Mudel on …

Mudel on empiiriliste andmete üldistus ja/või teoreetiline konstruktsioon, mis võimaldab prognoosida muutujate väärtusi.

- Mudel **üldistab ja lihtsustab**, ebaoluline jäetakse kõrvale (**üldistav funktsioon**).
- Mudeli koostamine annab teadmisi protsesside iseloomu ja statistiliste seoste kohta (**kirjeldav funktsioon**).
- Mudel annab näidiseid tegelikkuse **võrdlemiseks** teoreetiliste variantidega (juhul kui …) (**võrdlusfunktsioon**).
- Mudel annab (formaalselt) põhjendatud **ennustuse** (**prognoosiv funktsioon**).
- Mudelist saadud prognooside statistilist olulisust on võimalik hinnata, kui uurida lähteandmete ja korduvate prognooside varieeruvust (**tõestav funktsioon**).

Kas statistiline modelleerimine on kirjeldav või tõestav andmeanalüüs?

Säästvusreegel **Parsimoonia** ladina keeles **kokkuhoidlikkus**, parsimoonne = kokkuhoidlik, säästlik

Parsimoonia reegel ehk Occami (või Ockhami) habemenuga (*Occam's razor*) — keskaegsele filosoofile Ockhami Williamile omistatav loogikaprintsiip:

Entia non sint multiplicanda praeter necessitatem

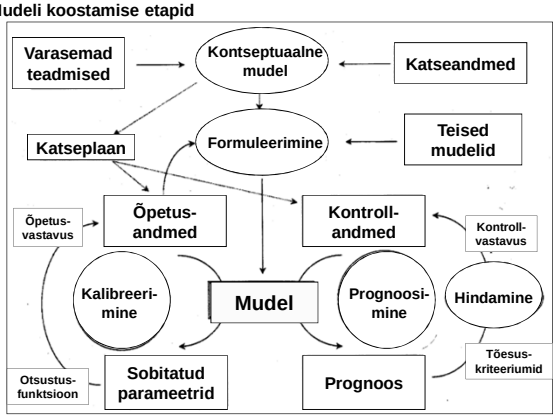
Üheski Occami kirjutises otseselt sellist sõnastus ei ole (Wikipedia).

keerukust ei tohiks eeldada ilma tungiva vajaduseta
ehk
kui mingit nähtust on võimalik seletada lihtsalt, siis ei ole põhjust kasutada keerukat seletust
ehk
tuleb eelistada lihtsamat seletust
ehk
tuleb olla nii laisk, kui võimalik.

Kuidas mõõta keerukust?
? Seletuses (mudelis) kasutatavate tunnuste arv.
? Teisenduste arv.
? Mudeli koostamiseks kulunud tööaeg ja töötasu.

Kuidas mõõta seletuse õigsust?
? Vastavus empiirilise kogemusega.
? Vastavus varasemate teadmistega (ettekujutustega).

Mudeli koostamise etapid



Statistilise mudeli koostamise etapid [Guisan ja Zimmermann, 2000]

Staatilised statistilised mudelid

Pidevad mudelid (funktsioontunnus on pidev)

Lineaarsed mudelid
- **Lihtsad lineaarsed mudelid**
- **Üldised lineaarsed mudelid**
- **Üldistatud lineaarsed mudelid**
- **Teised lineaarsed aditiivsed mudelid**

Mittelineaarsed pidevad mudelid

Aegridade mudelid

Klassifikatsioonimeetodid (f-tunnus on nominaalne)

Klassifikatsioonipuud, regressioonipuud, hägusad (tõenäosuslikud) klassifikatsioonid, diskriminantanalüüs, näidistele tuginevad klassifikatsioonid

Intellektitehnika (hinnangute aluseks on iteratiivne isekohanduv protsess — tehisõpe, võimalik on andmete pidev lisamine)

- **Tehisnärvivõrgud**
- **Ekspertsüsteemid**
- **Näidistele tuginevad süsteemid**
- **Otsuste puud**

Lihtsad lineaarsed meetodid ja mudelid

Funktsioontunnus on pidev

Regressioonanalüüs
pidevad argumenttunnused

Graafiline kujutis:

Regressioonijoon — joon läbi korrelatsioonivälja, millest hälvete (regressioonijääkide) summa on minimaalne. Regressioonijääkide mõõtmiseks on erinevaid variante.

