toodud tingimusi rahuldavat prognoosi $\tilde{Y}_{uus}$ kutsutakse juhusliku suuruse Y_{uus} parimaks lineaarseks nihketa prognoosiks (*BLUP*).

Tõestatud on (vt näiteks Isotalo jt [6]), et nõuetele 1.–3. vastav juhusliku suuruse $\hat{Y}_{uus}$ mingi realisatsiooni y_{uus} prognoos on kujul

$$\begin{aligned} \hat{y}_{uus} &= X_{uus}\hat{\beta} + CV^{-1}(y - X\hat{\beta}) \\ &= X_{uus}\hat{\beta} + (Z_{uus}GZ^T + K)V^{-1}(y - X\hat{\beta}), \end{aligned} \tag{1.7}$$

kus $\hat{\beta} = (X^T V^{-1} X)^- X^T V^{-1} y$.

Järgnevalt vaatleme mõningaid eriolukordi, et mõista juhuslike muutujate prognooside (1.7) üldkasutatavust.

1. Olgu $X_{uus} = X$ ja $Z_{uus} = 0$, $\varepsilon_{uus} = 0$, siis

[1] Termin 'prognoos' (*Predictor*) eristab juhuslike kordajate ja fikseeritud kordajate hinnangute nimetusi. Mõistet prognoos kasutatakse rõhutamaks, et tegemist on juhusliku suuruse realisatsiooni „ära arvamisega". Kui soovitakse teada fikseeritud mittejuhusliku parameetri tundmatut väärtust, siis seda väärtust hinnatakse, aga kui juhusliku suuruse realisatsiooni, siis prognoositakse. Tegemist ei ole sisulise, vaid traditsioonidest lähtuva eristamisega, vastavat arutelu vt näiteks Robinson [13].

$$BLUP(y_{uus}) = X\hat{\beta} + 0$$
$$= X\hat{\beta}.$$

Saadud prognoos langeb kokku üldise lineaarse fikseeritud kordajatega mudeli *BLUE* hinnanguga.

2. Võttes $Z_{uus} = I$ ning $X_{uus} = 0$, $\varepsilon_{uus} = 0$, saame prognoosi segamudeli juhusliku parameetri γ jaoks, see on sama tulemus, mis paljudes segamudelite kasutamist kirjeldavates artiklites nn Hendersoni võrrandeid kasutades [9,13,14]:

$$BLUP(y_{uus}) = 0 + (IGZ^T + 0)V^{-1}(y - X\hat{\beta})$$
$$= GZ^TV^{-1}(y - X\hat{\beta}).$$

3. Korreleeritud jääkidega mudeli $Y = X\beta + \varepsilon$ korral, kui $Z = Z_{uus} = 0$ saame

$$BLUP(y_{uus}) = X_{uus}\hat{\beta} + KR^{-1}(y - X\hat{\beta}).$$

Nagu punktis 1.1 juba mainitud, ei ole tegelikud β ja kovariatsioonimaatriksid *G* ja *R* üldjuhul teada. Eeldades asümptootilist normaaljaotust saab suurima tõepära meetodil hinnata nii β kui ka kovariatsioonimaatriksid, nagu seda tehakse näiteks edaspidi kasutatava SAS MIXED protseduuri korral [15]. Juhul kui Y_{uus} prognoosimisel kasutatakse tegelike β ja kovariatsioonimaatriksite *G* ja *R* asemel valimi põhjal hinnatud $\hat{\beta}$ ja kovariatsioonimaatrikseid $\hat{G}$ ja $\hat{R}$, nimetatakse hinnanguid vastavalt *EBLUE* ja *EBLUP*, rõhutamaks empiirilist lähenemist (*Empirical BLUP*).

Erilist tähelepanu väärivad segamudelite kasutamisel prognoosipiirid, sest kovariatsioonistruktuuri arvesse võttes peaksid prognoosid olukordades, kus lähedased mõõtmised korreleeruvad, täpsemad tulema (vt ka tulemusi antud töö punkt 3.2.2). Lineaarsete fikseeritud faktoritega mudeli puhul saadakse iga mistahes faktori taseme korral teatav keskmine prognoos, mõõtmistevahelisi kaugusi arvestades (kovariatsioone arvesse võttes) aga on võimalik olemasolevaid mõõtmisi kasutades täpsustada erinevusi selle keskmise suhtes. Kasutades juhusliku suuruse $\hat{Y}_{uus} - Y_{uus}$ asümptootilist normaaljaotust teame, et

$$\frac{\hat{Y}_{uus} - Y_{uus}}{s(\hat{Y}_{uus} - Y_{uus})} \sim t(df),$$

kus *s* on standardhälbe hinnang ja $t(df)$ on *df* vabadusastmega *Studenti* jaotus. Eelnevat arvestades leitakse $(1-\alpha/2)\cdot 100$-protsendilised prognoosipiirid järgmisest tingimusest

$$P(\hat{Y}_{uus} - t_{(\alpha/2,df)} \cdot s(\hat{Y}_{uus} - Y_{uus}) < Y_{uus} < \hat{Y}_{uus} + t_{(\alpha/2,df)} \cdot s(\hat{Y}_{uus} - Y_{uus})) < 1 - \alpha .$$

Standardhälbe prognoos *s* saadakse võttes ruutjuure järgmisest dispersioonist

$$Var(\hat{Y}_{uus} - Y_{uus}) = \hat{V}_{uus} - \hat{C}\hat{V}^{-1}\hat{C}^{T} + (X_{uus} - \hat{C}\hat{V}^{-1}X)(X^{T}\hat{V}^{-1}X)^{-} (X_{uus} - \hat{C}\hat{V}^{-1}X)^{T},$$

vt. näiteks Henderson [5].

Kovariatsiooniparameetrite kohta statistiliste järelduste tegemiseks kasutatakse tõepäral baseeruvaid statistikuid, nagu näiteks *Wald'i Z*, mis arvutatakse kui parameetri hinnangu ja asümptootilise standardvea suhe. Asümptootiline standardviga arvutatakse tõepärafunktsiooni teist järku tuletiste maatriksi pööramisel, arvestades igat kovariatsiooniparameetrit. Suurte valimite korral töötab *Waldi Z* valiidselt, probleeme võib olla väikeste valimimahtude korral. Teine võimalik statistik on tõepärasuhtel baseeruv χ^2 statistik, mida kasutatakse enamasti kovariatsioonimudelite võrdlemisel (vt Lisa 1, [15]).

Mõned märkused probleemidest, mis seostuvad segamudelite kasutamisega. Kuna me tegelikke kovariatsioone ei tea ning prognooside leidmisel kasutatakse valimi põhjal hinnatud kovariatsioonimaatrikseid, mis leitakse vaatluste asümptootilist normaaljaotust eeldades, tekib vajadus kontrollida eelduse paikapidavust. Väga oluline on valimi maht, samuti mudeli (disainimaatriksite *X* ja *Z*) valik. Enne ülesande formuleerimist on vajalik selgitada välja, kas vajatakse uuritavaid faktoreid fikseeritute või juhuslikena [3,13]. Seni on kehtinud arusaam, et olukorras, kus tahetakse teada iga konkreetse faktori taseme mõju, peaks valima fikseeritud faktori. Jälgides näiteks Robinson jt [13] diskussiooni, ei pruugi see alati kehtida.

2. RUUMILISTE ANDMETE MODELLEERIMINE SEGAMUDELITE ABIL

Ruumiliste andmete omapäraks on vaatluste võimalik korreleeritus. Positiivne ruumiline korrelatsioon leiab aset olukorras, kus ruumis üksteisele lähemal paiknevates punktides tehtud mõõtmised on sarnasemad kui kaugemad. Statistilise analüüsi ülesanne on kirjeldada ja modelleerida nii üldist varieeruvust kui ka vaatluste vahelise korrelatiivse sõltuvuse iseloomu.

Klassifitseerides näiteks Banerjee jt. [2] järgi saame ruumiliste andmete jaoks järgmised võimalikud lähenemised.

1. Punktiga määratud (*point-level*) andmed, kus juhuslik vektor $Y(s)$, mille argument s varieerub fikseeritud r-mõõtmelise reaalarvulise alamhulga S piires, $s \in S \subset \Re^r$. Selliseid andmeid nimetatakse geostatistilisteks, nende hulka võib arvata ka magistritöös kasutatud Peipsi järve veeproovide kontsentratsiooni näitajad. Tihti nimetatakse selliseid andmeid ka geostatistilisteks andmeteks.
2. Piirkonnapõhised ehk võreandmed (*areal-level*), kus S on r-mõõtmelise reaalarvude hulga alamhulk, kuid nüüd tükeldatuna teatud arvuks selgelt eraldatavateks piirkondadeks, millede andmed on tavaliselt nende tükkide keskmised, punke eraldi ei ole esitatud.
3. Teatud mustri järgi tekitatud piirkonnad (*point processes*), kusjuures S on ise juhuslik piirkond. Sellisel juhul $Y(s)$ on 1, kui $s \in S$, ja 0 mujal.

Huvitagu meid niisiis juhuslik protsess $\{Y(s) : s \in S \subset \Re^r\}$. Paneme tähele, et aegridade korral $r = 1$ ja juhul $r > 1$ on tegemist "tõeliselt" ruumilise protsessiga. Matemaatikuid huvitab protsessi $Y(s)$ kohta järelduste tegemine ning selle põhjal väärtuste prognoosimine mõõtmata kohtadesse. Samal ajal tehakse prognoosimisel selle protsessi kohta mitmesuguseid eeldusi. Üks oluline ja tihedamini kasutatavaid eeldusi on mõõtmiste korrelatsiooni sõltuvust mõõtmistevahelisest kaugusest.

2.1 Ruumiliste andmete korreleeritus

Toome sisse mõningad geostatistika põhimõisted. Eeldame, et protsess $Y(s)$ omab lõplikku keskväärtust ja dispersiooni iga $s \in S$ korral. Protsess on rangelt statsionaarne (*strictly stationary*), kui suvalise $h \in \Re^r$ ja punktide $\{s_1,...,s_n\}$ valiku korral $(Y(s_1),...,Y(s_n))$ jaotus on sama, mis $(Y(s_1+h),...,Y(s_n+h))$ jaotus.

Protsess on nõrgalt statsionaarne (*weakly stationary*), kui tema keskväärtus igas punktis on konstantne ja $Cov(Y(s),Y(s+h)) = C(h)$ iga $h \in \Re^r$ korral, kui $\{s, s+h\} \in S$. Nõrk statsionaarsus tähendab seda, et kovariatsioon kahe erineva punkti vahel on esitatav ainult nihkest *h* sõltuva kovariatsioonifunktsioonina $C(h)$.

Olemas on veel kolmas statsionaarsuse tüüp, mida nimetatakse sisemiseks statsionaarsuseks (*intrinsic stationarity*). Sel korral eeldatakse, et protsessi keskväärtus igas punktis $s \in S$ on konstantne, $E(Y(s+h)-Y(s)) = 0$. Juhul, kui suurus

$$2\gamma(h) := Var(Y(s+h)-Y(s)) \tag{2.1}$$

sõltub ainult nihkest *h*, nimetatakse protsessi sisemiselt statsionaarseks. Kui protsess on rangelt või nõrgalt statsionaarne, siis on tal ka sisemise statsionaarsuse omadus. Funktsiooni $2\gamma(h)$ nimetatakse **variogrammiks** ning funktsiooni $\gamma(h)$ **semivario-grammiks.**

Semivariogrammi ja kovariatsioonifunktsiooni vahel kehtib järgmine seos

$$\gamma(h) = C(0) - C(h),$$

sest

$$\begin{aligned} 2\gamma(h) &= Var(Y(s+h) - (Y(s)) \\ &= Var(Y(s+h)) + Var(Y(s)) - 2Cov(Y(s+h, Y(s)) \\ &= C(0) + C(0) - 2C(h) \\ &= 2(C(0) - C(h)). \end{aligned}$$

Käesolevas töös vaadeldakse ka juhtu, mil meid huvitav tunnus *Y* on mõõdetud ebatäpselt, st *i*. ja *j*. vaatlus, tehtud vastavalt punktides s_i ja s_j, ei pruugi olla (täpselt) võrdsed ka siis, kui on tehtud samas kohas (s_i=s_j). Sellisel juhul tuleb lisaks kovariatsioonifunktsioonilevaadelda ka vaatluste dispersiooni, $Var(Y(s))$. Mõõtmisvigade olemasolu korral $Var(Y(s)) \neq C(0)$ ja semivariogrammi ja kovariatsioonifunktsiooni vaheline seos teiseneb kujule: $\gamma(h) = (Var(Y) - C(h))$.

