www.emu.ee
Eesti Maaülikool
Estonian University of Life Sciences

Diskriminantanalüüs
Klassifitseerimis- ja regressioonipuud
Mitme tabeli koosanalüüs

Tanel Kaart
Mitmemõõtmelise statistika koolitus
Jaanuar 2016, Tartu

Diskriminantanalüüs

Diskriminantanalüüsi eesmärgiks on objektide rühmitamine nendel mõõdetud tunnuste alusel.

Seejuures on objektide klassidesse kuuluvus enne analüüsi teada (erinevalt peakomponent või klasteranalüüsist).

Diskriminantanalüüs

Fisheri lineaarne diskriminantanalüüs

Objektidel mõõdetud tunnuste alusel kostrueeritakse nn diskremineeriv funktsioon, mis eristaks gruppe võimalikult selgelt:

$$d = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

Näiteks

2,3×"jalgade pikkus" + 6,7×"saba pikkus" – 2,8×"kere pikkus" + 5,1×"noka pikkus"

Kui saadud väärtus <10,2, siis on ilmselt tegu isase isendiga.

Bayesi diskriminantanalüüs

Hinnatakse tunnuste vektori **x** tihedusfunktisoon $f_t(\mathbf{x})$ igas grupis t ning arvutatakse iga objekti mingisse gruppi kuulumise tõenäosus Bayesi valemist kujul

$$\mathrm{P}(t \mid \mathbf{x}) = \frac{q_t f_t(\mathbf{x})}{\sum_u q_u f_u(\mathbf{x})},$$

misjärel määratakse iga objekt tõenäoliseimasse gruppi (suurus q_t eelnevas valemis on objekti gruppi t kuulumise algtõenäosus).

Diskriminantanalüüs – näide

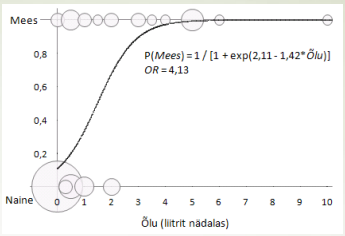
Jaanika Hõimra

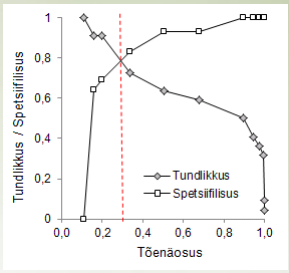
Morfomeetriliste tunnuste varieeruvuse sõltuvus keskkonnatingimustes kortslehe (*Alchemilla* L.) viiel mikroliigil eksperimendi tingimustes

Tabel 8. Klassifitseeriv diskriminantanalüüs 18 tunnuse alusel (2001). Ridades on antud empiiriliselt määratud liigid ja tulpades prognoos.

	%	KAREDA-KARVANE	VÄIKE	TERAVA-HÕLMINE	KÜÜT
KAREDAKARVANE	93,2	**262**	19	0	0
VÄIKE	99,7	1	**286**	0	0
TERAVAHÕLMINE	91,5	0	0	**259**	24
KÜÜT	87,4	0	0	36	**250**
KOKKU	92,9	**263**	**305**	**295**	**274**

Diskriminantanalüüs *vs* logistiline regressioon





Logistilise regressioonanalüüsi tulemus:

nii mehed kui ka naised identifitseeritakse õigesti 80%-lise tõenäosusega.

Et diskriminantanalüüs loeb objekti kuuluvaks suurima tõenäosusega gruppi, ei pruugi saadav klassifitseerimiseeskiri olla optimaalseim.

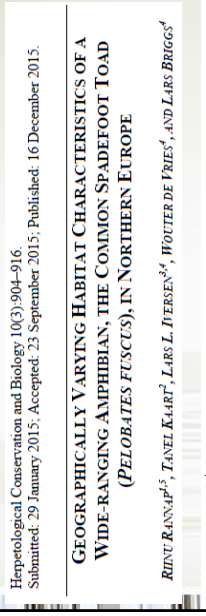
Diskriminantanalüüsi tulemus:

meestest identifitseeritakse õigesti 59,1%, naistest 92,9%.

Number of Observations and Percent Classified into SUGU

From SUGU	0	1	Total
0	39 92.86	3 7.14	42 100.00
1	9 40.91	13 59.09	22 100.00
Total	48 75.00	16 25.00	64 100.00

Diskriminantanalüüs, näide: Rannap jt.