Vastavuspind (*response surface*) — Kahe ja enama pideva argumenttunnuse korral.

Kovariantsanalüüs (*analysis of covariance* — ANCOVA)

Osa **argumenttunnuseid on pidevad ja osa nominaalsed**. Annab teavet, kuidas sõltub pideva argumenttunnuse prognoosimisvõime nominaalse argumenttunnuse klassidest.

Graafiline kujutis:
eraldi **regressioonijoon kategoorilise muutuja iga väärtusklassi puhul.**

Dispersioonanalüüs (*analysis of variance* — ANOVA)
funktsioontunnus on pidev, **kõik argumenttunnused nominaalsed**.

Dispersioonanalüüs on meetod pideva tunnuse varieeruvuse jaotamiseks etteantud gruppide **vaheliseks** ja gruppide **siseseks** varieeruvuseks.

Lihtne lineaarne regressioon võrrandina

Funktsioontunnuse ootus (keskväärtus).

Mudeli võrrand:

$$Y = b_0 + b_1 X_1 + \text{viga}\{\sim N(0,\sigma^2)\} \quad \textbf{ehk} \quad EY = b_0 + b_1 X$$

Funktsioontunnus ehk sõltuv tunnus ehk prognoositav tunnus. Selle väärtusi arvutatakse.

Normaaljaotusega juhuslik viga (saab kasutada prognoosile usalduspiiride lisamiseks). Muude faktorite mõju

Argumenttunnus ehk prognoosiv ehk seletav tunnus (selle väärtused on ette antud).

Vabaliige (inercept) (funktsioontunnuse prognoos juhul, kui argumenttunnus=0).

Regressioonikordaja (regression coefficient). Näitab funktsioontunnuse muutust argumentunnuse ühikulise muutuse korral. Sõltub tunnuste mõõtühikutest!

Korrelatsioon on vastastikune seos.
Regressioon on ühe tunnuse sõltuvus teisest, s.t. et reeglina ei ole regressioonimudel pööratav.

Regressioonimudeli näide

S = – 5.26 + 6.36 · elupymbr + 0.035 · teedzn + 0.013 · kaezn + 0.047 · asustuszn

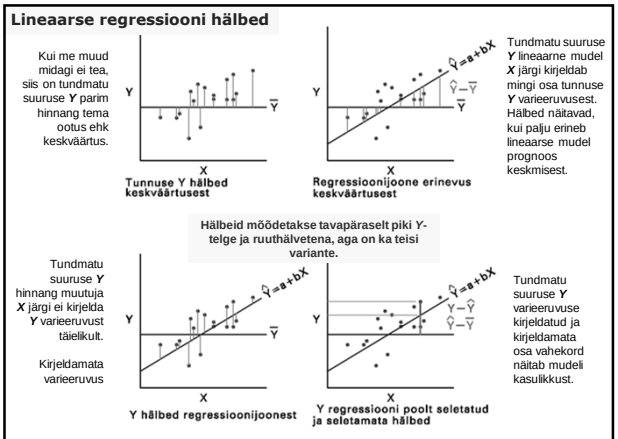
S = logit(põdra esinemistõenäosus) = koha sobivus põdra elupaigana

elupymbr – põdra elupaiga kaugusega kaalutud osakaal ümbruses,
teedz – kaugus maanteedest,
kaezn – kaugus kaevanduste lõhketöödest,
asustuszn – kaugus inimasustusest.

P = exp(S) / {1+ exp(S)}

NB! Võrrandis kasutatud tähistus tuleb ära seletada!

Lineaarse regressiooni hälbed



Standardviga

Andmekogumi varieeruvuse kirjeldamiseks arvutatakse **dispersioon** keskväärtuse suhtes.

$$\sigma^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{N}$$

Keskväärtus kui juhusliku muutuja esmane esindaja (ootus)

Mudeli puhul leitakse prognoosi hälbed vaatlustest.

Kui soovitakse hinnata mudeli täpsust **väljaspool õpetusandmeid** (väljaspool valimit), peaks dispersiooni ja standardhälbe arvutamisel nimetajas olema vabadusastmete arv =

vaatluste arv (*n*) – tunnuste arv mudelis (*t*).

