Multivariate Statistical Methods
(application of eigenvalue analysis)

Considered by Tõnu Möls

Journal Club, 26th May 2009

Topics:

- Multivariate data (matrix presentation, simulation, standardization, second-order moments)
- Eigenvalue analysis
- Principal Component Analysis (Example)
- Canonical Correlation Analysis(3 Examples)

Multivariate Data

Data vector: $X = (x_1, x_2, \ldots, x_p)$

Observations:

$$X_1 = (x_{11}, x_{12}, \ldots, x_{1p})$$
$$X_2 = (x_{21}, x_{22}, \ldots, x_{2p})$$
$$\cdots\cdots\cdots\cdots\cdots$$
$$X_n = (x_{n1}, x_{n2}, \ldots, x_{np})$$

Observations:

$$X_1 = (x_{11}, x_{12}, \ldots, x_{1p})$$
$$X_2 = (x_{21}, x_{22}, \ldots, x_{2p})$$
$$\cdots\cdots\cdots\cdots$$
$$X_n = (x_{n1}, x_{n2}, \ldots, x_{np})$$

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Generation of random simulated data

```
Data U;
        do i=1 to 10;
                x=normal(1000022);
                y=0.5*x+normal(0);
                u=0.5*y+normal(0);
                v=0.5*u+normal(0);
                output;
        end;
run;
```

Data, generated with random number progamm

i	x	y	u	v
1	0.88510	-0.83597	-0.67749	-2.14817
2	2.04206	1.99571	3.08504	1.11369
3	0.25260	-1.32018	0.15191	1.29817
4	-0.65648	-1.57101	-1.54206	-1.24360
5	-2.23955	-2.26853	-3.47058	-0.92672
6	0.69201	-0.62193	0.62507	0.33981
7	1.67152	0.17235	-0.97509	-1.26482
8	0.16825	0.42110	0.81163	-0.18010
9	1.29836	-0.97789	1.89865	1.62785
10	-0.10090	-0.84233	-2.18998	-1.41218

Standardization

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
i	10	5.5000000	3.0276504	1.0000000	10.0000000
x	10	0.4012983	1.2405408	-2.2395533	2.0420578
y	10	-0.5848662	1.1961874	-2.2685271	1.9957091
u	10	-0.2282901	1.9522768	-3.4705840	3.0850429
v	10	-0.2796067	1.3146844	-2.1481750	1.6278512

```
proc standard data=U mean=0 std=1 out=US;
var x y u v;
run;
```

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
i	10	5.5000000	3.0276504	1.0000000	10.0000000
x	10	-4.44089E-17	1.0000000	-2.1287905	1.3226163
y	10	-6.10623E-17	1.0000000	-1.4075227	2.1573335
u	10	0	1.0000000	-1.6607758	1.6971636
v	10	1.110223E-17	1.0000000	-1.4213055	1.4508866

Principal Component Analysis

SAS: proc Princomp

Example 1

Collapse of the vendace *Coregonus albula* (L.)