Herpetological Conservation and Biology 10(3):904–916.
Submitted: 29 January 2015; Accepted: 23 September 2015; Published: 16 December 2015.

GEOGRAPHICALLY VARYING HABITAT CHARACTERISTICS OF A WIDE-RANGING AMPHIBIAN, THE COMMON SPADEFOOT TOAD (*PELOBATES FUSCUS*), IN NORTHERN EUROPE

RÄNE RANNAP[1,5], TANEL KAART[2], LARS L. JYERSEN[3,4], WOUTER DE VRIES[4], AND LARS BRIGGS[4]

TABLE 2. Results of canonical discriminant analyses (CDA) of the aquatic and terrestrial habitat characteristics measured.

Country	Netherlands		Denmark		Estonia	
Variables in CDA [a]	All	Selection	All	Selection	All	Selection
Prediction ability of canonical variables according to the logistic regression analysis [b]						
Sensitivity (%)	100.0	92.3	73.3	64.7	87.5	77.8
Specificity (%)	93.8	72.7	83.5	78.9	91.5	87.0
AUC	0.983	0.916	0.892	0.838	0.967	0.909
Description of canonical discriminant functions (CDF)						
Mean CDF (with larvae)	2.25	1.40	1.51	1.11	2.23	1.95
Mean CDF (without larvae)	-0.91	-0.55	-0.19	-0.14	-0.38	-0.29
Raw canonical coefficients [c]						
Type of water body						
Natural depression	-1.749	-	0.265	-	-1.851	-
Lake	2.475	1.892	-1.096	-	-1.479	0.098
Man-made	0.138	-	0	-	-1.382	-
Beaver pond	-	-	-	-	0	1.818
Meander	0	-	-	-	-	-
Aquatic characteristics						
Area ($x10^{-6}$)	-4.358	-	-0.071	-	0.002	-
Shallow area ($x10^{-6}$)	8.608	1.198	3.665	4.103	0.050	-
Maximum depth	0.318	-	0.477	-	0.400	-
Uncultivated area ($x10^{-6}$)	-0.326	-3.960	-1.786	-	0.535	-
Average slope ($x10^{-6}$)	-1.172	-	-0.730	-0.486	-0.770	-0.494
Shadow ($x10^{-6}$)	0.552	-	-1.712	-	-0.910	-
Sediment						
Peat	-1.778	-	-0.717	-	-0.009	-
Mud	0.715	-	-0.349	0.122	0.038	-
Clay	2.536	1.944	0.556	1.001	0.643	0.335
Sand	0	-	0.521	-	0	-
Water						
Brown	3.505	-	1.490	-	-0.962	-
Clear	2.228	-	1.409	-	-1.405	-
Muddy	0	-1.041	1.274	-	-0.876	-
Algae-green	-	-	0	-	0	-
Number of water bodies						
<100 m	-0.238	-	-	-	0.274	-
100-200 m	0.394	-	-	-	0.229	0.169
200-500 m	0.206	-	-	-	0.062	-
Forest edge	0.0012	-	-	-	0.010	0.0107
Habitat within 50 m (presence)						
Coniferous forest	0.390	-	0.163	-	-0.214	-0.100
Deciduous forest	-0.178	-0.701	-0.200	-	-0.370	-0.573
Bogs/swamps	-	-	0.037	-	0.430	-
Crop field	2.294	1.354	0.394	-	0.016	-
Vegetable garden/field	1.034	1.394	-0.938	-1.191	-0.706	-
Gravel/sand pit	0.154	0.085	-	-	-	-
Meadow/fen	-1.112	-0.586	-1.098	-	0.611	0.391
Presence of fish	1.145	-	-0.688	-0.972	-1.478	-1.127

[a] For different levels of categorical variables ("Type of water body," "Sediment," and "Water") numerical dummy variables were used. For each country two analyses were performed: first, all aquatic and terrestrial characteristics observed were involved ("All"); second, only characteristics showing higher prediction ability (R^2>2.5% in univariate ANOVA performed by SAS procedure CANDISC) were considered ("Selection"). Symbol "-" denotes the characteristics were not observed or did not vary in specific countries ("All"), or did not show the prediction ability over fixed threshold ("Selection").