Dispersiooni arvutamisel valimis jagati (n – 1)

$$s^2 = \frac{\sum_i^n (x_i - pred_i)^2}{n - t}$$

$$RMSE = \sqrt{s^2}$$

Ruutkeskmine viga (*standard error of residuals = standard residuals = root mean squared error = RMSE)* ehk prognoosijääkide standardhälve ehk **standardviga** kirjeldab prognoosi oodatavat täpsust.

Vaatluste varieeruvust kirjeldab standardhälve.

Vabadusastmete arv

Sõltumatute muutujate arv mudelis või tunnusruumi dimensioonide arv.

0 mõõtmelises ruumis (punkt) koha määramatust ei ole
1 mõõtmelises ruumis (joonel) on koht määratav ühe arvuga (kaugus 0-punktist)
2 mõõtmelises ruumis (tasapinnal) on koha määratamiseks vaja kahte parameetrit (x ja y või suund ja kaugus või kaks suunda)
3 mõõtmelises ruumis on koha määratamiseks vaja kolme parameetrit (x, y ja z või kaks suunda ja kaugus või suund ja kaks kaugust või kolm suunda)

Lineaarne mudel: $Y = a + bx$
ühe seletava tunnusega (sobitatavad parameetrid *a* ja *b*) kirjeldab kahe andmepunkti paiknemist hälveteta. Läbi kahe punkti saab alati tõmmata sirge.

Lineaarne mudel: $Y = a + b_1x_1 + b_2x_2$
kahe seletava tunnusega (x_1 ja x_2, mudeli parameetrid *a*, b_1 ja b_2)
kirjeldab kolme andmepunkti paiknemist hälveteta.

Kui vaatlusi ei ole rohkem kui mudelis vabadusastmeid, saab kogu varieeruvuse ära seletada, aga me ei tea midagi mudeli kehtivusest väljaspool neid vaatlusi. => **parsimoonia reegel (eelista lihtsat seletust).**

Vähimruutude meetodil sobitatud lihtsate lineaarsete mudelite eeldused

- Faktorite mõjud ja koosmõjud on aditiivsed (liituvad), faktorite vahel puuduvad teistsugused koosmõjud.
- Kõik olulised seletavad tunnused on mudelis sees.
- Mudeli viga on konstantse dispersiooniga kõigi argumenttunnuse väärtuste korral (ei ole heteroskedastsust).
- **Vaatlused on sõltumatud ja ühekordsed**. Vaatlusvead on sõltumatud (ei sõltu vaatluste sarnasusest).
- Mõõtmisvigade keskväärtus on nihketa (valimi keskväärtus on üldkogumi keskväärtuse parim hinnang, valim on **esinduslik**).
- Vead on mudeli suhtes normaaljaotusega.
- Seos funktsioon- ja argumenttunnuste vahel on lineaarne või on andmed lineaarsuse saavutamiseks teisendatud.

Kui mõni eeldus ei ole täidetud, siis peaks otsima teisi modelleerimise meetodeid

Mittelineaarsed regressioonimudelid

Näiliselt mittelineaarsed (loomult lineaarsed) mudelid (*intrinsically linear*)

Päriselt mittelineaarsed mudelid

Argumenttunnuse teisendamisega lineaarseks muutuvad mudelid	**Funktsioontunnuse** teisendamisega lineaarseks muutuvad mudelid	**Mudeli viga ei allu muutujatega samale teisendusele**
Polünoomregressioonid $N = b_0 + b_1x + b_2x^2$ ⇩ $N = b_0 + b_1x_1 + b_2x_2$	$N = e^{b_1 \times aeg}$ ⇩ $\log(N) = b_1 \times aeg$ $N = e^{b_1 \times aeg} \times viga$ ⇩ $\log(N) = b_1 \times aeg + viga$ **Log-mudelis on vead aditiivsed**	$N = e^{b_1 \times aeg} + viga$ Murdepunktiga mudelid $N = b_0 + b_1x(x < 100)$ $N = b_{00} + b_2x(x \geq 100)$

Erinevus on eeldatavas vigade jaotuses

Üldised lineaarsed mudelid (GLM)

- **Mitu funktsioontunnust** — funktsioontunnuste **plaanimaatriks**
- Argumenttunnuste mõjud võivad olla omavahel **korreleerunud**
- Andmed võivad sisaldada **korduvmõõtmisi**
- **Pidevad** ja **nominaalsed** argumenttunnused võivad olla kombineeritud
- Eeldatakse funktsioontunnuse (regressioonivigade) **normaaljaotust**
- Funktsioontunnuste **teisendused** tuleb teha **enne** mudeli sobitamist

Nominaalsed argumenttunnused **kodeeritakse** kas numbriliselt või täheliselt.