Juhul kui kauguse arvestamisel pole oluline suund, on tegemist isotroopse protsessiga. Antud magistritöös piirdumegi vaid isotroopsete protsessidega. Seega saame (semi)variogrammi kirjeldada ka kui funktsiooni, mis sõltub ainult vaatlustevahelisest kaugusest. Peipsi järve näitel peaksid lähemalasuvate punktide veeproovidest saadud mõõtmistulemused olema sarnasemad, kui üksteisest kaugemal asuvates punktides saadud mõõtmistulemused. Tähistame d_{ij} *i*-nda ja *j*-nda proovipunkti koordinaatide vahelise kauguse, mis avaldub järgmiselt:

$$d_{ij} = \sqrt{(pl_i - pl_j)^2 + (ip_i - ip_j)^2}$$

Iseloomustame järgnevalt ruumiliste andmete modelleerimisel kasutatavaid kovariatsiooniparameetreid ning toome sisse vajalikud tähistused:

1) Tunnuse tegeliku väärtuse hajuvuse või dispersiooni tähistame σ^2.
2) Seriaalkorrelatsioon, mis näitab vaatlustevahelise korrelatsiooni sõltuvust mõõtmispunktide vahelisest kaugusest. Seda korrelatsiooni iseloomustatakse funktsiooni $\rho(\theta, d_{ij})$ abil. Korrelatsioonifunktsiooni hinnatav parameeter θ võimaldab välja valida algandmetega kõige paremini sobiva mudeli, seda nimetatakse ka korrelatsiooni- või kovariatsiooniparameetriks. Korrelatsiooni-funktsioon peaks olema selline, et $\rho(\theta, 0) = 1$, ning mida kaugemal vaatlused asuvad, seda nõrgemaks korrelatsioon muutub. Suurust $\phi = 1/\theta$ nimetatakse skaalaparameetriks. Skaalaparameeter iseloomustab paljudel juhtudel funktsiooni kuju (lamedust).
3) Mõõtmisveast tulenev varieeruvus, mis näitab, et mõõtmisprotsess ise on hajuvuse allikaks, tähistame τ^2.

Kasutusel on mitmeid standardseid kovariatsioonistruktuure [2,15], antud töös vaadeldakse neist kolme, millede lähemad kirjeldused on esitatud punktis (2.3). Põhimõtteliselt on võimalik ka ise uusi ruumiliste andmete analüüsiks sobivaid korrelatsioonistruktuure konstrueerida.

Paljudel juhtudel on ruumilised vaatlused tehtud gruppidena, kusjuures eri gruppidesse kuuluvad vaatlused on omavahel sõltumatud. Eeldatakse, et kõikides vaatluste gruppides (näiteks erinevad järved või sama järve andmed erinevatel aegadel) on ruumiline kovariatsioonistruktuur sama. Parameetrite prognoosimisel on võimalik sel juhul valimi kovariatsioonimaatriks esitada teatava plokkmaatriksina, iga faktori taseme ruumilised

vaatlused eri plokkides. Plokisisesed vaatlused on korreleeritud, kuid erinevate plokkide omad mitte. Segamudeli hindamine sellise plokkstruktuuri korral on tunduvalt kiirem (pöördmaatriksite leidmine toimub plokkide kaupa). Järgmises punktis esitame näite, kuidas selline plokkidesse jagamine prognoose mõjutab. Etterutates mainime, et antud magistritöö algandmete puhul on faktoriks aeg, mille tasemeteks on ekspeditsioonid.

2.2 Empiiriline ja teoreetiline variogramm

Ruumiliste andmete modelleerimisel on väga tähtis õige mudeli valik. Kontrollimaks, kuidas üks või teine korrelatsioonistruktuuriga mudel algandmetega sobib, kasutatakse suurima tõepära meetodil hinnatud semivariogrammi graafilist võrdlemist algandmete põhjal hinnatud empiirilise või robustse semivariogrammiga. Viimased arvutatakse järgmiselt:

empiiriline semivariogramm

$$\tilde{\gamma}(h) = \frac{1}{2\,|N(h)|}\sum_{N(h)} |Y(s_i) - Y(s_j)|^2,$$

kus $|N(h)|$ on selliste paaride arv, millede vaheline kaugus on *h*, kusjuures $h \in \left(d_{ij} - \delta, d_{ij} + \delta\right)$, see tähendab valitakse välja mingi vahemik, kuhu kuuluvad kaugused loetakse samadeks;

robustne semivariogramm

$$\tilde{\gamma}(h) = \frac{\Psi^4(h)}{2(0{,}457 + 0{,}494/|N(h)|)},$$

kus

$$\Psi(h) = \frac{1}{|N(h)|}\sum_{N(h)} |Y(s_i) - Y(s_j)|^{1/2}.$$

Empiiriline ja robustne semivariogramm on tegeliku semivariogrammi kaks erinevat hinnangut, mis ei tee eeldusi vaatlustevahelise sõltuvuse funktsiooni kuju kohta..

Uurimaks semivariogrammi hinnangute käitumist ning mudeli sobitamist variogrammide võrdlemise abil tehti magistritöö koostaja poolt simuleerimiskatse. Arvestades töö kirjutamisel kasutatud andmestiku struktuuri ja hinnangute esialgseid ligikaudseid väärtusi, võeti järgmised valimid:

1) väiksema valimimahuga *N*=3*56 andmestik, kuhu võeti Peipsi andmestikust viiekümne kuue erineva punkti koordinaadid, nende korral kolm sama eksponentsiaalse kovariatsioonistruktuuriga uuritava tunnuse väärtuste plokki (ütleme, et kolm erinevat ekspeditsiooni), kõik standardsest normaaljaotusest;

2) suurema valimimahuga andmestik *N*=2*200, mille koordinaadid simuleeriti standardsest normaaljaotusest, kusjuures plokke võeti 2. Eksponentsiaalse kovariatsioonistruktuuri parameetrid anti ette järgnevad (vt tähistused lk 13): $\sigma^2 = 0{,}5$; $\theta = 1$; $\tau^2 = 0$;

SAS protseduuri VARIOGRAM abil (vt Lisa 1 ja [16]) telliti uuritava tunnuse empiiriline semivariogramm kaht erinevat andmestruktuuri eeldades:

1. võttes kõik andmed sõltuvatena, sellisel juhul erinevate plokkide samade punktide väärtuste erinevus tuleneks mõõtmisveast;
2. võttes plokid omavahel sõltumatutena, eeldades, et plokisiseselt on ruumiline kovariatsioonistruktuur sama.

Seejärel leiti SAS MIXED protseduuriga simuleeritud andmete jaoks kovariatsiooniparameetrite hinnangud $\hat{\sigma}^2$, $\hat{\theta}$, $\hat{\tau}^2$. Siinkohal märgime, et tegelikult olid need andmete simuleerimisel ette antud, seega katsega on võimalik ühtlasi kontrollida seda, kui täpsed MIXED protseduuri suurima tõepära hinnangud tulevad.

Tulemused on esitatud järgnevas tabelis.

Tabel 1. PROC MIXED abil saadud parameetrite hinnangud erinevate valimite korral, kus tegelikud väärtused on $\sigma^2 = 0{,}5$; $\theta = 1$; $\tau^2 = 0$.

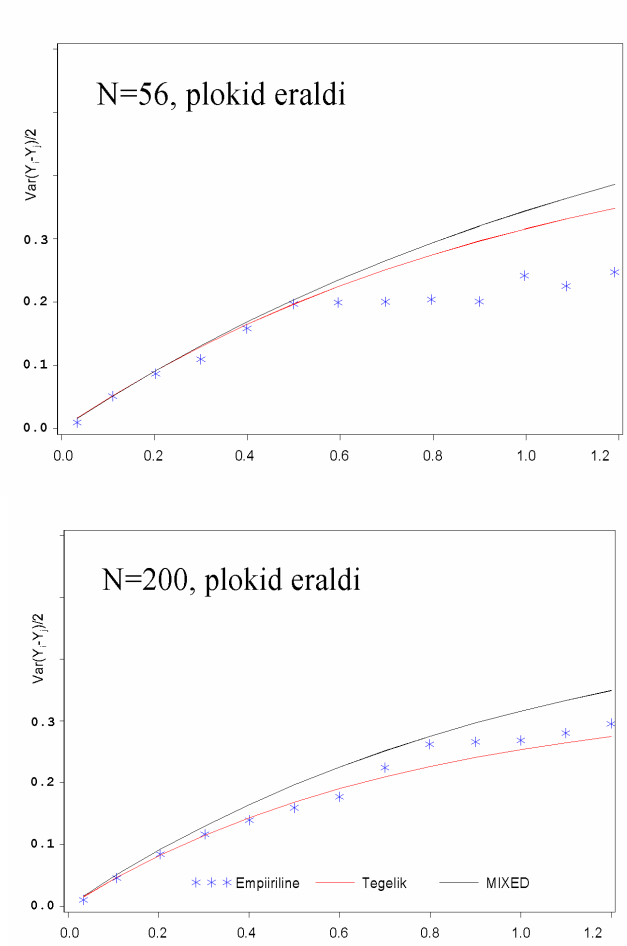
N	SUBJ=	$\hat{\sigma}^2$	$\hat{\theta}$	$\hat{\tau}^2$
3*56=168	INTERCEPT	0,017	0,626	0,31
3*56=168	PLOKK	0,669	1,403	0
2*200=400	INTERCEPT	0,26	1,307	0,15
2*200=400	PLOKK	0,34	0,75	0

Tabelist näeme, et sama struktuuriga ühendatud plokkide hindamine ühe, kõikide mõõtmiste omavahelist sõltuvust eeldava andmestikuna ei anna õigeid tulemusi, sest samas kohas tehtud mõõtmiste erinevus kantakse tunnuse üldise hajuvuse arvelt mõõtmisvea arvele. Kõige täpsemad prognoosid saadi suurema valimimahuga plokkide korral, seega valimimaht on väga oluline. Kasutades suurima tõepära meetodil prognoositud

parameetrite väärtusi arvutati suurima tõepära hinnangute abil variogramm, eksponentsiaalse kovariatsioonistruktuuri korral näiteks valemi (2.2) põhjal:

$$2\hat{\gamma}(d_{ij}) = (\hat{\sigma}^2 + \hat{\tau}^2) - \hat{\sigma}^2 \cdot \exp\left(-\frac{d_{ij}}{\hat{\theta}}\right).$$

Semivariogramm $\hat{\gamma}(d_{ij})$, algandmete põhjal koostatud empiiriline semivariogramm ja teoreetiline semivariogramm paigutati samale joonisele. Tulemusi illustreerib Joonis 1. Sellelt on näha, et piisavalt suure valimi korral langevad semivariogrammid päris hästi kokku ning plokk-kovariatsioonistruktuuri hindamine annab täpsemad tulemused. Seda võis järeldada ka sellest, et hinnanguid leides 3*56-se valimi korral langesid protseduuriga MIXED saadud hinnangud etteantutega kokku pooltel juhtudel, 2*200-se korral peaaegu kõigil.



Joonis 1. Empiiriline, tegelik ja suurima tõepära meetodil hinnatud semivariogramm.

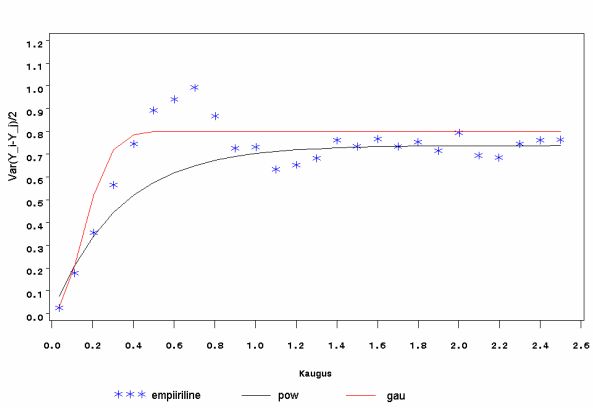
Teine katse tehti vastupidiselt eelmisele järgmiselt: simuleeriti üks suur etteantud kovariatsioonistruktuuriga andmestik ja kontrolliti, kas selle hindamine plokkidena annab õiged parameetrite hinnangud. Hinnates etteantud kovariatsioonistruktuuriga andmestikku (N=1000) jagatuna erinevaks arvuks plokkideks saadi tulemused, mis on esitatud tabelis 2.

Tabel 2. Suurima tõepära hinnangud kovariatsioonistruktuuri parameetritele erinevalt moodustatud andmete korral, kui tegelikud väärtused on $\sigma^2 = 0{,}5$; $\theta = 1$; $\tau^2 = 0$.

Plokke	σ^2	θ	τ^2
1	0,49	1,04	0
2	0,45	0,94	0
20	0,39	0,75	0
50	0,38	0,74	0

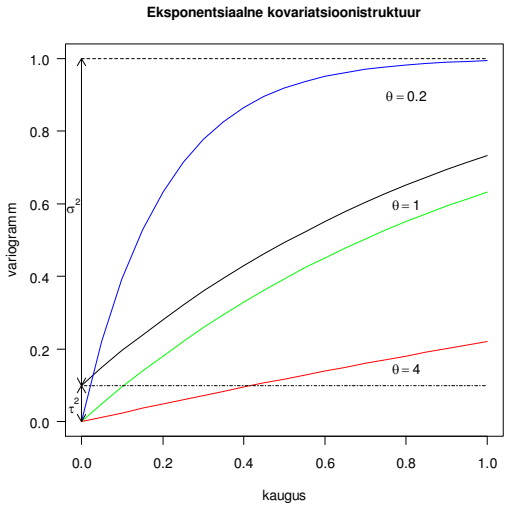
Tabelist on näha, et hinnangud tulevad ka plokkstruktuuri kasutades (antud juhul vale kovariatsioonistruktuur) õigetele väärtustele lähedased. Katsed erinevate valimimahtudega näitasid, et erinevus tegelikest parameetritest oli vahemikus 0,05…0,4, kusjuures arvuti tööaeg plokkstruktuuriga andmestikku hinnates oli tunduvalt lühem.