[b] Sensitivity and specificity indicate the proportion of water bodies with and without Common Spadefoot Toad larvae predicted correctly by canonical variable (the most optimal threshold corresponding to the maximum sum of sensitivity and specificity was used), respectively; AUC is the area under the ROC-curve;

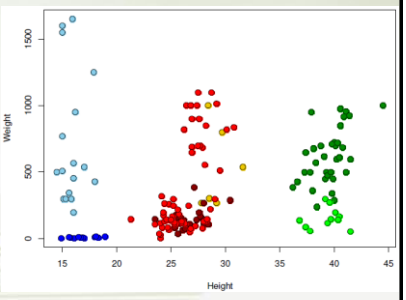
[c] Raw canonical coefficients are the multipliers of variables in discriminant function; in case of dummy variables and complete set of levels, they present the difference from the last level of present characteristic (these coefficients must be interpreted in conjunction with mean values of canonical discriminant functions in water bodies with and without larvae).

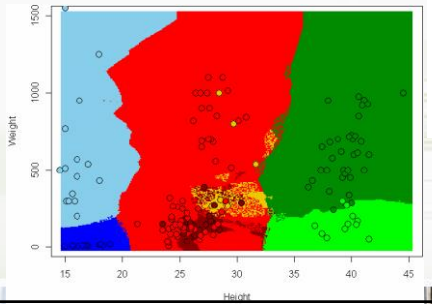
Diskriminantanalüüs

Naabritel põhinev diskriminantanalüüs
[*k-Nearest Neighbour Classification*]

Meetodi idee: otsitakse *k* antud vaatlusele kõige sarnasemat vaatlust ja lahterdatakse uus vaatlus sinna gruppi, kuhu kuulus (kõige enam) temaga sarnaseid vaatluseid.

R-s:
library("class")
knn(…)





Otsustuspuud (*decision tree*), vähe spetsiifilisemalt ka

klassifitseerimis- ja regressioonipuud

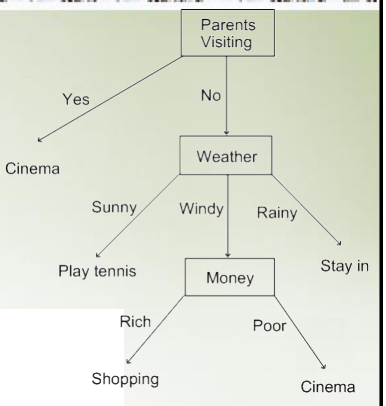
(CART, *classification and regression tree*)

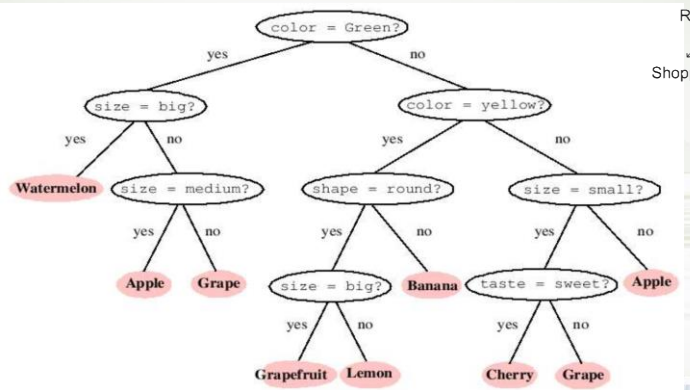
Kujutavad enesest alternatiivi erinevatele mitmemõõtmelise statistika meetoditele ja mudelitele – diskriminantanalüüsile, dispersioonanalüüsile jm.

Võtavad arvesse nii seoste võimaliku mittelineaarsuse kui ka interaktsioonid.

Otsustuspuu (*decision tree*)

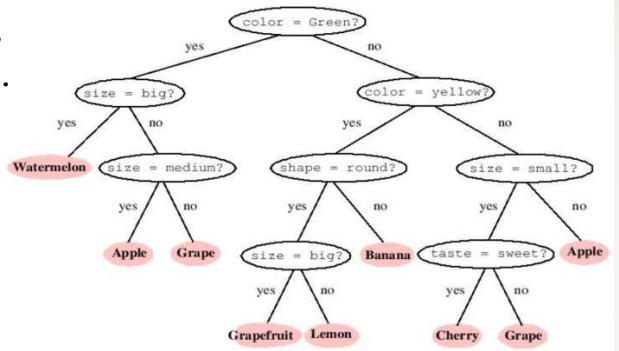
- Otsustuspuu on otsustusprotsessi kujutav joonis.
- Matemaatiliselt on tegu prognoosimudeliga.