Standardiseeritud kodeerimise (*sigma restricted*) korral omistatakse binaarsele tunnusele koodid 1 ja –1.

Üleparametriseeritud mudelis kodeeritakse kõik variandid omaette tunnusteks väärtustega 0 ja 1 (esineb / puudub).

Üldistatud lineaarsed mudelid (GLMZ)

- ⊗ **Mitu funktsioontunnust** — funktsioontunnuste **plaanimaatriks.**
- ⊗ Mõjud võivad olla omavahel **korreleerunud.**
- ⊗ Pidevad ja **nominaalsed** tunnused võivad olla kombineeritud.
- ⊗ Andmed võivad sisaldada **korduvmõõtmisi.**
- **Funktsioontunnuse jaotus võib olla mitmesugune**, k.a. nominaalne.
- Jaotust ja lineariseerivat teisendust (seosefunktsiooni) tuleb eelnevalt **teada.**

Seosefunktsioonid (*link functions*)

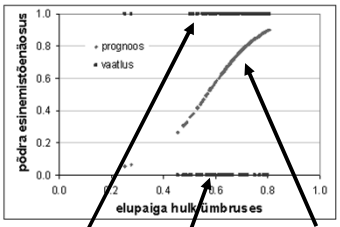
normaal-, gamma-, pöördnormaal- ja Poissoni jaotus:
Identity link: f(z) = z kui eeldatakse vigade normaaljaotust
Log link: f(z) = log(z) loendusandmed, Poissoni jaotusega vead
Power link: f(z) = z^a, funktsioontunnus astendatakse
binoom- ja järjestatud multinominaaljaotus:
Logit link: f(z)= log(z/(1-z))
Probit link: f(z)=*invnorm*(z) *invnorm* on standardiseeritud kumulatiivse normaaljaotuse pöördväärtus
Log-log link: f(z)=-log(-log(z))
Complementary log-log link: f(z)=log(-log(1-z))
multinominaaljaotus:
Generalized logit link: f(z1|z2,…,zc)=log(x1/(1-z1-…-zc)) mudelis on c+1 kategooriat

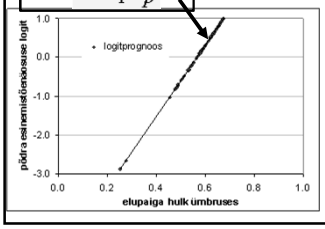
Logitfunktsioon on logaritm šansside suhtest (*odds ratio*) ehk esinemistõenäosuse puudumistõenäosuse vahekorrast ehk tõenäosuste suhtest. **Logitfunktsioon kirjeldab ühesuunalist tendentsi binaarsetes andmetes, kus mõlema variandi esinemise tõenäosus on vahemikus 0…1.**
Probitregressiooni mudel on $y=N(b_0+\Sigma_i b_i x_i)$, kus N on standardiseeritud kumulatiivne normaaljaotus (normaaljaotuse kõvera alune pind). Probitfunktsioon kasutab normaaljaotust pideva muutuja teisendamiseks vahemikku 0…1.

Logitmudel

Logitfunktsioon on logaritm šansside suhtest (*odds ratio*) ehk esinemise puudumise vahekorrast ehk tõenäosuste suhtest. Logit-funktsioon kirjeldab ühesuunalist tendentsi binaarsetes andmetes, milles mõlema variandi esinemise tõenäosus on vahemikus 0…1.

Esinemistõenäosuse logit $= \log \frac{p}{1-p}$





Põdraga vaatlusalad

Põdrata vaatlusalad

Logitmudeliga (logistilise mudeliga) prognoositud esinemistõenäosus

$$p = \frac{e^{\text{logit}(p)}}{1+e^{\text{logit}(p)}}$$

Kas logitmudel on lineaarne mudel või ei ole?

Sõltub vigade jaotusest. Kui logit-teisenduse järel saadakse normaaljaotusega hälbed, siis on mudel näiliselt mittelineaarne, aga sisuliselt siiski lineaarne.