Joonis 2 illustreerib semivariogrammide hinnanguid, mis kasutavad õiget ja vale kovariatsioonistruktuuriga mudelit. Esitatud on kolm erinevat semivariogrammi hinnangut: Gaussi kovariatsioonistruktuuriga (vt valem 2.4) andmete empiiriline semivariogramm; Gaussi mudelit kasutades suurima tõepära meetodil hinnatud semivariogramm; astendajaga kovariatsioonistruktuuri (vt valem 2.3) kasutava mudeli põhjal suurima tõepära meetodil hinnatud semivariogramm (hindamisel kasutati vale kovariatsioonistruktuuri). Jooniselt on näha, et viimane erineb empiirilisest oluliselt väikeste kauguste korral.

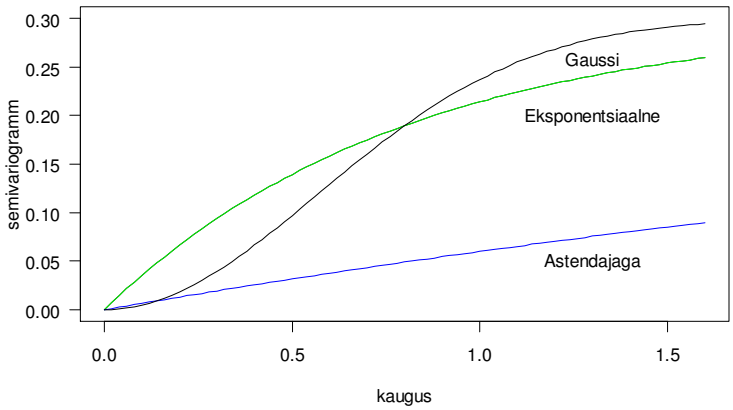


Joonis 2. Mudeli sobitamine variogrammi abil Gaussi kovariatsioonistruktuuriga andmete näitel.

Vaatleme nüüd eksponentsiaalse kovariatsioonistruktuuri teoreetilise variogrammi näiteid parameetrite erinevate väärtuste korral. Nimelt saab variogrammi põhjal teha järeldusi, kas vaatlused on korreleeritud ning milliste vahemaade tagant tehtud mõõtmised on sisuliselt sõltumatud. Variogrammi kuju määrab skaalaparameetri ($1/\theta$) suurus, alguspunkti määrab mõõtmisvea hajuvus ning maksimaalse väärtuse mõõtmisvea ja tunnuse hajuvuste summa. Joonis 3 on $\theta = 1$ korral hinnatud kaks variogrammi: roheline mõõtmisveata ning must mõõtmisveaga, mistõttu viimane paikneb lihtsalt kõrgemal. Punase joone järgi on näha, et väikese skaalaparameetri väärtuse korral korreleeruvad mõõtmised isegi maksimaalsel kaugusel. Vähimat kaugust, mille juures variogramm stabiliseerub (joonisel esitatud sinise joone korral umbes 0,7), nimetatakse mõjupiiriks (*range*).



Joonis 3. Variogrammid eksponentsiaalse kovariatsioonistruktuuri korral, skaalaparameetrite erinevate väärtustega. Kaugus on juhuslik valim standardsest normaaljaotused.



Joonis 4. Semivariogrammide võrdlus erinevate kovariatsioonistruktuuride korral (antud parameetrite väärtustel $\theta = 0{,}8$; $\sigma^2 = 0{,}3$; $\tau^2 = 0$).

2.2 Kovariatsioonistruktuurid ruumiliste andmete jaoks

Järgnevalt esitame magistritöös kasutatud erinevate ruumiliste struktuuride semivariogrammide arvutuseeskirjad ja kovariatsioonifunktsioonid, kasutades kõikide struktuuride korral punktis 2.1 (lk 13) toodud tähistusi.

1) Eksponentsiaalse kovariatsioonistruktuuri mudel

$$\gamma(d_{ij}) = \begin{cases} \tau^2 & , d_{ij} = 0 \\ \tau^2 + \sigma^2 - \sigma^2 \cdot \exp\left(-\frac{d_{ij}}{\theta}\right) & , d_{ij} \neq 0 \end{cases}, \tag{2.2}$$

$$C(d_{ij}) = \begin{cases} \sigma^2 & , \quad d_{ij} = 0 \\ \sigma^2 \cdot \exp\left(-\frac{d_{ij}}{\theta}\right) & , \quad d_{ij} \neq 0 \end{cases}.$$

2) Astendajaga kovariatsioonistruktuuri mudel, mis on esimest järku autoregressiivse kovariatsioonistruktuuri üldistus selles mõttes, et kui kaugused oleks naturaalarvud, siis oleks tegemist 1.-järku autoregressiivse (AR1) protsessi poolt tekitatud korrelatsioonistruktuuriga:

$$\gamma(d_{ij}) = \begin{cases} \tau^2 & , d_{ij} = 0 \\ \tau^2 + \sigma^2 - \sigma^2 \theta^{d_{ij}} & , d_{ij} \neq 0 \end{cases}, \tag{2.3}$$

$$C(d_{ij}) = \begin{cases} \sigma^2 & , d_{ij} = 0 \\ \sigma^2 \theta^{d_{ij}} & , d_{ij} \neq 0 \end{cases}.$$

Astendajaga mudel reparametriseerib eksponentsiaalse mudeli (2.2).

3) Gaussi kovariatsioonistruktuuriga mudel, mis erineb eksponentsiaalsest vaid selle poolest, et kaugus ja korrelatsiooniparameeter on võetud ruutu:

$$\gamma(d_{ij}) = \begin{cases} \tau^2 & , d_{ij} = 0 \\ \tau^2 + \sigma^2 - \sigma^2 \cdot \exp\left(-\frac{{d_{ij}}^2}{\theta^2}\right) & , d_{ij} \neq 0 \end{cases}, \tag{2.4}$$

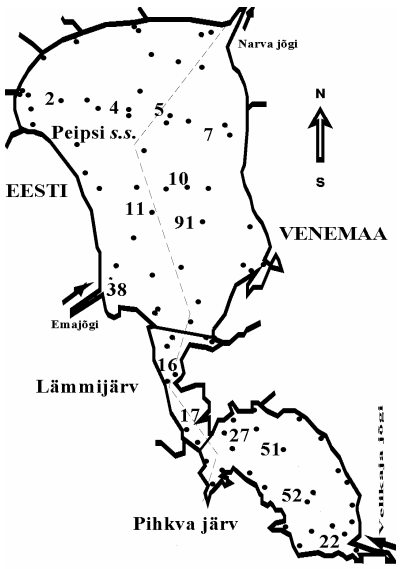
$$C(d_{ij}) = \begin{cases} \sigma^2 & , d_{ij} = 0 \\ \sigma^2 \cdot \exp\left(-\frac{{d_{ij}}^2}{\theta^2}\right) & , d_{ij} \neq 0 \end{cases}.$$

Vaadeldes Joonisel 4 olevaid semivariogramme, märkame, et Gaussi mudel kirjeldab teistest paremini olukorda, kus nullist alates kauguse suurenedes korrelatsioon vaatluste vahel püsib pikemat aega tugev ning siis järsult väheneb.

3. PEIPSI JÄRVE MUDELID

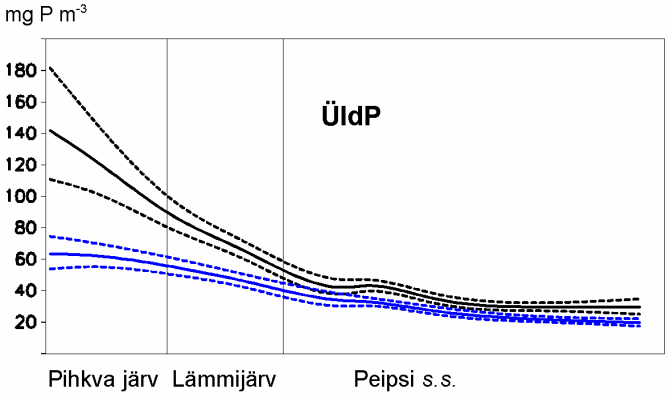
3.1 Ülevaade varasematest uuringutest

Peipsi järve vee andmete statistilise modelleerimisega on tegeldud regulaarselt 1989. aastast alates. Seni on kõik uuritavad mudelid olnud üldised lineaarsed mudelid, mis võimaldavad hinnata vee koostise näitajate erinevaid ajalisi ja ruumilisi keskmisi ning nende muutumise suunda ja kiirust. Kasutusel on olnud kuni 151 hinnatava parameetriga mudeleid, kus on arvestatud proovivõtu kohta, aega ja sügavust ning nende erinevaid koosmõjusid. Aja näitajatena on vaadeldud aastaid ja proovivõtu päevi, viimaseid on on teisendanud kui "mitmes päev aastas". Päeva number iseloomustab aastasisest sesoonsust, mis on suure tähtsusega faktor hüdrokeemiliste ja -bioloogiliste vaatluste korral. Proovi võtmise koht kui faktor on olnud vaatluse all mitut moodi. Algul diskreetsena kui järveosa (eraldi Peipsi suurjärv (lühendiga *s.s – sensu stricto*), Lämmijärv, Pihkva järv ning mõne ülesande korral ka Emajõe mõjuala). Veidi hiljem jagati järv teatavatesse asukoha suhtes sarnastesse "aspektidesse" Möls jt. [12]. Viimasel ajal on punkti asukoha koordinaadid mudelis sees lineaarselt, pidevate suurustena. Sellisel viisil saadud mudelit on kasutatud Peipsi sünteetilise andmebaasi loomisel Möls jt. [10] ning erinevate artiklite Haldna jt [4], Milius jt [8] koostamisel. Järgmises alapunktis 3.1.1. on toodud mudeli parameetrid ja nende kirjeldused ning edaspidi nimetatakse seda lühidalt Mudel0.

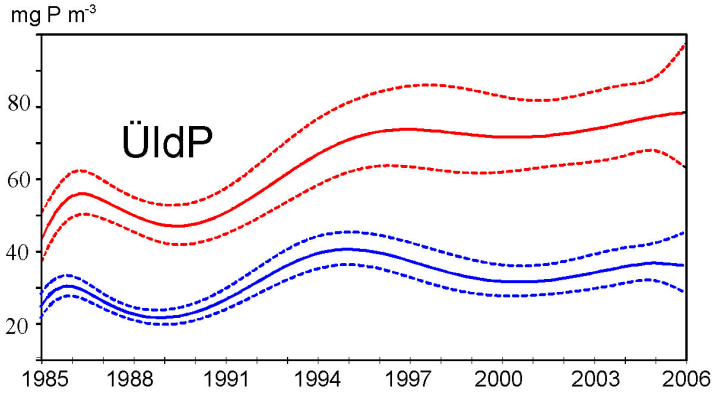


Joonis 5. Järve kaart koos Peipsi seire (1996-2006) proovipunktidega. Järve pikkus põhjast lõunasse on maksimaalselt 152 km.

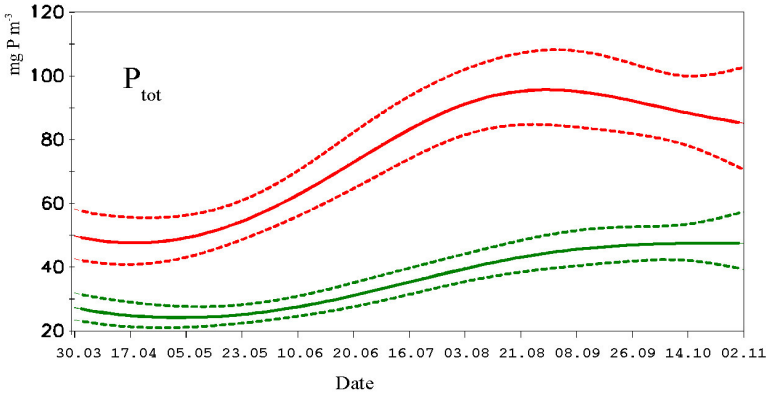
Järgnevalt mõned Mudel0 abil saadud tulemused.



Joonis 6. Üldfosfori kontsentratsiooni muutumine piki järve 19. juuli seisuga 1987. a (sinine) ja 2006. aastal, prognoosid koos keskväärtuse 95% usalduspiiridega.



Joonis 7. Üldfosfori kontsentratsiooni pikaajaline muutumine Lämmijärves (punane) ja Peipsis *s.s* 19. juuli seisuga, prognoosid koos keskväärtuse 95% usalduspiiridega.



Joonis 8. Üldfosfori sesoonne muutumine Lämmijärves (punane) ja Peipsis *s.s.* 2003.a. seisuga, prognoosid koos keskväärtuse 95% usalduspiiridega.

3.1.1 Üldised lineaarsed mudelid

Eelmises punktis toodud Jooniste 6 – 8 tegemisel on kasutatud üldist lineaarset mudelit Mudel0. Joonis 6 ja Joonis 7 on tehtud toimetamisel oleva Peipsi järve monograafia jaoks, Joonis 8 on võetud ajakirjale Hydrobiologia saadetud artiklist [4]. Toome siinkohal mudeli parameetrid ja nende kirjeldused.