population in Lake Peipsi: the result of extreme

weather events, climate change and

Predator-prey interactions

K. KANGUR, T. MÖLS, A. KANGUR and P. KANGUR

Foto Külli Kanguri kogust

SAS program calculating principal components for fish data

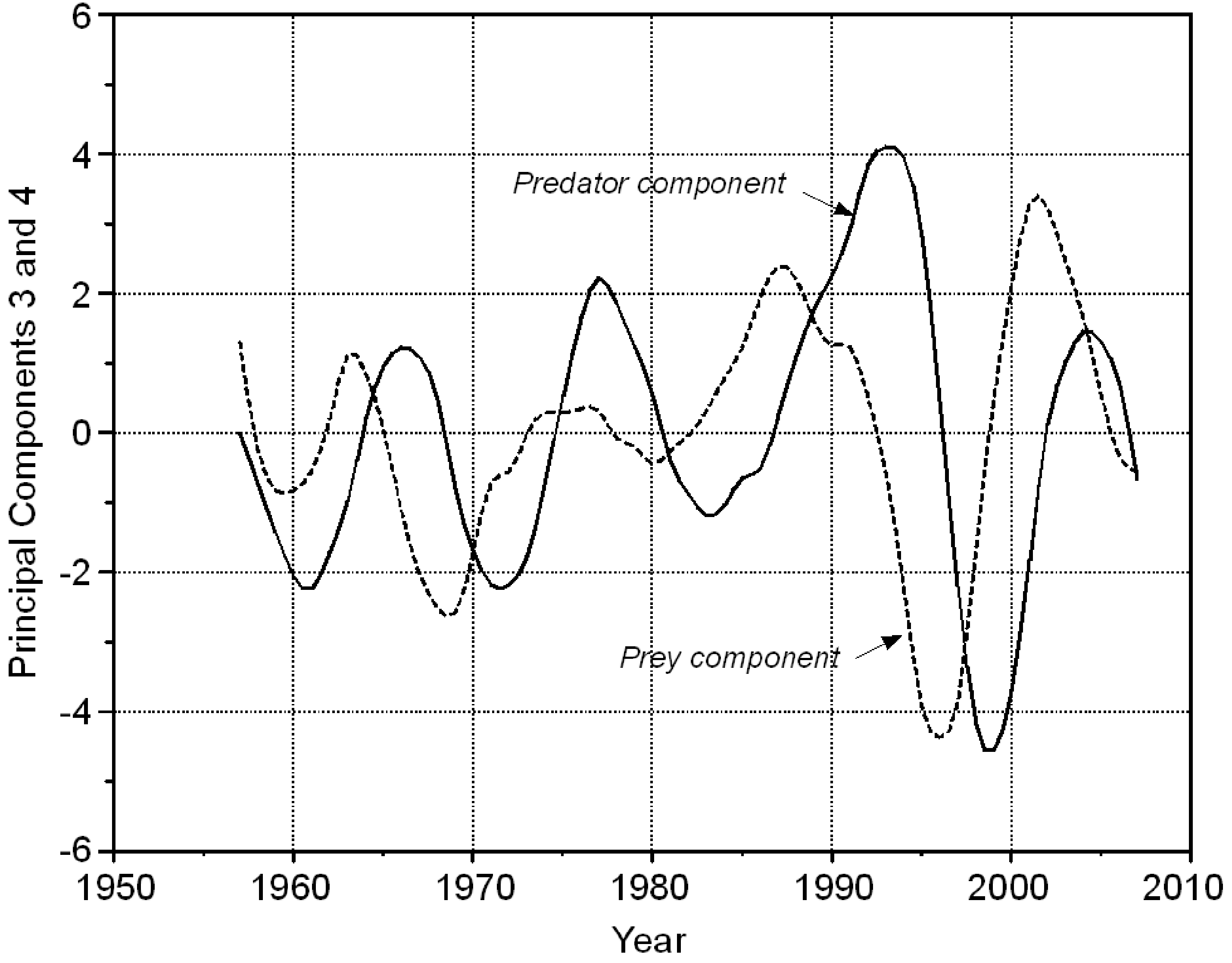
```
proc princomp data=Lkoos123n out=PLkoos123n;
   var
       pikeperch_0  pikeperch_1 pikeperch_2 pikeperch_3
            pikeperch_4 pikeperch_5 pikeperch_6
       bream_0  bream_1 bream_2 bream_3 bream_4 bream_5 bream_6
       burbot_0  burbot_1 burbot_2 burbot_3 burbot_4 burbot_5 burbot_6
       perch_0  perch_1 perch_2 perch_3 perch_4 perch_5 perch_6
       pike_0  pike_1 pike_2 pike_3 pike_4 pike_5 pike_6
       vendace_0  vendace_1 vendace_2 vendace_3 vendace_4
            vendace_5 vendace_6;
run;
```

Ten of 42 eigenvalues of the (42 x 42) correlation matrix of 6 fish species catches in current and 6 previous years.