Probleemid lineaarsete mudelite kasutamisel

Probleem	Tagajärg	Kontrollimisviis	Abinõu
Jäägid ei ole normaaljaotusega.	F-statistikul baseeruvad testid ei anna õigeid tulemusi. Nihkega hinnangud.	Jääkide graafik, normaaljaotuse testid.	Funktsioontunnuse **teisendamine**. Üldistatud lineaarse mudeli või **mitteparameetriliste meetodite** kasutamine.
Jääkide hajuvus on ebaühtlane.	Regressiooni täpsuse hinnangud hälbivad.	Jääkide graafik.	Funktsioontunnuse **teisendamine**. Kaalude omistamine vaatlustele. **Mitteparameetrilised meetodid.**
Jäägid ei ole sõltumatud.	Regressiooni tugevus ülehinnatud.	Jääkide graafik, Morani test ruumi-andmete puhul, Durbin-Watsoni test kõigi järjestuvate andmete puhul.	Iteratiivsed sobitamismeetodid. **Üldistatud lineaarne mudel.**
Seose mittelineaarsus.	Mudeli vähene kirjeldav võime, sisutud tulemused.	Korrelatsiooniväli.	Muutujate **teisendamine**. Mittelineaarne mudel. **Mitteparameetrilised meetodid.**
Argumenttunnuste multikollineaarsus.	Ebastabiilne mudel.	Korrelatsioonimaatriks.	Argumenttunnuste arvu vähendamine, argumenttunnuste **teisendamine**, riff-regressioon.
Võimalike argumenttunnuste suur hulk.	Raske valida optimaalset tunnuste komplekti.	Tunnuste loend.	Samm-sammuline regressioon, muutujate **teisendamine**, mudeli lihtsustamine.
Erindid.	Mudel on erinditest liigselt sõltuv.	Kirjeldav andmeanalüüs.	Erindite **eemaldamine**. **Mitteparameetrilised meetodid.**
Vead andmetes.	Sisutud tulemused.	Kirjeldav andmeanalüüs.	Vigaste andmete **eemaldamine** või **asendamine**.
Lünklikud andmed.	Olemasolevate andmete ebaefektiivne kasutus.	Risttabel.	Puuduvate andmetega vaatluste **eemaldamine** või **asendamine** prog-noositud väärtustega või keskmisega.
Nominaalsed tunnused.	Normaaljaotust eeldav regressioon ei ole kasutatav.	Väärtuste jaotumise vaatlus.	Logistiline regressioon või muu **üldistatud lineaarse mudeli** variant.

Ebasobiva meetodi rakendamise tagajärjed

- Mitteparameetrilise meetodi kasutamine, kui parameetrilise meetodi eeldused on täidetud => otsitav seos või erinevus jääb avastamata.
- Kaalude omistamine vaatlustele => tulemuste (meelevaldne) muutmine.
- Keerukama meetodi kasutamine kui piisaks lihtsamast => uurimus ei jäta usaldusväärset muljet, risk eksida keeruka meetodi paljude parameetrite ja variantide hulgast sobivaima valikul, suurem arvutuste maht ja kulutatud aeg, tulemusi on keerukam kontrollida.
- Tunnuste teisendamine => tunnuste muutumispiirkondade mõju tulemustesse muutub.
- Erindite eemaldamine => erindid ei pruugi olla vigased vaatlused, võib olla esindavad mingit olulist nähtust. Tulemuse sobitamine oma suva järgi erindeid valides.
- Vigaste andmete asendamine => subjektiivsus.
- Puuduvate väärtuste asendamine keskmistega => varieeruvus väheneb. Kas varieeruvust on tarvis hinnata?
- Puuduvate väärtuste asendamine juhuslike väärtustega => seoste, varieeruvuse ja keskmise muutumine.
- Lünklike vaatluste eemaldamine => kasutamata jäävad ka mõõdetud tunnuste väärtused.
- Väärtuste juhuslik segipaiskamine (permuteerimine) => kasutatav vaid jaotuse genereerimiseks nullhüpoteesi kehtimisel.

Näidiste järgi prognoosimine

Näidiste järgi saab nii **klassifitseerida** (klassikuuluvust prognoosida) kui ka pideva muutuja väärtusi (puude katvust, liigi esinemistõenäosust, mingi protsessi intensiivsust) **prognoosida**.