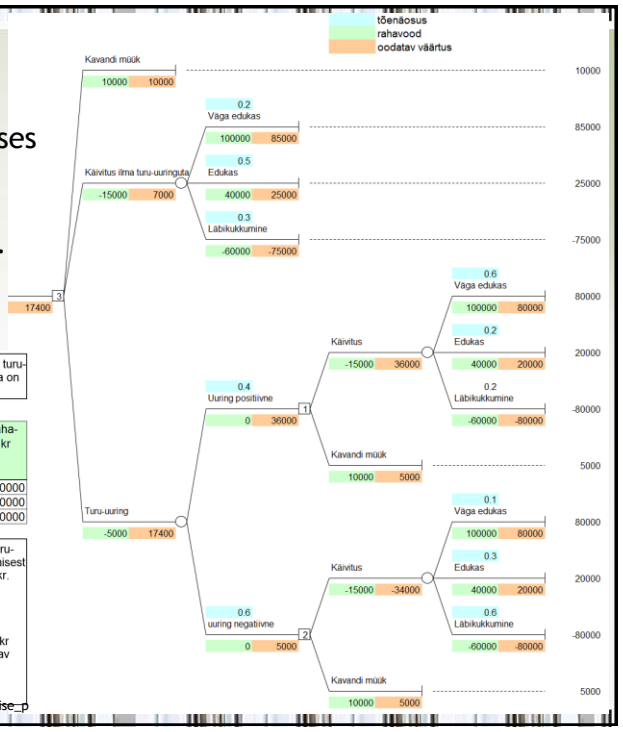
Otsustuspuu (*decision tree*)

- Puu igas **sõlmes** e **tipus** (ingl. ***node***) paikneb küsimus.
- Sõlmest lahknevad puu **harud** (***branches***), kusjuures igale küsimuse vastusele vastab eraldi haru.
- Viimast tippu, kust edasi harunemist enam ei toimu, nimetatakse **leheks** (***leaf*** või ***final node*** või ***terminal node***).



Otsustuspuu

Majanduses (turunduses jmt) ühendatakse otsustuspuu sageli kasumiga (vmt-ga).



Toimub uue toote arendus. Tootmise käivituskulud on 15 000 kr, turu-uuring maksab 5000 kr. Tõenäosused toote müügiedukuse kohta on toodud alljärgnevas tabelis.

Toote müügiedukus	Erinevate sündmuste tõenäosused p Ilma turu-uuringuta	Turu-uuring annab positiivse tulemuse	Turu-uuring annab negatiivse tulemuse	Tulu (raha-voog), kr
Väga edukas	0.2	0.6	0.1	100000
Keskmine	0.5	0.2	0.3	40000
Läbikukkumine	0.3	0.2	0.6	-60000

Eelmiste sedalaadsete toodete kogemus näitab: tõenäosus, et turu-uuringu tulemus on positiivne, on 0,4. Kui loobuda toote käivitamisest ja müüa ideekavand maha, on võimalik saada selle eest 10 000 kr.

Otsustamiseks kasutatav otsustamise puu on toodud lehel "Puu"
Võimalikud variandid
1. Müüa ideekavand maha. Oodatav väärtus 10 000 kr
2. Käivitada tootmine ilma turu-uuringuta. Oodatav väärtus 7000 kr
3. Tellida turu-uuring ja otsustada siis, kas toota või mitte. Oodatav väärtus 17 400 kr

Soovitav valida 3. variant (suurima oodatava väärtusega).

http://www.sauga.pri.ee/qme/otsustamise_puu.html

Klassifitseerimis- ja regressioonipuu (CART, *Classification and Regression Tree*)