Mudel0 parameetrid:

```
intercept
a1 a2 a3 a4 a5 a6                 – aasta teisendused
pl ip pl*pl pl*pl*pl              – teisendatud koordinaadid
pl*ip pl*pl*ip
t35 t44 t53                       – sesoonsuse teisendused
a1*pl a1*pl*pl a4*pl a4*pl*pl     – interaktsioonid
a2*pl a2*pl*pl a5*pl a5*pl*pl
a3*pl a3*pl*pl a6*pl a6*pl*pl
a1*ip a1*t35 a1*t44 a1*t53
a2*ip a2*t35 a2*t44 a2*t53
a3*ip a3*t35 a3*t44 a3*t53                                  (Mudel0)
a4*ip a4*t35 a4*t44 a4*t53
a5*ip a5*t35 a5*t44 a5*t53
a6*ip a6*t35 a6*t44 a6*t53
pl*t35 pl*t44  pl*t53
pl*pl*t35 pl*pl*t44 pl*pl*t53
ip*t35 ip*t44 ip*t53.
```

Mudeli parameetrite kirjeldused: `a1,a2,…,a6` on teatavad normaaljaotuse tihedusel põhinevad aastateisendid, mille arvutusvalemid on järgmised:

$a_i = \exp\left(-\frac{(a-\mu_i)^2}{2\sigma^2}\right)$, i=1,…6, kus μ_i on kasutaja poolt valitud teatavad "keskmised aastad", et hõlmata kogu ajavahemik, millal andmed kogutud on. Arvutuste stabiilsuse tagamiseks astmeteisenduste korral võetakse (vt. Möls, T. [11]) $a = (\text{aasta} - 1920)/10$. Universaalseimad (sobivad enamiku ainete kontsentratsioonide jaoks) ja parimaid tulemusi andnud väärtused on μ =(3; 4,5; 6; 7; 8; 8,4) ja σ^2=1,4. Sesoonse aja teisendused on välja töötatud beetafunktsioonidena, aluseks algne päeva teisendus *t* = päeva number aastas/365, parimateks ja piisavaiks osutusid *t35, t44, t53*, mis arvutatakse valemiga $tij = t^i(1-t)^j$ (vt Aruvee [1]).

Mudel0 on kasutatav paljude Peipsi järve vee keemiliste ja bioloogiliste näitajate jaoks. Ta on kohandatav ka teiste madalate suurjärvede vee näitajate uurimiseks, mille näiteks võib tuua T. Mölsi poolt tehtud EMÜ Limnoloogiakeskuse ja Hollandi teadlaste andmete (Ijsselmeeri järv) võrdlus. Antud mudeliga on väga hea hinnata faktorite mõjude suurust ja suunda erinevate spetsiaalsete parameeterfunktsioonide abil, samuti huvipakkuvaid keskmisi, kontsentratsioonide ajalisi ja ruumilisi muutusi jms. Parameeterfunktsioonide kasutamine sellise suure arvu parameetrite korral nõuab muidugi erilist tähelepanelikkust ja täpsust programmi tellimuse vormistamisel (SAS paketis `ESTIMATE` lause kirjutamine), samuti on äärmiselt tähtis, et mudel oleks võimalikult "õige", vastasel korral eksisteerib oht teha eksitavaid järeldusi.

Võrdluseks koostatud lihtsama mudeli koostas magistritöö autor. Lihtsamaks nimetame seda parameetrite arvu tõttu – väljatöötamisel arvestati, et esindatud oleksid vaid olulised fikseeritud parameetrid ja võimalikult vähe interaktsioone.

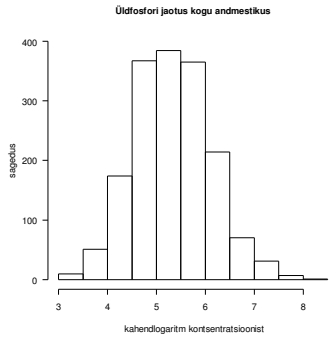
Lihtsa mudeli Lmud parameetrid on järgmised:

```
a a*a a*a*a a*a*a*a
pl*pl pl*pl*pl                                    (Lmud)
t44 t53
a*pl a*a*pl a*pl*pl .
```

Aastasõltuvus võeti mudelisse vaid astmelisena (ei vaja lisaparameetreid), sesoonsuse näitajad *t44* ja *t53* ning põhjalaius jäeti samad kui Mudel0 korral. Edaspidi kasutatakse lühendeid Mudel0 ja Lmud antud töö tabelites ja tekstis pikemalt kirjeldamata.

3.1.2 Üldfosfori andmete kirjeldus

Üldfosforit on mõõdetud alates 1985. a. (va 1991) kuni tänapäevani. Proovipunktide arv Peipsil on olnud aastati kõikuv, üldiselt on püütud haarata kohad, mis järve võimalikult täpselt iseloomustaksid. Andmestik koosneb 137 ekspeditsiooni (nimetatakse edaspidi eestipäraselt „reid") tulemustest, kokku on mõõdetud fosforit vee pinnakihis 1549 korral. Ülevaate lihtsustamise huvides jagatakse järv järgnevate jooniste esitamisel neljaks osaks: Peipsi suurjärve Eesti-pool ja Venemaa-pool, Lämmijärv ja Pihkva järv (Joonis 5). Üldfosfori kontsentratsioon on fikseeritud mõõtmisel milligrammides kuupmeetri kohta. Normaaljaotusele lähendamise eesmärgil on kontsentratsioonid kogu statistilise analüüsi käigus kahendlogaritmitud.

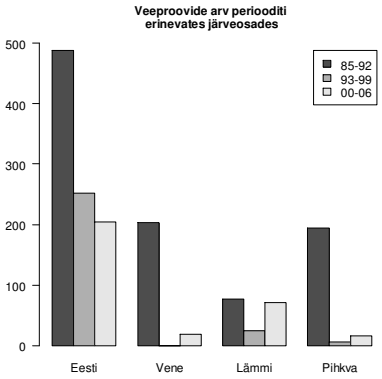


Joonis 9. Peipsi üldfosfori kontsentratsioonide kahendlogaritmi sagedusjaotus.

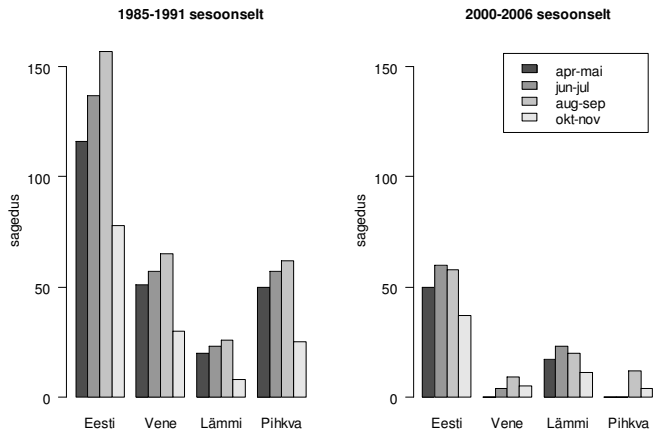
Veeproovide arvu järgi võib eraldada kolm ajalist perioodi (Joonis 10):

- 1985-1992 (joonisel lühendatult 85-92) iseloomustab tihedam proovipunktide arv, kusjuures Venemaa-poolsed ning Pihkva järve punktid on esindatud proportsionaalselt Eesti-poolsetega;
- 1993-1999, mil Peipsi ega Pihkva järve Venemaa-poolsel osal ei käidud kordagi, Pihkva järve esindab vaid mõni Värska lahe mõõtmine;
- 2000-2006, kui hakati koostööprojektide raames uuesti nii Pihkva järve, kui ka Peipsi suurjärve Venemaa-poolset osa uurima.

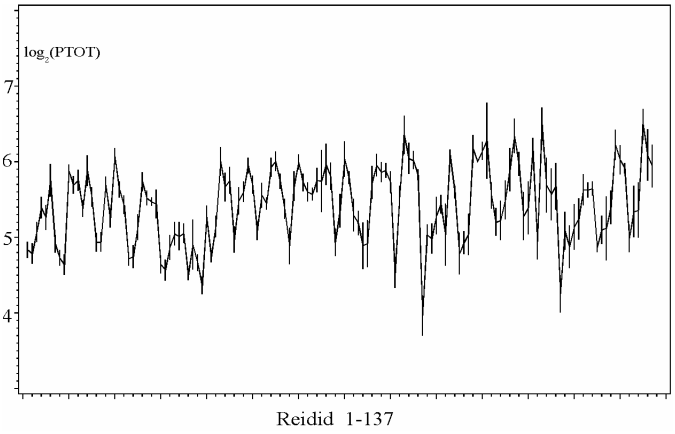
Vee keemiliste ja bioloogiliste näitajate korral varieerub tunnuse väärtus sesoonselt, seetõttu on töös toodud ka joonised esimese ja viimase ajalise perioodi proovide sageduste kohta kahe kuu kaupa jäävabal perioodil (enamasti aprill-november). Tulpade kõrgust mõjutab küll ka proovipunktide arv erinevates järveosades (Joonis 5), kuid kokkuvõttes on siiski näha, et viimastel aastatel on Vene-poolsetes punktides käidud tunduvalt vähem, kui varasematel aastatel.



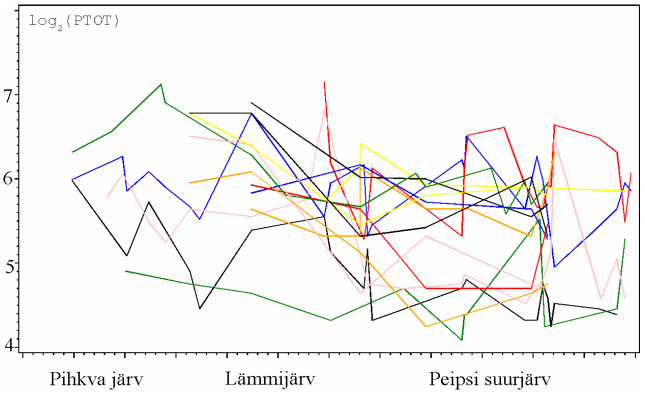
Joonis 10. Fosfori veeproovide arv Peipsi järve osadest erinevatel perioodidel.



Joonis 11. Üldfosfori veeproovide arv Peipsi järve veest kahel erineval perioodil ernevates järveosades Eesti-poole suurjärve osa, Venemaa-poolne suurjärve osa, Lämmijärv ning Pihkva järv (joonisel lühendatult).



Joonis 12. Üldfosfori (PTOT) kontsentratsiooni muutumine andmete kogumise aja jooksul, reidi keskmine koos kahekordse standardveaga.



Joonis 13. Üldfosfori kontsentratsiooni muutumine reidide jooksul, erivärviliselt seitsme erineva reidi mõõtmised 1987. aasta maist kuni novembrini.

3.2 Kovariatsioonistruktuuridega mudelid

Punktis 3.1.1 kirjeldatud suure arvu parameetritega Mudel0 sobib kahtlemata hästi keskmise fosforitaseme prognoosimiseks mistahes argumenttunnuste väärtuste korral. Samas võimaldaksid olemasolevad vaatlused (naabruses asuvad punktid) täpsustada, kuidas üks või teine vaatlus keskmisest erineda võiks. Sellise ühtse veekogu puhul mingi käitumise sarnasus kindlasti eksisteerib. Oletame, et järvel uurimisreidil viibides võetakse järjestikustest punktidest veeproove. Korrelatsioonid punktide vahel tekivad sel juhul kahel põhjusel:

1) kaks mõõtmist on omavahel korreleeruvad, kuna nad on tehtud samal reidil, samades ilmastiku jne tingimustes;
2) asukoha poolest lähemal olevad mõõtmised korreleeruvad rohkem kui kaugemal asuvad. Seda võiks nimetada reidisiseseks punktidevaheliseks kovariatsiooniks.

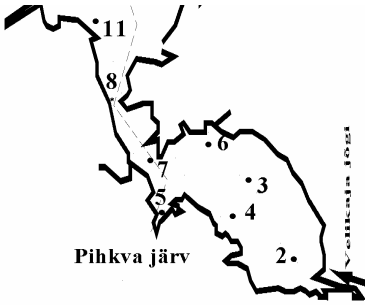
Järgnevalt toome ühe osa valimi põhjal näite, kuidas üldfosfori vaatluste-vahelised Pearsoni korrelatsioonid kaugustega seotud on. Esindatud on ETA Zooloogia ja Botaanika Instituudis aastatel 1985-1990 kogutud andmed kaheksa proovipunkti kohta (paiknemist vt Jooniselt 14). Tulemused on esitatud järgmises tabelis.

Tabel 3. Proovipunktide vahelised kaugused (vasakpoolne alumine kolmnurk) ja vastavate punktide üldfosfori sisalduste omavahelised korrelatsioonid Mudel0 jääkide põhjal (ülemine kolmnurk).

	p2	**p4**	**p5**	**p3**	**p7**	**p6**	**p8**	**p11**
p2		**0.507**	**0.206**	**0.719**	**0.412**	**0.414**	**0.466**	**0.337**
p4	0.595		**0.327**	**0.416**	**0.282**	**0.169**	**0.494**	**0.594**
p5	1.015	0.243		**0.091**	**0.402**	**0.631**	**0.381**	**0.482**
p3	0.356	0.268	0.687		**0.624**	**0.528**	**0.595**	**0.289**
p7	1.087	0.493	0.134	0.743		**0.611**	**0.521**	**0.374**
p6	0.654	0.152	0.430	0.297	0.485		**0.442**	**0.347**
p8	1.242	0.655	0.304	0.891	0.177	0.596		**0.528**
p11	1.387	0.811	0.481	1.031	0.352	0.733	0.176	

Paljudel juhtudel on näha, et kauguse suurenedes korrelatsioon punktide vahel väheneb. Vajadus mingi ruumilise korrelatsioonistruktuuri kasutamise järele on seega olemas.