Kuna enamasti täpselt vastavat näidist ei leidu, siis kasutatakse pidevate andmete puhul prognoosina sarnasemate näidiste **sarnasusega kaalutud keskmist**.

Sarnasuse järgi prognoosimisel kasutatakse:

kõiki vähemalt teatud sarnasusega näidiseid	**teatud hulka kõige sarnasemaid näidiseid**
***d*-lähima naabri meetod ehk d-*NN* meetod**	***k*-lähima naabri meetod ehk *k*-*NN* meetod**

d-lähima naabri meetodi puhul varieerub näidiste hulk, mille järgi prognoos arvutatakse. Kui näidised ei paikne tunnusruumis ühtlaselt, on hinnangu usaldusväärsus varieeruv.

Kui piisavalt sarnaseid näidiseid ei ole, siis prognoosi ei saa.

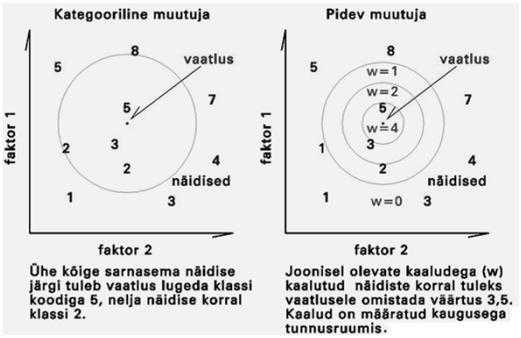
k-lähima naabri meetod on kasutatav igasuguse andmete tiheduse juures. k-lähimat naabrit on lähimad naabrid ka siis, kui nad väga kaugel on.

Prognoos saadakse ka juhul, kui sarnaseid näidiseid ei ole, hinnangu täpsus varieerub.

Näidiste kaalumine

Klassifitseerimine näidiseid kaalumata

Pideva tunnuse prognoos kaalutud näidiste järgi



Ühe kõige sarnasema näidise järgi tuleb vaatlus lugeda klassi koodiga 5, nelja näidise korral klassi 2.

Joonisel olevate kaaludega (w) kaalutud näidiste korral tuleks vaatlusele omistada väärtus 3,5. Kaalud on määratud kaugusega tunnusruumis.

Kas siin kujutatu on k-NN või d-NN meetod?

Näidistebaas

Näidiste järgi prognoosimiseks on mudeli asemel tarvis **näidiste baasi** ja **tarkvara**, mis leiaks andmebaasist kiiresti iga juhtumi puhul sovivaimad näidised ning tuletaks näidistest hinnangu või ennustuse.

Näidistega sarnasusele tuginevaks tehisõppeks, hinnangute ja hinnangukaartide arvutamiseks on geoinformaatika õppetoolis loodud tarkvarasüsteem *Constud*.

http://www.geo.ut.ee/CONSTUD

Aja modelleerimine

Aeg kui **funktsioontunnus**
uuritavaks suuruseks on aeg, mis kulub teatava sündmuseni (surmani, haigestumiseni, detaili riknemiseni jmt). Näiteks ***(elu)kestusanalüüs***.

Aeg kui **müratunnus**
Vaatlused peaksid pärinema samast ajamomendist, kuid mingitel põhjustel on vaatlusaeg veninud pikemaks. Üldine probleem kaartide ja fotode kasutamisel ja ka väliandmete kogumisel. Lahendus võib olla aja lisamine argumenttunnuste hulka.

Aeg kui **argumenttunnus**
Ajas muutuva tunnuse või tunnuste väärtusi nimetatakse **aegreaks**.
Aegridade mudelites on aeg põhiline (tihti ainus) argumenttunnus.

Tavaliselt on aegreas:
iga ajahetke kohta vaid üks vaatlus;
vaatluste koguarv suur;
aega käsitletakse ühtlaselt diskreetsena (fikseeritud sammuga aegrida).

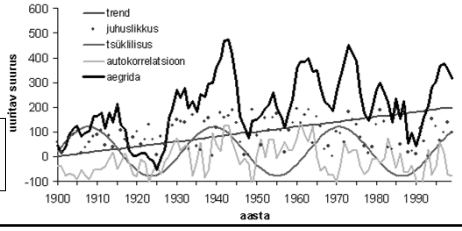
Aegrea komponendid

Andmete varieeruvus aegreas

Argumenttunnustega seotud varieeruvus.
Seda soovitakse mürast eraldada

Müra (juhuslikud hälbed)
Prognoosida saab müra keskmist tugevust ja üksikvaatluse ootust, aga mitte üksikvaatluste juhuslikke hälbeid.