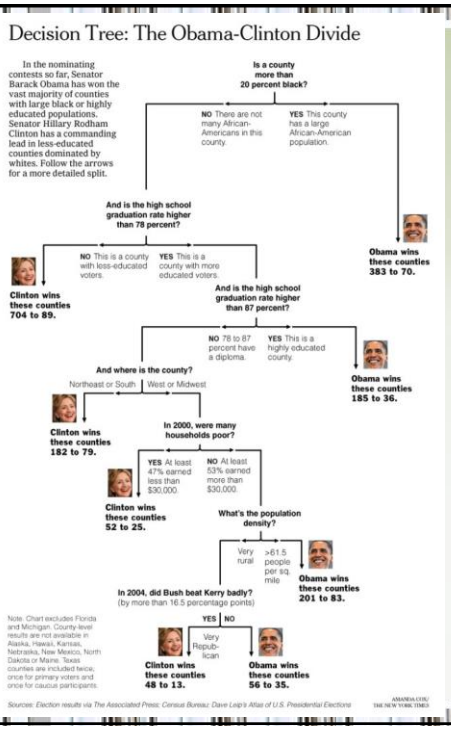
- CART annab suhteliselt arusaadavad prognoosimise eeskirjad ka siis, kui potentsiaalseid mõjutegureid on palju ning need on nii prognoositava suurusega kui ka omavahel mittelineaarselt seotud.
- Rakendamine on lihtne – vaja vaid ühte prognoositavat, ükskõik kas diskreetset (kategoorilist) või pidevat näitajat, ja potentsiaalseid riskifaktoreid, mis võivad samuti olla nii pidevad kui ka diskreetsed.
- Paindlik – puuduvad eeldused selle kohta, kuidas potentsiaalsed riskifaktorid peaksid prognoositavat näitajat mõjutama.
- Ebaolulised riskifaktorid jäetakse automaatselt kõrvale.

Klassifitseerimis-puu

Prognoositav näitaja on kategooriline.

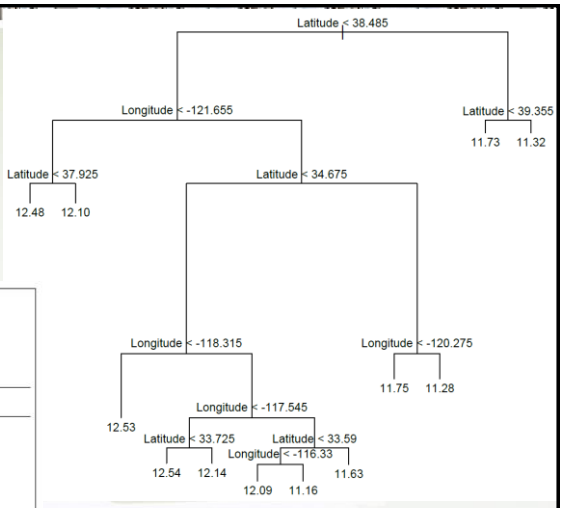
Näide.

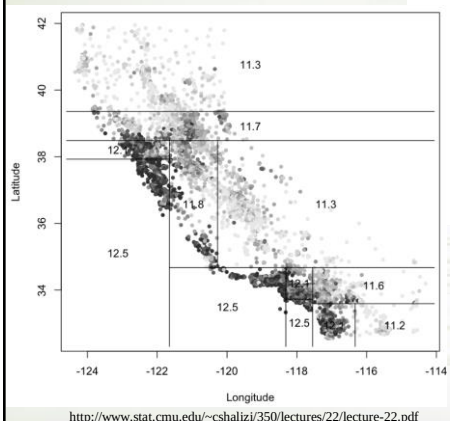
USA demokraatliku partei esinumbri valimiste otsustuspuu maakondade andmete alusel 2008. aastal (Amanada Cox, New York Times).



Regressioonipuu

Prognoositav näitaja on arvuline (pidev).





http://www.stat.cmu.edu/~cshalizi/350/lectures/22/lecture-22.pdf

```
require(tree) # vajalik pakett 'tree'
treefit = tree(log(MedianHouseValue) ~ Longitude+Latitude, data=calif)

plot(treefit)
text(treefit,cex=0.75)

price.deciles = quantile(calif$MedianHouseValue,0:10/10)
cut.prices = cut(calif$MedianHouseValue,price.deciles,include.lowest=TRUE)
plot(calif$Longitude,calif$Latitude,col=grey(10:2/11)[cut.prices],pch=20,
  xlab="Longitude",ylab="Latitude")
partition.tree(treefit,ordvars=c("Longitude","Latitude"),add=TRUE)
```

Detailsemad otsused

• Mitmeks võib tipp jaguneda igal sammul – kas lubada vaid kaheks jagunemisi?

- Klassikaliselt kaheks – iga enam kui kaheks jagunemine on esitatav mitme üksteisele järgneva kaheks jagunemisena!

• Millise riskifaktori kohta „küsida“ järgmisena e. kuidas hinnata jagunemise headust?