Vastavalt teises peatükis toodud ruumiliste andmete kovariatsioonistruktuuride kirjeldustele uuriti magistritöö käigus kõiki kolme punktis 2.3. esitatud kovariatsioonistruktuure.



Joonis 14. Tabelites 3 ja 4 toodud proovipunktid.

Tabelis 4 näeme suurima tõepära meetodil hinnatud korrelatsioone. Hindamisel on kasutatud kõiki olemasolevaid vaatlusi.

Tabel 4. SAS MIXED protseduuriga hinnatud üldfosfori korrelatsioonid Mudel0 eksponentsiaalse kovariatsioonistruktuuri korral 1. reidi kaheksa proovipunkti jaoks koos vastavate kaugustega (vasakpoolne alumine kolmnurk).

Estimated R Correlation Matrix for CLUSTER 1

Row	p2	p4	p5	p3	p7	p6	p8	p11
p2		**0.356**	**0.283**	**0.406**	**0.272**	**0.345**	**0.250**	**0.231**
p4	0.595		**0.391**	**0.426**	**0.376**	**0.453**	**0.345**	**0.317**
p5	1.015	0.243		**0.338**	**0.458**	**0.380**	**0.417**	**0.379**
p3	0.356	0.268	0.687		**0.328**	**0.419**	**0.303**	**0.281**
p7	1.087	0.493	0.134	0.743		**0.384**	**0.447**	**0.407**
p6	0.654	0.152	0.430	0.297	0.485		**0.356**	**0.330**
p8	1.242	0.655	0.304	0.891	0.177	0.596		**0.447**
p11	1.387	0.811	0.481	1.031	0.352	0.733	0.176	

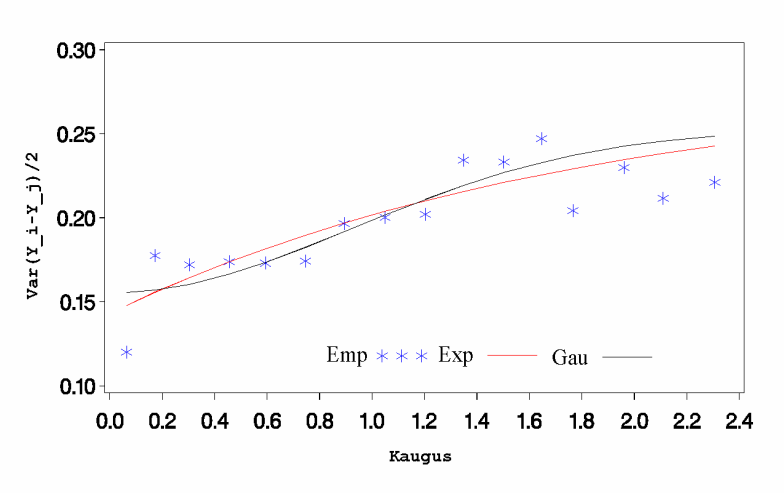
Samas võib võtta veel üldisemalt, et kõik järjestikused proovid, alates 1985. aastast kuni tänapäevani, on omavahel korreleeritud. Sellisel juhul peaksime samaaegselt modelleerima nii ruumilist kui ajalist korrelatsiooni vaatluste vahel. Paar katset modelleerida ka ajalist struktuuri vaatluste vahel ei andnud paremat tulemust, kui vaatlusalused mudelid, mis eeldavad, et eri reididel tehtud vaatlused on sõltumatud. Eeldades kõikide vaatluste vahelist sõltuvust tekkisid probleemid ka arvutamise ajaga, sest ühe hindamise jaoks kulus arvutil enam kui tund aega.

3.2.1 Mudeli valik

Magistritöö eesmärgid tulemuste praktilise kasutamise seisukohast on järgmised:

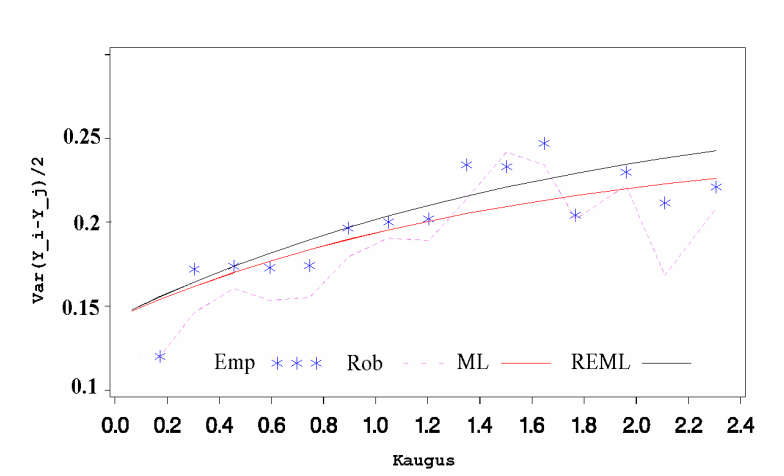
1. uurida, kas varemkasutatud Peipsi mudeleid annab parandada, kui võtta arvesse vaatlustevahelist sõltuvusstruktuuri;
2. katsetada, kas kovariatsioonistruktuuri kasutamine võimaldab vähendada tundmatute fiktiivsete parameetrite arvu mudelis (kas saadud prognoosid on sama head). Näiteks üldfosfori käitumise modelleerimiseks on hetkel kasutatava mudeli Mudel0 puhul tarvis hinnata 61 tundmatu parameetri väärtused;
3. parima mudeli abil joonistada Peipsi pinna kaart üldfosfori sisalduse kohta.

SAS MIXED poolt hinnatud parameetrite väärtuste ja valemite (2.2) – (2.4) abil saab konstrueerida erinevaid variogramme testimaks nende kokkulangevust empiirilise variogrammiga. Kasutades simuleerimisülesandega saadud teadmisi (punkt 2.2), koostati Peipsi vee üldfosfori semivariogrammid selliselt, et reidid loeti sõltumatuteks. Kauguse arvutamisel võeti arvesse maakera ellipsoidiline kuju, mistõttu Peipsi järve idapikkuste vaheline kaugus on põhjalaiuste vahelise kaugusega võrreldes umbes kahekordne. Parim ehk andmetega kõige rohkem kooskõlas olev mudel valiti välja andmete põhjal hinnatud empiirilist ja suurima tõepära hinnangutega arvutatud semivariogramme võrreldes. Võrreldud semivariogrammid on esitatud Joonisel 15. Silma järgi vaadeldes paistab, et mõlemad struktuurid on küllalti hästi lähendavad. Gaussi mudel sobib paremini väikeste kauguste korral, kuid suuremate kaugustega on erinevus empiirilisest variogrammist suurem. Arvestades ka Tabelis 6 toodud kooskõlakordajate ja katvusprotsentide näitajaid, loeti paremaks siiski eksponentsiaalse kovariatsioonistruktuuriga mudel, mis andis kõige paremaid tulemusi lihtsa mudeli Lmud korral. Arvestamaks igasuguseid trende, kasutati empiirilise variogrammi joonistamiseks valitud fikseeritud kordajatega mudelite Mudel0 või Lmud jääke, et järgida MIXED protseduuri REPEATED käsu kovariatsiooniparameetrite hindamise eeldusi jääkide korreleerituse osas. Järgmised Joonised 15 ja 16 on tehtud Mudel0 jääkide põhjal, Lmud jääkide korral olid empiirilised semivariogrammid samasugused.



Joonis 15. Peipsi üldfosfori andmete empiiriline (Emp) semivariogramm, ja erinevatele kovariatsioonistruktuuridele vastavad mudeli poolt hinnatud semivariogrammid (Exp – eksponentsiaalne ja Gau – Gaussi).

Jooniselt 15 puudub astendajaga mudeli variogramm, kuna prognoositud parameetrite väärtused andsid sama semivariogrammi kui eksponentsiaalse mudeli korral.



Joonis 16. Üldfosfori algandmete põhjal koostatud empiiriline (Emp) ja robustne (Rob) variogramm ning suurima tõepära meetodil (ML) ja kitsendatud suurima tõepära meetodil (REML) hinnatud semivariogrammid eksponentsiaalse kovariatsioonistruktuuriga mudeli korral.

Empiiriliselt variogrammilt võib lugeda välja järgmist: mõõtmisvea dispersioon on umbes 0,14, tunnuse tegeliku väärtuse varieeruvus 0,26 – 0,14=0,12. Punktide vaheline korreleeritus väheneb kuni kauguseni 1,6; kauguste 1,3 – 1,7 juures leidub kõige rohkem erinevaid vaatlusi, kaugemate punktide vaatlused muutuvad aga jälle sarnasemaks. Kaugus 1 on näiteks punktide 2 ja 16 vahel (vt Joonis 5), mis on umbes 70 km. Algandmeid uurides selgus, et kaugeimad punktid asetsevad Peipsi järve loodeosa jõesuudmete ning

Pihkva järve Velikaja jõe suudme lähedal, seega on tegemist mudelisse mittearvestatud olulise faktoriga, nagu jõesuue.

Järve põhjaosa jõesuudmete välja eraldamine nõudnuks täiendavat lisatööd algandmete valdajate poolt. Antud magistritöö eesmärki silmas pidades polnud mudeli täiendamine siiski hädavajalik. Katsetati ka sügavuse lisamisega, kuid varasemate aastate sellekohase informatsiooni puudumise tõttu ei saanud jälle jõesuudmete punkte järve omadest täpselt eraldada. Sügavus osutus küll oluliseks faktoriks, kuid sügavusega mudeli kovariatsiooniparameetrite hinnangud tuli küllaltki sarnased ilma sügavuseta mudeliga hinnatutele (vt Tabel 5). Ka mudeli prognooside katvusprotsent sügavuse arvesse võtmisel ei suurenenud.

Tabel 5. Kovariatsiooniparameetrite *REML* hinnangud erinevate mudelite korral.

Mudel	σ^2	θ	τ^2
Lmud	0,1465	1,6553	0,1443
Sügavusega	0,1392	1,8327	0,1409
Velikajata	0,1395	1,8556	0,1450
Mudel0	0,1391	1,8349	0,1432

Kõik järgmised tulemused on saadud kogu andmestikku kasutades.

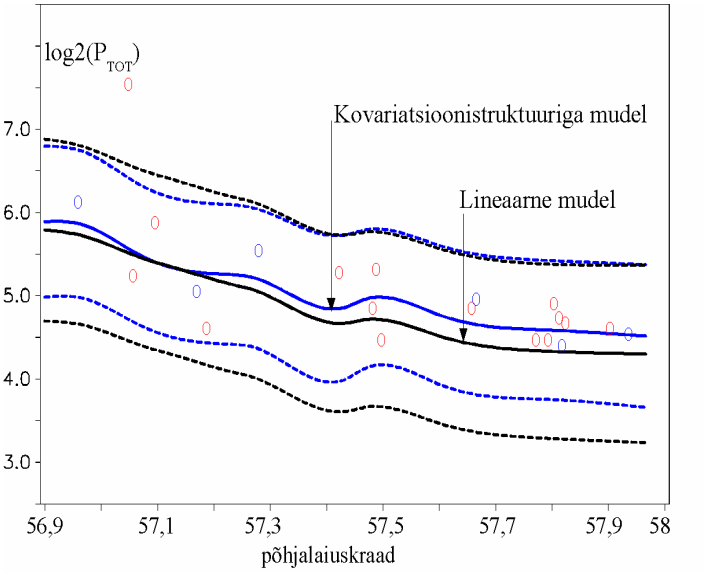
Mudeli prognoosivõime ja headuse kirjeldamiseks tehti mitmeid uuringuid. Kõigepealt jagati olemasolev andmestik kaheks. Poolt olemasolevast andmestikust (800) kasutati tundmatute parameetrite hindamiseks, teist poolt (749) aga mudeli prognoosivõime kontrollimiseks. Paremuse kriteeriumiks võeti mõõdetud väärtuste ja prognoositud väärtuste erinevuste ruutude summa (SSE) ning protsent „õigetest“ prognoosidest (Protsent), st mitu protsenti tegelikest mõõtmistulemustest sattus 95%-sse prognoosiintervalli. Hea mudeli korral peaks 95% uutest vaatlustest (mida mudeli hindamisel ei kasutatud) paiknema antud intervalli sees. Kovariatsioonistruktuurita mudeleid võrreldi eksponentsiaalse, astendajaga ning Gaussi kovariatsioonistruktuuriga mudelitega. Mudelite nimetustega Lmud ja Mudel0 kohta vt punkt 3.1.1. Nimele lisatud täiendid _exp, _gau ja _pow tähistavad vastavalt eksponentsiaalset, Gaussi ja astendajaga kovariatsioonistruktuuri mudelit. Tabel 6 põhjal on näha, et kovariatsioonistruktuuri lisamine parandab hälvete ruutude summat tunduvalt, samuti väheneb Akaike kordaja (AIC), prognoosiintervalli katvusprotsent on parim eksponentsiaalse kovariatsioonistruktuuri korral. Lmud ja Mudel0 lihtsate lineaarsete variantide võrdlemisel

on Mudel0 parem, kuid mitte eriti oluliselt (samas katvusprotsendi poolest on Lmud veidi parem).