Aegridade modelleerimiseks kasutatakse:
trendi (kindlasuunalist muutust),
sesoonsust (tsüklilisust),
autokorrelatsiooni (endaga sarnasust).



Aegrida on siin joonisel komponentide summa

Mudeli kontrollimine ja parima mudeli valik

Otsustusfunktsioon (*objective function*) — parema mudeli valimise aluseks olev kriteerium.

Liigsobitumine (*overfitting*) — olukord, milleni jõutakse mudeli liigsel sobitamisel andmetele. Liigsobitunud mudeli lahend annab õpetusandmete puhul tunduvalt parema prognoosi kui väljaspool õpetusandmeid.

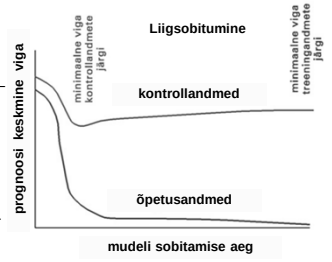
Treeningandmestik ehk **õpetusandmestik** (*training data set*) — vaatlusandmed, mille järgi mudelit kalibreeritakse.

Kontrollandmestik (*validation data set*) — vaatlusandmed, mille järgi mudeli kalibreeringut ja seletavat võimet kontrollitakse.

Ristkontroll (*cross-validation*) — prognoosi täpsuse kontroll õpetusandmetest **sõltumatu** kontrollandmestikuga.

Kontrollandmestikke võib olla üks või mitu.

Vaatluste ükshaaval kõrvalejätmine *(LOOC — leave-one-out cross-validation).*



Üleõppimine ja tühitargutamine tekib siis, kui vähestest faktidest üritatakse saada kaugeleulatuvaid järeldusi.

Prognoositäpsuse hindamine ja mudelite võrdlemine

Mõõdupuud (otsustusfunktsioon)

Normaaljaotusega muutuja — ruutkeskmine viga või determinatsioonikordaja, s.o. mudeli poolt ärakirjeldatud varieeruvuse osa muutuja varieeruvusest keskväärtuse suhtes.
Sobib pideva tunnuse mudelite puhul, kui funktsioontunnuse hälvete ruutude kasutamine on põhjendatud.

Muudel juhtudel:
Hinnatakse jaotuste (klassifikatsioonide) vastavust —
Kapa,
vigade maatriks,
χ^2 test,
sarnasuskordajad,
jt. meetodid (tõepärasuhe, tõepäraskoor, Akaike informatsiooni kriteerium).

Nendest oli juttu eelmistes loengutes

Mõõtmismeetodid:
1) ristkontroll
2) üksikvaatluste kordamööda kõrvalejätmine (*jack-knife*),
3) korduvad tagasipandad juhuslikud valimid (*bootstrap*),
4) juhuandmetest saadud prognooside varieeruvusega võrdlemine.

Tulemus

Treeningtäpsus (*training accuracy*)
Hinnatakse samadest andmetest, mille järgi mudel koostati.

Kontrolltäpsus (*test accuracy*)
Hinnatakse treeningandmetest sõltumatust kontrollandmestikust.

Võiks meelde jääda

- Mudel on **teooriale** tuginev üldistus. Mudel ei teki vaid andmetest.
- Säästvusreegli kohaselt tuleb eelistada lihtsamat mudelit ja seletust.
- Kalibreerimisel saadakse andmetele sobitatud mudeli parameetrid.
- Funktsioontunnuse ootust prognoosiv lihtne lineaarne regressioonivõrrand sisaldab
  - funktsioontunnust,
  - vabaliiget,
  - regressioonikordajaid,
  - argumenttunnuseid.
- Regressioonivõrrandi pööramisel ei saada oodatavalt parimat lahendit.
- Logit on logaritm esinemise ja puudumise tõenäosuste suhtest.
- Sarnasuse järgi prognoosimisel mudelit ei looda.
- Sõltumatu kontrollandmestik võimaldab hinnata mudeli üldkehtivust (kehtivust väljaspool õpetusandmeid).