	Eigenvalue	Difference	Proportion	Cumulative
1	21.1812474	15.5180738	0.5043	0.5043
2	5.6631736	2.5170227	0.1348	0.6392
3	3.1461509	0.7862884	0.0749	0.7141
4	2.3598624	0.7700933	0.0562	0.7702
5	1.5897692	0.4517293	0.0379	0.8081
6	1.1380399	0.1394331	0.0271	0.8352
7	0.9986068	0.2142200	0.0238	0.8590
8	0.7843867	0.1784644	0.0187	0.8776
9	0.6059223	0.0591546	0.0144	0.8921
10	0.5467677	0.0485481	0.0130	0.9051

Species	Prin3	Prin4	Species	Prin3	Prin4
Pikeperch_0	0.108417	0.087986	Perch_0	-.179977	-.040910
Pikeperch_1	0.090337	0.058202	Perch_1	-.114594	0.034167
Pikeperch_2	0.069006	0.028900	Perch_2	0.007264	0.090059
Pikeperch_3	0.040602	0.000671	Perch_3	0.146586	0.124461
Pikeperch_4	0.002551	-.023638	Perch_4	0.259620	0.119243
Pikeperch_5	-.043884	-.038656	Perch_5	0.321231	0.071001
Pikeperch_6	-.096278	-.029965	Perch_6	0.302627	-.011444
Bream_0	0.121610	0.305177	Pike_0	-.193707	0.334899
Bream_1	0.180748	0.286672	Pike_1	-.058405	0.344676
Bream_2	0.194272	0.197642	Pike_2	0.070288	0.238587
Bream_3	0.157495	0.103656	Pike_3	0.166339	0.101327
Bream_4	0.076187	0.017830	Pike_4	0.215130	-.074432
Bream_5	-.024183	-.043210	Pike_5	0.213066	-.242291
Bream_6	-.123155	-.086306	Pike_6	0.158037	-.360671
Burbot_0	-.126550	0.037999	Vendace_0	-.202050	-.008678
Burbot_1	-.090322	0.062272	Vendace_1	-.168405	0.119554
Burbot_2	-.038767	0.044690	Vendace_2	-.080035	0.207888
Burbot_3	0.013636	-.000766	Vendace_3	0.048480	0.229611
Burbot_4	0.064108	-.057969	Vendace_4	0.179957	0.162305
Burbot_5	0.089159	-.124169	Vendace_5	0.272195	0.017272
Burbot_6	0.086572	-.159150	Vendace_6	0.285737	-.153765

Lotka-Volterra-type dependence between the two Principal Components in Lake Peipsi



Canonical Correlation Analysis

Canonical correlation is a generalization of simple correlation for analysing the relationship between the two sets of variables

Explanatory variables
(p variables)

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Dependent variables
(q variables)

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & x_{nq} \end{pmatrix}$$

In canonical correlation, you examine the relationship between linear combinations of the set of **X** variables and linear combinations of a *set* of **Y** variables.

These linear combinations are called *canonical variables* or *canonical variates*.

A linear combination of p variables $x_1, \ldots, x_p$ looks like this:

$$a_1 x_1 + a_2 x_2 + \ldots + a_p x_p$$

Independent

Independent

$$w_1 = \underbrace{a_{11}x_1 + a_{21}x_2 + \dots + a_{p1}x_p}$$

Maximal Pearson correlation

$$v_1 = \overbrace{b_{11}y_1 + b_{21}y_2 + \dots + b_{q1}y_q}$$

$$w_2 = \underbrace{a_{12}x_2 + a_{22}x_2 + \dots + a_{p2}x_p}$$

Maximal Pearson correlation

$$v_2 = \overbrace{b_{12}y_2 + b_{22}y_2 + \dots + b_{q2}y_q}$$

The eigenvalues of $\mathbf{S}_{YY}^{-1}\mathbf{S}_{XY}'\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}$ are the squared canonical correlations, the right eigenvectors are raw Canonical coefficients for the **Y** variables:

Eigenvalue *Coefficients*

$$\mathbf{S}_{YY}^{-1}\mathbf{S}_{XY}^{T}\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}\, v = \lambda\, v$$

(qxq)(qxp)(pxp)(pxq)(qx1)

The eigenvalues of $\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-1}\mathbf{S}_{XY}'$ are the squared canonical correlations, the right eigenvectors are raw canonical coefficients for the **X** variables:

$$\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-1}\mathbf{S}_{XY}^{T}\, v = \lambda\, v$$

Either set of variables can be considered explanatory or response variables, since the statistical model is symmetric in the two sets of variables.