Tabel 6. Üldiste lineaarsete ja kovariatsioonistruktuuridega mudelite võrdlus.

Mudel	AIC	SSE	Protsent
Lmud	1247,8	238	93
Lmud_exp	1047,6	162,0	95
Lmud_pow	1047,6	162,0	95
Lmud_gau	1047,7	163,1	94,5
Mudel0	1224,9	226,9	92,5
Mudel0_exp	1068,3	161,6	94,1
Mudel0_pow	1068,3	161,6	94
Mudel0_gau	1067,0	161,7	93,5

Joonis 17 illustreerib, kuidas kovariatsioonistruktuuriga mudeli Mudel0_exp prognoosijoon paikneb olemasolevatele mõõtmistele lähemal kui üldise lineaarse mudeli prognoosijoon.



Joonis 17. Esimese reidi (1985.a. mai) prognoosijooned koos piiridega kindlat teekonda jälgides (lõunast-põhja), nulliga on märgitud olemasolevad mõõtmised (sinisega kindla teekonna lähedased, punasega kõik ülejäänud sama reidi omad).

Tabel 7 sisaldab eri mudelitega arvutatud prognoosivahemike laiusi. Esimeses veerus on prognoosiintervalli laius kogu Peipsit katva võrgustiku jaoks, iga aasta kohta sesoonsed

mõõtmised alates 5. maist kuni 20. oktoobrini (iga 20 päeva järel), nimetame neid fiktiivseteks andmeteks. Fiktiivsete proovivõtu punktide vahekaugused paiknevad umbes 2 kilomeetri raadiuses. Teine veerg kirjeldab prognoosivahemike laiusi nendes proovipunktides, kus on tehtud reaalsed mõõtmised. Usalduspiiride tellimise kohta MIXED protseduuri abil vt Lisa1, [15].

Tabel 7. Võrdlev tabel prognoosipiiride laiuse kohta.

Mudel	**Prognoosivahemikud**					
	Fiktiivsed			**Olemasolevad**		
	min	**keskm.**	**max**	**min**	**keskm.**	**max**
Lmud	2,11	2,12	2,24	2,11	2,12	2,21
Lmud_exp	1,60	1,93	2,28	1,60	1,67	1,78
Mudel0	2,03	2,09	2,75	2,03	2,06	2,32
Mudel0_exp	1,64	2,04	3,59	1,58	1,66	1,79

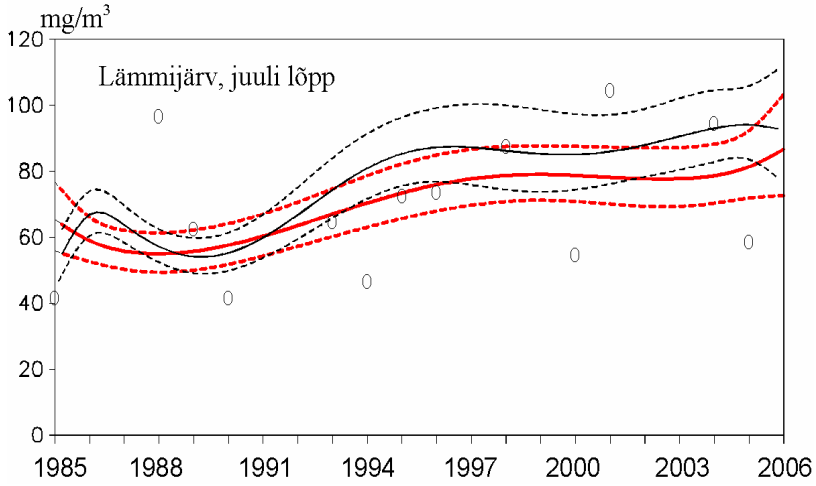
Tabelilt 7 on näha, et kovariatsioonistruktuuride lisamine vähendab oluliselt prognoosivahemiku laiust (väljaarvatud maksimaalne laius fiktiivsete mõõtmiste korral). Kasutades ka Tabeli 6 tulemusi, on näha, et vaatamata kitsamale prognoosipiirile on kovariatsioonistruktuuri kasutavate mudelite katvusprotsent parem, st mudelite Lmud ja Mudel0 jaoks raporteeritud prognoosipiirid on liiga kitsad, tegelikult peaks laiemad olema. Olemasolevaid andmeid prognoosivad kovariatsioonistruktuuriga mudelid paremini.

Tabel 8. Võrdlev tabel keskväärtuse usalduspiiride laiuse kohta.

Mudel	**Keskväärtuse usaldusintervall**					
	Fiktiivsete korral			**Olemasolevate korral**		
	min	**keskm.**	**max**	**min**	**keskm.**	**max**
Lmud	0,13	0,22	0,76	0,13	0,19	0,67
Lmud_exp	0,26	0,34	0,85	0,26	0,33	0,76
Mudel0	0,18	0,47	1,82	0,19	0,38	1,11
Mudel0_exp	0,41	0,72	3,09	0,43	0,67	1,36

Tabelis 8 toodud keskväärtuse usalduspiirid tulevad laiemad kovariatsioonistruktuuriga mudelite korral. See võib olla tingitud ka sellest, et vaatlused on tehtud justkui 137 korral (reidide arv), kuna reidisisesed vaatlused on arvestatud korduvatena.

Järgneval joonisel toome näite hinnatud keskväärtuse ja selle usalduspiiride kohta, lisatud on ka valimis olnud reaalsed mõõtmistulemused. Sellelt jooniselt paistab, et lihtne mudel koos kovariatsioonistruktuuriga prognoosib keskväärtust paremini kui ilma kovariatsioonistruktuurita mudel Mudel0.



Joonis 18. Üldfosfori kontsentratsiooni pikaajalised muutused Lämmijärves kasutades Lmud_exp (punane) ja kovariatsioonistruktuurita Mudel0. Joonisel keskväärtus koos 95% usalduspiiridega. Nullidega on tähistatud mõõdetud vaatlused.

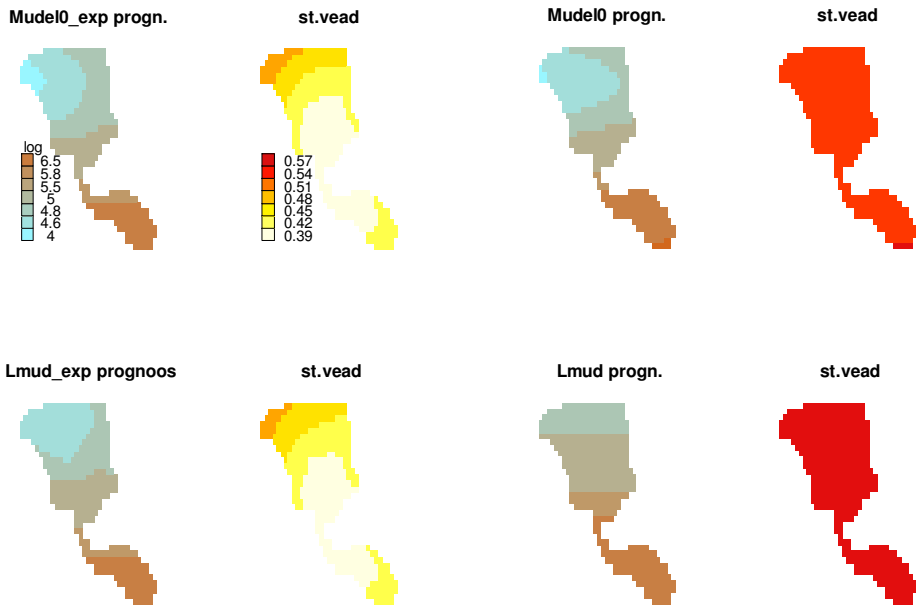
3.2.2 Tulemused

Parimaks mudeliks võib Tabelite 5 – 8 ja Jooniste 17 – 20 põhjal pidada oluliste fikseeritud kordajatega eksponentsiaalse kovariatsioonistruktuuriga mudelit (Lmud_exp). Mudeli kirjeldus koos hinnatud parameetritega on toodud Lisas 1. Mudeli kovariatsiooniparameetrid hinnati järgnevalt: ruumiline skaalaparameeter 1,655, mis näitab, et mõõtmised, mis paiknevad lähemal kui 1,655 on oluliselt korreleeritud. Kaugus 1,7 on näiteks proovipunktide 4 ja 22 vahel (vt joonis 5). Hajuvuse parameetri σ^2 hinnanguks saadi ligikaudu 0,15 ja mõõtmisvea dispersiooni hinnang 0,14. Empiiriline variogramm, tehtud küll Mudel0 jääkide põhjal, näitas ligikaudu sama (Joonis 15), koguvarieeruvus jääb empiirilise variogrammi pealt vaadates siiski alla 0,29.

Joonis 19 näitab ilmekalt, et päevadel, mil on tehtud reaalsed mõõtmised, annab kovariatsioonistruktuuriga mudel täpsemaid prognoose. Joonis 20 näitab, et prognoosi standardvead on oodatult suuremad, kui algandmeid ei ole. Lineaarseid ja kovariatsioonimudelitega mudelite saadud prognooside standardvigu võrreldes näeme, et need on oluliselt väiksemad, kui me arvestame prognoosimisel kovariatsioonistruktuuriga. Autori poolt välja pakutud lihtne mudel Lmud ei sobi ilma kovariatsioonistruktuuri arvestamata prognoosimiseks eriti hästi, nagu oligi arvata. Tõenäoliselt kirjeldavad Mudel0 faktorite koosmõjud võrreldes lihtsa mudeliga ära jääkide struktuurist teatava osa, samas ruumilise korrelatsioonistruktuuri arvestamine omakorda parandab ka Mudel0 prognooside standardvigasid oluliselt. Jooniste vaatamisel tasub meeles pidada, et kovariatsioonistruktuurita mudelite puhul on standardvead arvutatud lähtuvalt vaatluste sõltumatuse eeldusest ja seetõttu üleliia optimistlikud.

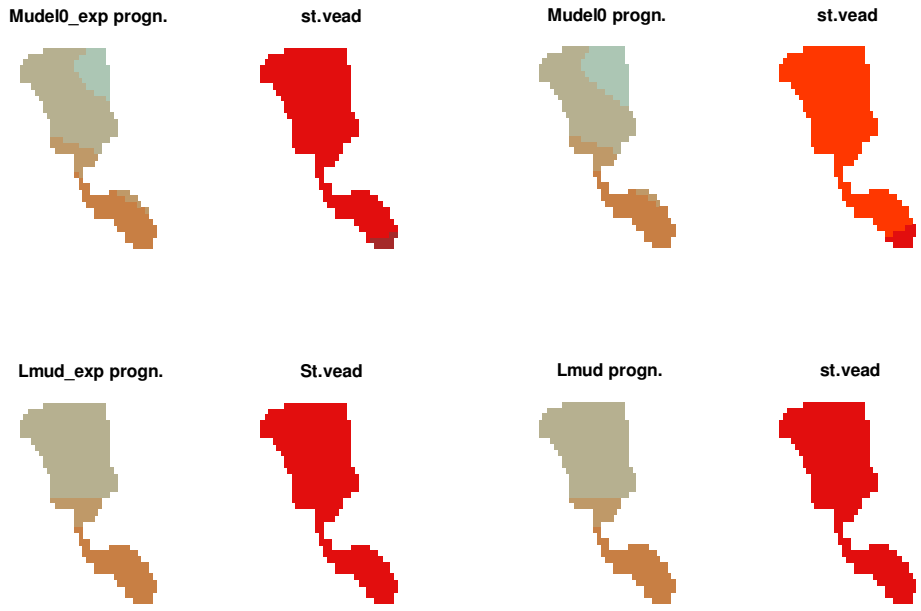
Saadud mudeleid kasutades saab tellida mistahes aja kohta analoogseid kaarte, samuti on võimalik teisendada andmed tagasi algskaalasse bioloogidele mugavamaks jälgimiseks (Joonis 21). Kahtlemata saab teha ka erinevaid väljavõtteid prognoosidest ning nende alusel Joonistega 6. – 8. analoogseid kahemõõtmelisi graafikud.