- Klassifitseerimise puhul õigesti klassifitseeritud väärtuste osakaalu alusel.
- Väärtuste prognoosimise (regressioonipuu) puhul prognoosivea alusel.
- Võimalikud on erinevad teisendused toodud suurustest – mitmesugused nn mudeli sobivuse/mittesobivuse karakteristikud.

Detailsemad otsused

• Millal peatuda – kuulutada „tipp/sõlm“ „leheks“?

- Kui sõlme kuuluvate vaatluste arv on väiksem edasiseks jagamiseks mõttetuna tunduvast arvust (R-i funktsioonil ‘rpart’ vaikimisi 20).
- Kui uue jagunemise lisandumine ei suurenda puu (mudeli) prognoosivõimet piisaval määral (R-i funktsioonil ‘rpart’ klassifitseerimispuu korral 0,01 ehk 1%).

• Kuidas liiga suurt puud „tagasi lõigata“ (kärpida, ingl. ***prune***)?

• Kuidas toimida puuduvate andmete / müra esinemisel?

Klassifitseerimispuu näide: Jõgi jt, 2016

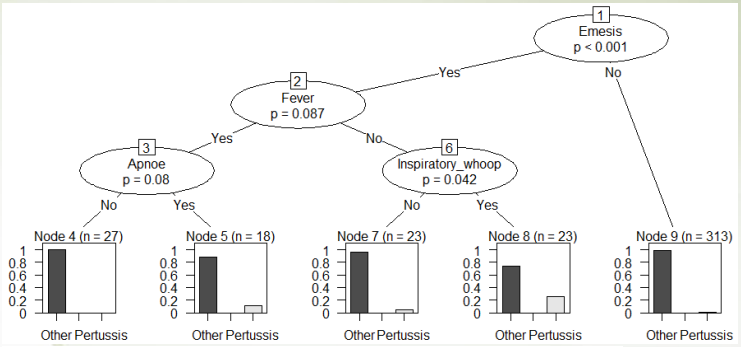


Figure 3: Classification tree of clinical characteristics as paroxysms, inspiratory whoop, posttussive emesis, apnoea and fever to predict the disease type on adult patients. The minimum splitting criteria was set as univariate p<0.1 and for each final node the distribution of patients according to their diagnosis is presented.

Regressioonipuu näide: Veromann jt, 2016

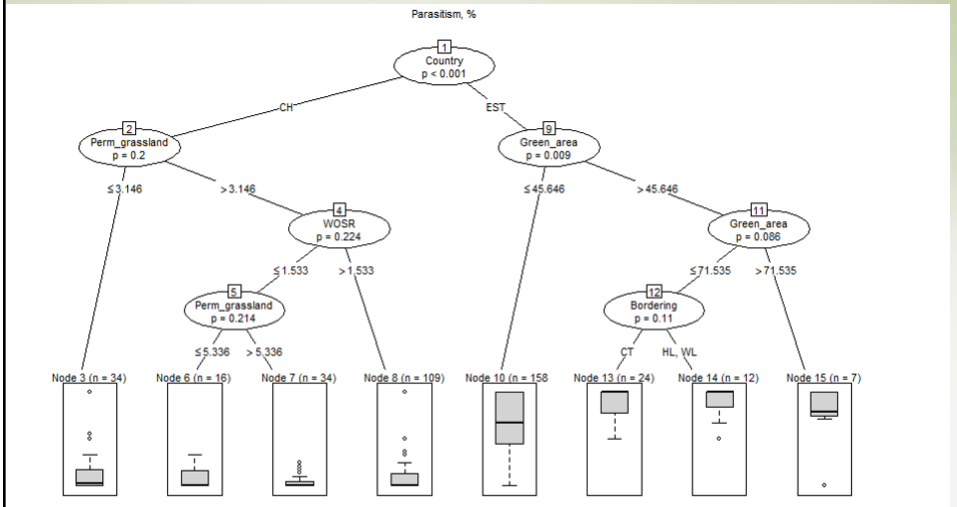


Figure 3. Classification tree of country (Estonia and Switzerland) and landscape parameters (percentage of permanent grassland, green area and agricultural land, WOSR, bordering type and distance from border) to predict the parasitism rate of larvae caught with funnels. To discover also less important but potentially interesting differences the minimum splitting criteria was set as univariate p<0.3. For each final node the box plot of parasitism rate is presented (the scale of y-axis is 0-100%).