Simple and multiple correlation are special cases of canonical correlation in which one or both sets contain a single variable.

Example: modelled data

Data, generated with random number program

i	x	y	u	v
1	0.88510	-0.83597	-0.67749	-2.14817
2	2.04206	1.99571	3.08504	1.11369
3	0.25260	-1.32018	0.15191	1.29817
4	-0.65648	-1.57101	-1.54206	-1.24360
5	-2.23955	-2.26853	-3.47058	-0.92672
6	0.69201	-0.62193	0.62507	0.33981
7	1.67152	0.17235	-0.97509	-1.26482
8	0.16825	0.42110	0.81163	-0.18010
9	1.29836	-0.97789	1.89865	1.62785
10	-0.10090	-0.84233	-2.18998	-1.41218

Calculating Canonical Correlations with SAS

```
proc cancorr data= Us out = Vs
       vprefix = Plant
       wprefix = Geo;
   var x y;
   with  u v;
   ods output RawCanCoefV=ccv;
   ods output RawCanCoefW=ccw;
run;
```

The CANCORR Procedure

Canonical Correlation Analysis

	Canonical Correlation	Standard Error
1	0.896235	0.065588
2	0.100402	0.329973

The CANCORR Procedure
Canonical Correlation Analysis

Standardized Canonical Coefficients
for the VAR Variables

	Plant1	Plant2
x	0.6026	1.4282
y	0.4609	-1.4800

Standardized Canonical Coefficients
for the WITH Variables

	Geo1	Geo2
u	1.3783	-0.5499
v	-0.6478	1.3351

Multivariate Statistics and F Approximations

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.19477982	3.80	4	12	0.0322
Pillai's Trace	0.81331718	2.40	4	14	0.0996
Hotelling-Lawley Trace	4.09243210	5.99	4	6.3077	0.0248
Roy's Greatest Root	4.08224899	14.29	2	7	0.0034

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

Example

Global gradients in moss (*Bryopsida*) and vascular plant diversity
Nele Ingerpuu, Ain Vellak, Kai Vellak and Tõnu Möls

Used data

Latrange	Precrange	Elev.Range	LArea	LPopulation	QCoastline	LMosses	LVasculars
11	400	3595	19.33	21.64	0.0	8.99	10.68
5	150	256	17.66	23.29	0.0	8.39	11.03
12	2950	6452	20.06	23.09	0.0	10.19	14.08
11	1930	4042	19.84	21.97	164.3	9.22	11.44
3	750	2925	16.75	22.80	18.8	9.05	11.80
10	3100	4095	18.85	24.04	20.0	8.49	13.01
8	1400	1085	19.25	22.03	0.0	8.06	11.81
16	7300	5775	20.11	25.37	56.6	9.89	15.64
2	700	1487	16.26	23.28	0.0	9.32	10.89
2	250	329	15.34	22.37	21.2	8.70	10.25
6	2600	6267	18.11	23.69	47.2	9.89	14.24
2	200	317	15.46	20.33	37.3	8.74	10.67
10	300	1328	18.36	22.31	67.8	9.36	10.10
1	1800	1575	18.03	20.44	29.7	7.86	12.69
8	1400	2966	18.44	26.29	48.8	9.69	11.38
.	. . .	. . .	. . .	. . .	. . .		
15	750	3144	18.33	26.33	58.6	9.21	13.35
11	875	2430	18.57	23.54	0.0	8.04	12.11

Canonical Correlation Analysis

	Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.870307	0.034652	0.757434
2	0.354421	0.124912	0.125614

Pearson Correlation Coefficients, N = 50

	Species Number Index (SNI)	Moss Prevalence Index (MPI)
Area Size Index	0.87031	0.00000
(p-value)	<0.0001	1.0000
Env. Div. Index	0.00000	0.35442
(p-value)	1.0000	0.0116