10. august 1985

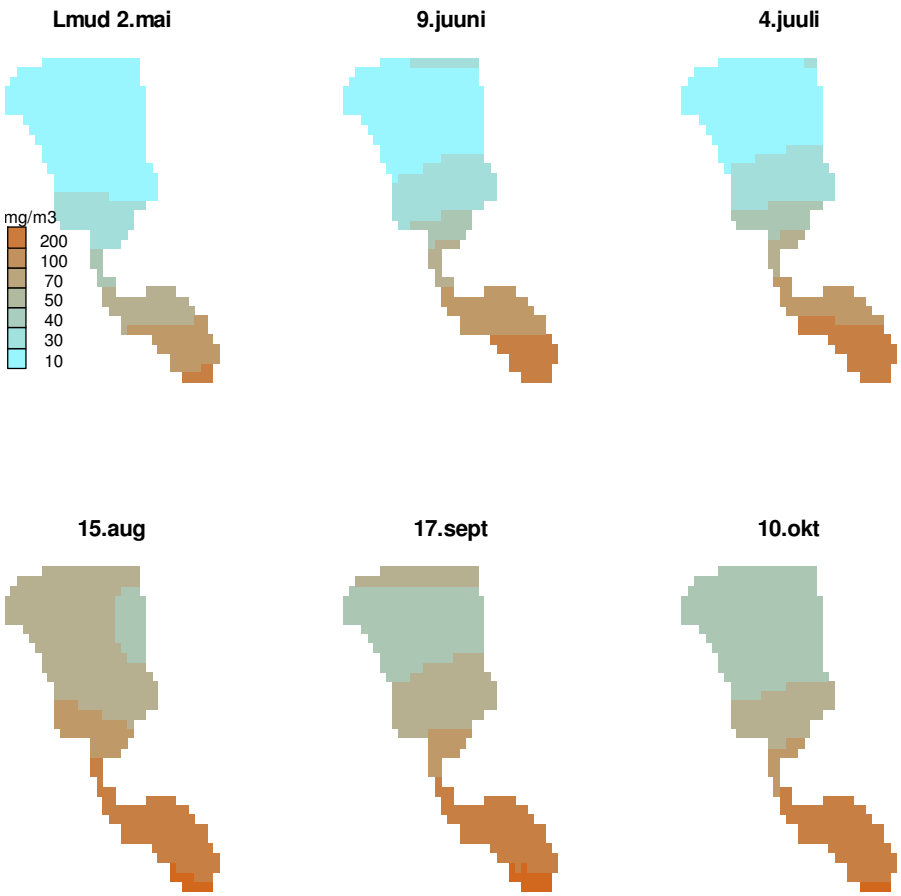


Joonis 19. Kaardid 1985.a üldfosfori prognooside ja nende standardvigade kohta nelja võrreldava mudeli abil**,** logaritmitud andmetega.

10. august 1991



Joonis 20. Kaardid 1991.a. (mil tegelikud vaatlused puuduvad) prognooside ja nende standardvigade kohta nelja võrreldava mudeli abil, logaritmitud andmetega.



Joonis 21. Prognoosid 2006. a. järjestikuste reidide jaoks.

Joonise põhjal võib väita, et fosfori sisaldus suureneb kevadest sügise poole, olles kõige kõrgem augustikuus, hiljem hakkab kontsentratsioon jälle vähenema. Sama näitas ka Mudel0 põhjal tehtud kahemõõtmeline Joonis 8, antud jooniselt saab aga välja lugeda rohkem infot korraga (esindatud on kogu järv).

KOKKUVÕTE

Antud magistritöö teema valikul oli algseks põhjuseks soov täiendada senikasutatud Peipsi järve vee kvaliteedinäitajate muutumist kirjeldavaid üldisi lineaarseid fikseeritud faktoritega mudeleid. Segamudelite abil on võimalik prognoose täpsustada, võttes arvesse vaatluste geograafilisest kaugusest tingitud korreleeritust. Eesmärgi saavutamiseks oli esmalt tarvis tutvuda segamudelite teooriaga, eriti korreleeritud jääkidega mudelite kasutamise võimalustega. Varasemate Peipsi järve uuringute käigus on olnud alust arvata, et lähedamates proovipunktides saadud mõõtmistulemused on rohkem sarnased, kui kaugemates punktides.

Töö esimeses osas on toodud segamudelite teooria osa, mis hõlmab korreleeritud jääkidega mudelite koostamist, mudeli parameetrite hindamist ja prognooside leidmist.

Teises osas antakse ülevaade ruumilistest kovariatsioonistruktuuridest ning empiirilistest ja teoreetilistest variogrammidest kui mudeli koostamise olulistest näitajatest mudeli sobivuse kontrollimisel.

Kolmandas osas on toodud Peipsi järve üldfosfori kontsentratsiooni näitel kogu modelleerimise protsess alates varasemate tulemuste ülevaatest ja uue mudeli väljatöötamisest kuni parima mudeli diagnostika ja rakendamiseni.

Kokkuvõttes võib öelda, et kovariatsioonistruktuuriga mudeli kasutamine on otstarbekas igal juhul. Teine tähelepanek antud töö tulemusena on see, et varemkasutatud küllastunud mudel koos kovariatsioonistruktuuri arvesse võtmisega ei pruugi anda täpsemaid prognoose kui ainult oluliste fikseeritud parameetritega kovariatsioonistruktuuriga mudel. Kuigi simuleeritud andmetega uuring näitas, et me võime kasutada kovariatsioonistruktuuride hindamisel plokkstruktuuri, ilma et prognoosid ebatäpsemaks läheksid, saaks antud mudelit kindlasti ajalisi kovariatsioonistruktuure arvesse võttes veel parandada. Arvestades Peipsi järve asendit, kuju ja vee voolusuunda võiks edaspidi uurida mudeli prognoosivõimet ka kauguste teistsuguse tõlgendamisega.

Marina Haldna

USING COVARIANCE STRUCTURES IN MODELLING WATER QUALITY PARAMETERS

Master Thesis

SUMMARY

Present thesis provides new applications of statistical linear mixed models to the analysis of spatial data of the main water quality parameters – the total phosphorus.Using of mixed models, it is possible to adjust the predictions with computation of correlations between observations.

First chapter presents the short overview of mixed linear models theory and statistical methods for predicting unobserved observations. Under discussion are models with correlated errors.

Second chapter presents different covariance structures used in spatial data analysis.

Under investigation is comparison theorethical and empirical semivariograms as method for estimating the applicability of the models.

Third chapter includes description of the dataset of phosphorus in Lake Peipsi. The research of the data is based on the model developed especially for Lake Peipsi in Estonian University of Life Sciences. One important problem considered in thesis was to examine, if using the mixed models with correlated errors improves the present good-working models. Hypothetical approach is that observations at close quarters are more correlated. Using correlation structures gives better predicted values and standard errors of predictions. Author worked with the three different spatial correlation models (exponential, power and gaussian) using SAS MIXED procedure.

Model diagnostics and results revealed, that the use of mixed models is justified. In the results there are added some maps of phosphorus of Lake Peipsi with predicted values using the model with exponential correlation structure which turned out to be the best.

KASUTATUD KIRJANDUS

1. Aruvee, E. Sesoonsuse ja selle muutuste modelleerimine. *Statistikameetodid keskkonnakaitses ja ökoloogias. Eesti Statistikasektsi teabevihik* 11: 8-24, Tartu, 2001.
2. Banerjee, S., Carlin, B., Gelfand A. *Hierarchical modeling and analysing for spatial data*, 1-94. 2004.
3. Diggle, P., Liang, K., Zeger, S. *Analysis of Longitudinal Data*. Oxford Science Publications, 1994.
4. Haldna, M., Milius, A., Laugaste, R.Nutrients and phytoplankton in L. Peipsi in two periods with different water level and temperature. *Hydrobiologia* (vastu võetud)
5. Henderson, C. R., *Applications of Linear Mixed Models in Animal Breedings* , University of Guelph, Ontario. 1984.
6. Isotalo, J., Puntanen, S., Styan, G. P. H. *Matrix tricks for linear statistical models:our personal Top Sixteen*. University of Tampere, 2006.
7. Littell, R. C, Pendergast, J., Natarajan, R. Modelling covariance structure in the analysis of repeated measures data. *Tutorial in Biostatistics* , 2: 161-185, 2004
8. Milius, A., Laugaste, R., Möls, T., Haldna, M. & Kangur, K. Weather conditions and water level as factors determining phytoplankton biomass and nutrient content in Lake Peipsi. *Proceedings of the Estonian Academy of Sciences. Biology Ecology*, 54: 5–17, 2005.
9. Möls, M. Linear Mixed Models with equivalent predictors. *Dissertationes Mahematicae Universitatis Tartuensis*, 36, Tartu, 2004.
10. Möls, T., Kangur, K., Haldna, M., Milius, A., Haberman, J., Laugaste, R. & Möls, M. The Synthetic Hydrochemical and Hydrobiological Database (SD) for Lake Peipsi Release 1.0). Tartu (CD). 2004.
11. Möls, T. *Lineaarsed statistilised meetodid eesti mageveekogude vee ja elustiku analüüsimiseks*. Eesti loodusuurijate Selts, Tartu, 2005.
12. Möls, T., Saan, T., Lindpere A. & Starast H. *Peipsi troofsusseire*. Tartu, 1990.
13. Robinson, G.K. That BLUP is a good thing-the estimation of random effect. *Statistical Science* , 6: 15-51.1990

14. Searle ,S. R., Casella, G, McCulloch, C. E. *Variance Components*. Wiley, New York, 1992.
15. http://www.asu.edu/sas/sasdoc/sashtml/stat/chap41/index.htm
16. http://v8doc.sas.com/sashtml/stat/chap70/sect8.htm

LISA 1. RUUMILISTE ANDMETE ANALÜÜSIL KASUTATUD SAS PROTSEDUURIDE TUTVUSTUS

MIXED

Protseduur MIXED võimaldab hinnata segamudeleid ning leida mitmesuguseid prognoose, usaldus- ja prognoosiintervalle; testida seonduvaid hüpoteese. Mudeli parameetrite hindamisel saab valida meetodi: antud töös kasutatakse põhiliselt suurima tõepära (method=ML) või kitsendatud suurima tõepära meetodit (method=REML). Mudelite omavaheliseks võrdlemiseks on soovitav kasutada ML meetodit, mis arvestab Akaike kooskõlakordaja arvutamisel nii fikseeritud kui ka juhuslike hinnatavate parameetrite arvu.

Tellimise näide, kus kasutatakse magistritöö koostamisel vaja läinud käske ja võimalusi.

```
proc mixed data= sygavusega covtest method=ML;
class reid ;
model y= syg a pl pl*pl t44 t53/outp=pr outpm=uspr ;
repeated /subject=reid type=sp(exp)(ip pl) local Rcorr;
estimate "sügis Lämmijärves" intercept 1 t53 0.85 pl 1.2;
run;quit;
```

Suurima tõepära meetodite abil hinnatakse dispersiooni- ja kovariatsioonimaatriksid *G* (RANDOM käsu korral) või *R* (REPEATED käsukorral) Rcorr väljastab *R*-le vastava korrelatsioonimaatriksi. REPEATED käsuga saab määrata ka korduvate mõõtmiste jaoks üksteisest sõltumatud vaatluste grupid ja kovariatsioonistruktuuri (näite korral ruumiline eksponentsiaalne). Antud näite korral on erinevatel reididel tehtud mõõtmised loetud sõltumatuteks. Reidisiseselt on uuritava tunnuse *y* mõõtmised korreleeritud, kuidas täpselt, seda määratakse TYPE= tellimisel. Mõõtmise asukoha koordinaadid on antud tunnustega ip ja pl. Sellisel tellimise juures eeldatakse, et ruumiline kovariatsiooniparameeter, dispersioon ja mõõtmisviga on kõikide reidide korral ühesugused.

SAS väljatrüki kovariatsiooniparameetrite hinnangute kohta saame COVTEST abil, seal on kirjas tõepäral baseeruva *Waldi Z* statistiku hinnang, mis arvutatakse kui parameetri hinnangu ja asümptootilise standardvea suhe. Asümptootiline standardviga arvutatakse tõepära teist järku tuletiste maatriksi pööramisel, arvestades igat kovariatsiooniparameetrit. Suurte valimite korral töötab *Waldi Z* valiidselt.

COVTEST, kooskõlakordajate ja ESTIMATE tulemuste väljatrükk

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Variance	reid	0.1753	0.01837	9.54	<.0001
SP(EXP)	reid	0.9324	0.1692	5.51	<.0001
Residual		0.1314	0.008681	15.14	<.0001

Fit Statistics

-2 Log Likelihood	2034.6
AIC (smaller is better)	2054.6
AICC (smaller is better)	2054.7
BIC (smaller is better)	2083.8

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
A	1	135	28.10	<.0001
pl	1	1407	89.81	<.0001
pl*pl	1	1407	47.14	<.0001
t44	1	1407	81.84	<.0001
t53	1	1407	81.56	<.0001

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
sügis Lämmijärves	2.6148	0.4607	1407	5.68	<.0001

SUBJECT=INTERCEPT korral eeldatakse, et kõik mõõtmised on omavahel sõltuvad.

ESTIMATE lause võimaldab hinnata fikseeritud faktorite tasemete lineaarkombinatsioonide väärtusi. Näite korral on hinnatud üldfosfori keskmine sügisene tase Lämmijärves. OUTP väljastab prognoosid koos vastavalt 95% prognoosipiiridega ja OUTPM keskväärtuse koos usalduspiiridega. OUTP korral peab teadma, st prognoosipiirid arvutatakse vaid neile objektidele, millel uuritava tunnuse väärtus puudub (tekstis nimetatud Y_{uus}). Juhul kui tahetakse võrrelda kovariatsioonistruktuuriga mudelit ilma kovariatsioonistruktuurita mudeliga on oluline SUBJECT määramine, seega tuleb valida ka viimasel juhul REPEATED käsk ning struktuuri tüüp *VC* (*Variance Components*), mille korral peadiagonaalil on tunnuse dispersioon ning kovariatsioonid 0).

VARIOGRAM

Protseduur võimaldab tellida etteantud valimi regulaarse ja robustse variogrammi hinnangud.