Osavähimruutude regressioon- ja korrelatsioonanalüüs

(*Partial Least Square Regression and Corelation, Reduced Rank Regression, Principal Component Regression, …*)

Kanooniline korrelatsioonanalüüs

(canonical correlation *analysis*)

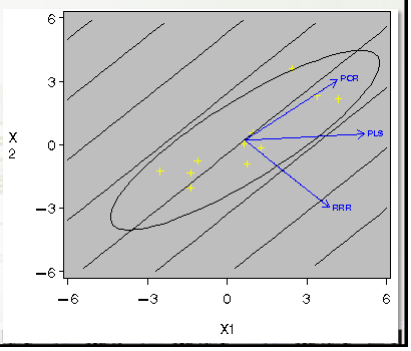
Koinertsusanalüüs

(*coinertia analysis*)

…

Erinevad lähenemised mitme tabeli koosanalüüsil

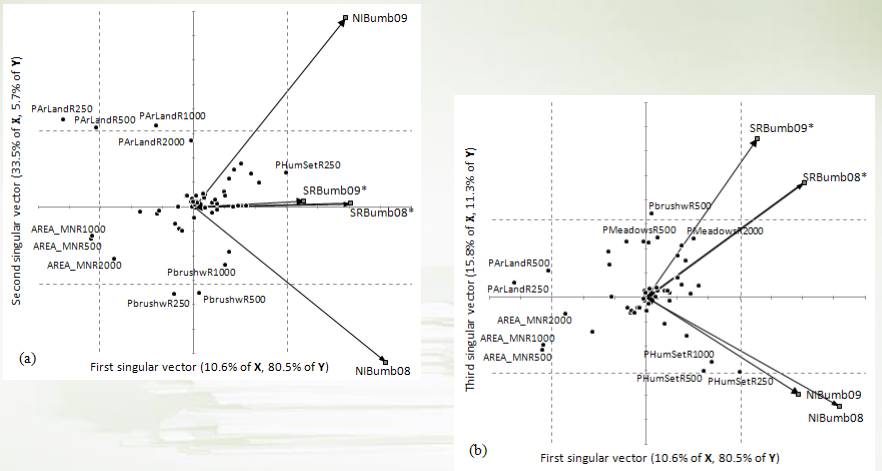
- Peakomponentide regressioon (*principal components regression*, *PCR*)
- Osavähimruutude korrelatsioon- ja regressioonanalüüs (*partial least squares correlation/regression*, *PLSC*), kanooniline korrelatsioonanalüüs (*canonical correlation*), koinertsusanalüüs (*co-inertia analysis*)
- *Reduced rank regression* (*RRR*)



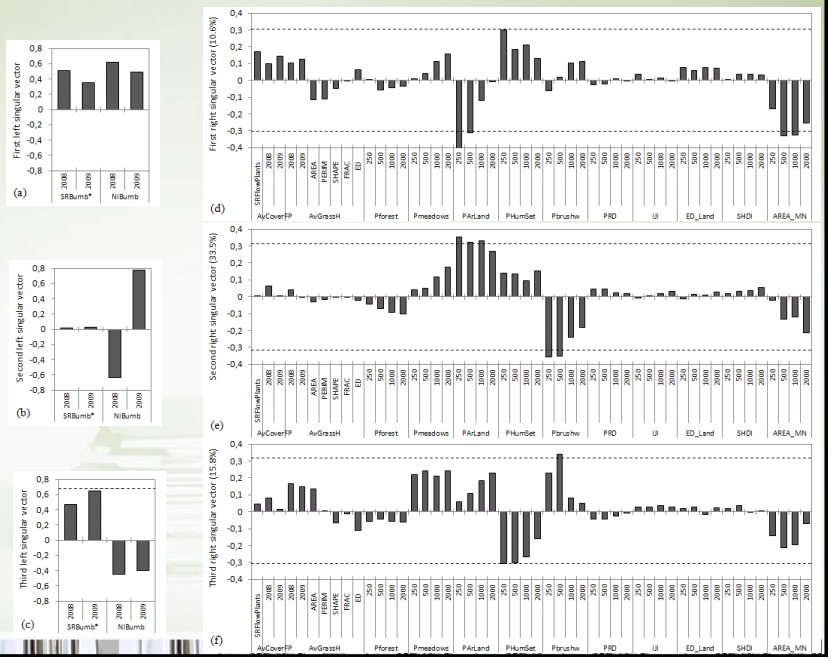
Joonis: SAS Online Doc

Osavähimruutude korrelatsioon (*partial least square correlation*, PLSC)