Näide tellimise kohta:

```
proc variogram data = ok outdistance = kaugus;
coordinates xc = ip yc = pl;
var res ;
compute novariogram nhclasses = 15 OUTPDISTANCE=5;
run;
proc gchart data = kaugus;
vbar lag /sumvar = count discrete;
run;quit;
proc variogram data = ok outvar = vario ;
var res;
compute lagdistance = 0.1 maxlags = 15 OUTPDISTANCE=5;
coordinates xcoord = ip ycoord = pl;
run;
```

Koordinaatide määramine COORDINATES XC= , YC= .

Empiirilise variogrammi arvutamiseks kasutatakse kaugusi. Kuna pidevate koordinaatide korral saadakse erinevaid kaugusi väga palju, siis on mõistlik kaugused koondada nn. kaugusklassidesse. Sobiva arvu kaugusklasside ning kaugusklassi laiuse leidmiseks kasutatakse käsku COMPUTE NOVARIOGRAM NHCLASSES=(klasside arv) OUTDIST=(väljundfail). Väljastatud kaugused trükitakse välja näiteks CHART käsu abil histogrammina ning selle põhjal tehakse otsus klasside arvu kohta. Soovitatav on valida vahemikud nii, et igasse kaugusklassi kuuluks vähemalt 30 paari. Kaugusklasside laiuse ja arvu valik on üsna eksperimentaalne tegevus.

Variogrammi tellimisel määratakse kaugusklassid ja intervalli laius järgmiste käskudega: LAGDISTANCE= ja MAXLAGS= . Esimene näitab kaugusklassi laiust, ehk seda, missuguse maksimaalse kaugusega mõõtmised arvatakse samasse klassi ning teine käsk määrab kaugusklasside arvu.

OUT=(andmefaili nimi) väljastab faili, mis sisaldab kaugusklasside kohta klassi numbri, pikkuse, mitu paari olemasoleva valimi mõõtmistest konkreetsesse kaugusklassi kuulub ja variogrammi väärtuse .COMPUTE OUTPDISTANCE =(arv) abil saab ette anda kauguse, millest väiksematel kaugustel asuvad punktipaarid empiirilise variogrammi arvutamisse kaasatakse.

LISA 2. SAS PROGRAMMI TEKST

Märkus. Ruumi kokkuhoiu mõttes ei ole siin kogu programmi tekst, vaid põhilised makrod koos ühe tellimisega.

```
*Lisame vajalikud makrod mudelitega,aasta ja sesoonsuse teisendusteks
ning Peipsi fiktiivsete koordinaatide moodustamiseks;

%include "C:\Mari\magtoo\makrod2.sas";
%include "C:\Mari\magtoo\mudelid.sas";

******** SALVESTATUD ANDMED SISSE    ;
data reidid2;set pe.reidid2;run;

******** Fiktiivsete andmete moodustamine, iga reidi jaoks    ;

proc sql;create table uus1 as
select distinct a,t,cluster from reidid2;quit;
proc sql;create table uus3 as select a.*,b.a,b.t,b.cluster,1 as in
from realv6re a,uus1 b
order by ip,pl,cluster;quit;
run;
*Fiktiivsed andmed vaatlusteta aasta jaoks;
data uued2;set realv6re;
in=1;cluster=200;a=7.1;t=0.6;output;
run;

data ptot2;set reidid2 uus3 uued2;
%aeg(2);%beta;
* Kuna maakera on lapik, siis idapikkuste vahelised kaugused on
umbes 2 korda suuremad, kui läänelaiuste vahelised;
ip2=ip*2;
run;

libname pe1 "C:/Mari/magtoo/prog1";*Lmud failid;
libname pe2 "C:/Mari/magtoo/prog2";*Mudel0 failid;

*****************************************************************;
******** Andmed valmis, nüüd MIXED protseduuriga hindama*********;
*Hindan parameetrid ja salvestan ühtlaselt üle järve paiknevad   ;
*                  prognoosid                                  ****** ;
*****************************************************************;
*******  Kõik mõõtmised sõltuvad    ****************************;
*****************************************************************;
proc mixed data=ptot2 covtest method=REML;
class cluster;
model y=&mudel0/outpred=pr outpm=uspr;
repeated /subject=int type=sp(exp)(ip pl) local;
run;quit;
data pe2.exp_int;set pr;run;
data pe2.usexp_int;set uspr;run;


*Teeta ei tule nullist oluliselt erinev!;
```

```
*******  Erinevad kovariatsioonistruktuurid *******
*******     koos prognooside salvestamisega ******;

%macro arvuta(struktuur);
proc mixed data=ptot2 covtest method=REML;
class cluster;
model y= &Lmud/outp=pr outpm=uspr;
parms 0.1,1,0.1;
repeated /subject=cluster type=sp(&struktuur)(ip2 pl)local;
run;
title " Prognoosipiirid Lmud &struktuur ";
data pr1;set pr;vahe=upper-lower;run;
proc means data=pr1 n min mean max;by in;var vahe ;run;
title " Usalduspiirid Lmud &struktuur ";
data pr1;set pr;vahe=upper-lower;run;
proc means data=pr1 n min mean max;class in;var vahe ;run;
* Salvestamine;
data pe1.&struktuur;set pr;run;
data pe1.us&struktuur;set uspr;run;

%mend arvuta;
%arvuta(exp);
%arvuta(pow);
%arvuta(gau);

*********************************************;
*****    R image jaoks salvestamine
                   algne skaala     **********;
*********************************************;

* Valikuga annan ette: aasta, sesoonse t*100,mudeli;
%macro valik1(aasta,aeg,mud);
data prog2(keep=pl ip pred pred2 se cluster a t );set pe&mud..exp ;
if in=1;
if abs(a-(&aasta-1920)/10)<0.01;
if abs(t*100-&aeg)<1;
se=stdErrPred;
*Prognoosid algskaalas;
pred2=round(2**pred);upp2=round(2**upper);low2=round(2**lower);run;
PROC EXPORT DATA= WORK.prog2
            OUTFILE= "C:\Mari\magtoo\prog&mud\exp_&aasta&aeg..csv"
            DBMS=CSV REPLACE;
RUN;
%mend;
%valik1(1985,62,1);
%valik1(1985,62,2);


*********************************************;
************Andmed variogrammi jaoks***********;
*********************************************;
proc mixed data=reidid2 covtest method=REML;
model y= &Mudel0 /outpred=jaagid  ;
run;quit;
```

```
*********************************************;
******************** VARIOGRAMM ***************;
*********************************************;
data ok;set jaagid;
ip2=ip*2;
*reidid sõltumatud;
pl=pl+10*cluster;ip2=ip2+10*cluster;
run;
proc variogram data = ok outdistance = kaugus;
coordinates xc = ip2 yc = pl;
var res ;
compute novariogram nhclasses = 25 OUTPDISTANCE=5;
run;
*proc print data=kaugus; *run;
proc gchart data = kaugus;
vbar lag /sumvar = count discrete;
run;quit;
proc variogram data = ok outvar = vario ;
var res;
compute lagdistance = 0.1 maxlags = 25  robust OUTPDISTANCE=5;
coordinates xcoord = ip2 ycoord = pl;
run;
proc print data=vario;run;
* Erinevate variogrammide arvutamine;
data vario3; set vario;
type = '0';vario = variog; output;
vario=rvario; type = '1rob'; output;
* Eksponentsiaalne reididega;
type = 'ML';
sigma=0.1088; theta=1.5837;err=0.1428;
vario = (2*(sigma+err)-2*sigma*(exp(-distance/theta)))/2;output;
type = 'REML';
sigma=0.1391; theta=1.8349;err=0.1432;
vario = (2*(sigma+err)-2*sigma*(exp(-distance/theta)))/2;output;
run;
goptions reset=symbol;

%macro gr4(fail);
 goptions /*reset=global*/
             cback=white colors=(black blue pink orange yellow green
red);
             filename gg "C:\Mari\magtoo\variog\&fail..CGM";
             goptions gsfname=gg device=cgmwpwa
             gsfmode=replace gsflen=80 lfactor=2;
 %mend gr4;
*proc print data=Vario3(obs=50); *run;
%gr4(ekspuus);
title;
axis1 label = (angle=90 rotate=0 "Var(Y_i-Y_j)/2")
order = (0.1 to 0.3 by 0.05)
value=(f=swiss h=1.5) width=2 major=(w=2)
minor=none length=25;
axis2 label = ("Kaugus") order = (0 to 2.5 by 0.2)
value=(f=swiss h=1.5) width=2 major=(w=2)
minor=none length=65;
proc gplot data=vario3;
plot vario*distance=type / frame vaxis=axis1 haxis=axis2;
symbol1 i=none v=star c=blue ;
symbol2 i=join l=2 c=violet ;
symbol3 i=join l=1 c=red ;
symbol4 i=join l=1 c=black ;run;quit;
```

```
data vario3; set vario;
if distance<0.1 then variog=0.12;
type = '0(emp)';vario = variog; output;
type = 'exp';
sigma=0.1391; theta=1.8349;err=0.1432;
vario = (2*(sigma+err)-2*sigma*(exp(-distance/theta)))/2;
output;

type = 'gau';
sigma=0.09779; rho=1.3105;res=0.1553;
vario = (2*(sigma+res)-2*sigma*(exp(-distance**2/rho**2)))/2;
 output;
run;
goptions reset=symbol;
%gr4(uusiperi);
title;
axis1 label = (angle=90 rotate=0 "Var(Y_i-Y_j)/2")
order = (0.1 to 0.3 by 0.05)
value=(f=swiss h=1.5) width=2 major=(w=2)
minor=none length=25;
axis2 label = ("Kaugus") order = (0 to 2.4 by 0.2)
value=(f=swiss h=1.5) width=2 major=(w=2)
minor=none length=65;
proc gplot data=vario3;
plot vario*distance=type / frame vaxis=axis1 haxis=axis2;
symbol1 i=none v=star c=blue ;
symbol3 i=join l=1 c=red ;
symbol4 i=join l=1 c=black ;
run;
quit;

* Mudeli katvusprotsendi leidmine;
* katse-andmebaasi tegemine;

data juhuslikud;do i=1 to 1549;juh=ranuni(1896548);output;end;run;
data ptot2b (drop=i);merge sygavusega juhuslikud;run;
proc sort data=ptot2b;by juh;run;
data ptot_alg;set ptot2b;if _n_<801;in=0;run;
proc sort data=ptot_alg ;by cluster ip pl;run;
data ptot_uus;set ptot2b;if _n_>800;yuus=y;y=.;in=1;run;
proc sort data=ptot2 ;by cluster ip pl;run;
data ptot2;set ptot_alg ptot_uus;run;
proc sort data=ptot2 ;by cluster ip pl;run;

%macro mixreid(struktuur);
title '';
proc mixed data=ptot2 covtest method=ML;
class cluster;
model y= &Lmud/outpred=prexp solution;
repeated /subject=cluster type=sp(&struktuur)(ip pl)local;
run;
title " &struktuur Fiktiivsete jaoks n=749";
data pr1;set prexp;if in=1;vahe=upper-lower;s2=(pred-yuus)**2;
if yuus>lower and yuus<upper then okey=1;
run;
proc means data=pr1 n sum mean std;var vahe s2 okey;run;
%mend mixreid;
%mixreid(exp);
```

LISA 3. PARIMA MUDELI KIRJELDUS

Kasutades järgnevat SAS MIXED protseduuri väljatrükki saab kirja panna valemi prognoosi keskväärtuse ja vaatluste vahelise kovariatsiooni leidmiseks.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Variance	CLUSTER	0.1465	0.01691	8.67	<.0001
SP(EXP)	CLUSTER	1.6553	0.3388	4.89	<.0001
Residual		0.1443	0.008239	17.51	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	2231.40	866.62	132	2.57	0.0111
A	-1175.04	463.49	132	-2.54	0.0124
A*A	231.30	92.7114	132	2.49	0.0138
A*A*A	-20.1821	8.2192	132	-2.46	0.0154
A*A*A*A	0.6601	0.2725	132	2.42	0.0168
t44	-0.3955	0.04262	1405	-9.28	<.0001
t53	0.4181	0.04897	1405	8.54	<.0001
pl*pl*pl	2.2701	0.4548	1405	4.99	<.0001
pl*pl	-15.3093	2.2843	1405	-6.70	<.0001
A*pl	5.2314	0.8748	1405	5.98	<.0001
A*A*pl	-0.5350	0.08529	1405	-6.27	<.0001
A*pl*pl	0.8874	0.2055	1405	4.32	<.0001

$$E(\log_2(Ptot)) = 2231{,}4 - 1157a + 231{,}3a^2 - 20{,}18a^3 + 0{,}66a^4 - 0{,}4t^4(1-t)^4 + 0{,}42t^5(1-t)^3 - 15{,}31pl^2 + 2{,}27\,pl^3 + 5{,}23a \cdot pl - 0{,}53a^2 \cdot pl + 0{,}89a \cdot pl^2$$

$$Cov(Ptot_i, Ptot_j) = 0{,}1465 \cdot \exp\left(- \frac{\sqrt{(pl_i - pl_j)^2 + 4(ip_i - ip_j)^2}}{1{,}6553} \right),$$

kus

$Ptot_i$ on üldfosfori i-nda mõõtmise kontsentratsioon, mõõdetud mgP/m^3;

a=(aastaarv-1920)/10;

t=(päeva number aastas)/365;

pl=(põhjalaiuskraad kümnendarvuna)-57;

ip=(idapikkuskraad kümnendarvuna)-26.