Näide: Isabel Diaz-Forero jt, 2012



PLSC näide: Isabel Diaz-Forero jt, 2012



PLSC näide: Harro jt, 2014

Contents lists available at ScienceDirect

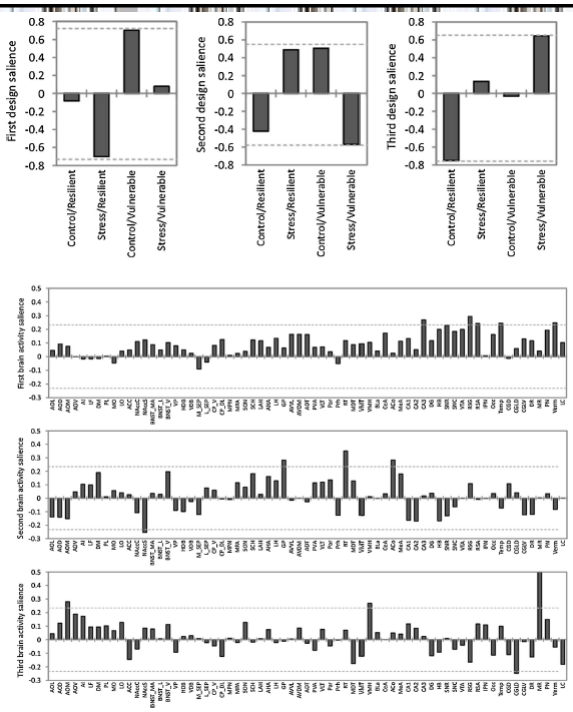
Behavioural Brain Research

journal homepage: www.elsevier.com/locate/bbr

Research report

Revealing the cerebral regions and networks mediating vulnerability to depression: Oxidative metabolism mapping of rat brain

Jaanus Harro[a,*], Margus Kanarik[a], Tanel Kaart[b], Denis Matrov[a], Kadri Kõiv[a], Tanel Mallo[a], Joaquin Del Río[c], Rosa M. Tordera[c], Maria J. Ramírez[c]



Koinertsusanalüüsi näide: Rannap jt, 2016

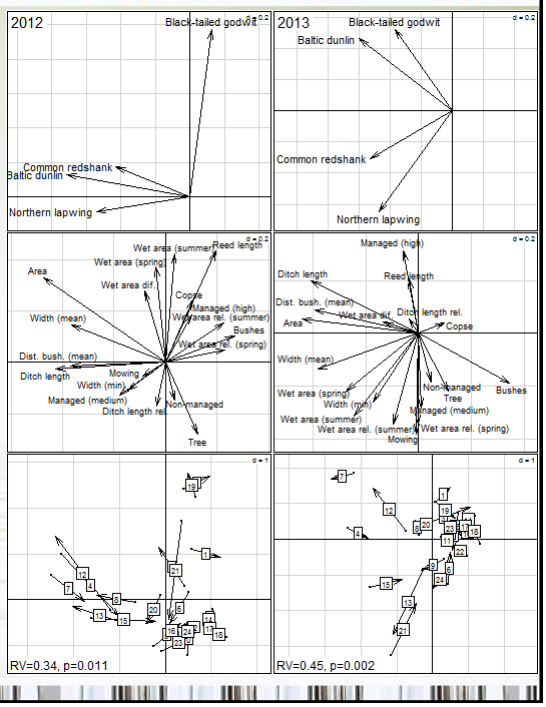


Figure 3. Results of co-inertia analysis of landscape parameters and breeding of four wader species in 2012 (left column of figures) and 2013 (right column of figures). The lengths and directions of arrows in first and second rows of figures denote the weights of initial variables concerning the first two factors (describing 51.3% and 31.6%, and 82.6% and 14.8% of overall covariance between landscape parameters and four wader species in 2012 and 2013, respectively). The last row of figures represents the concordance between landscape parameters and breeding of four wader species: the beginning and end of arrows denote the relative location of meadows according to their landscape parameters and number of breeding waders, respectively; the numerical measure of concordance is given by RV-coefficient, which p-value is evaluated by permutation test.