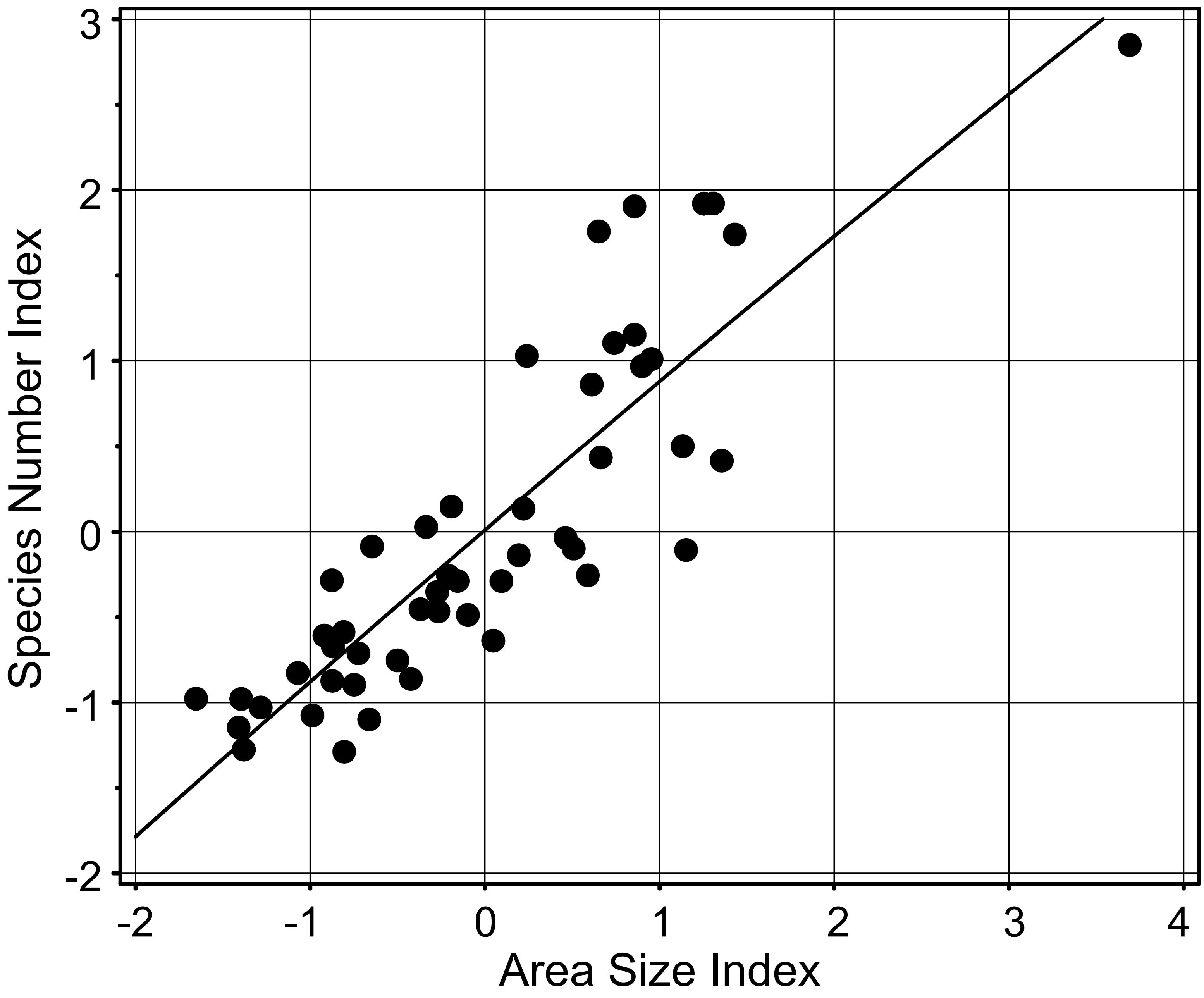
The CANCORR Procedure
Canonical Correlation Analysis

Standardized Canonical Coefficients for the species richness

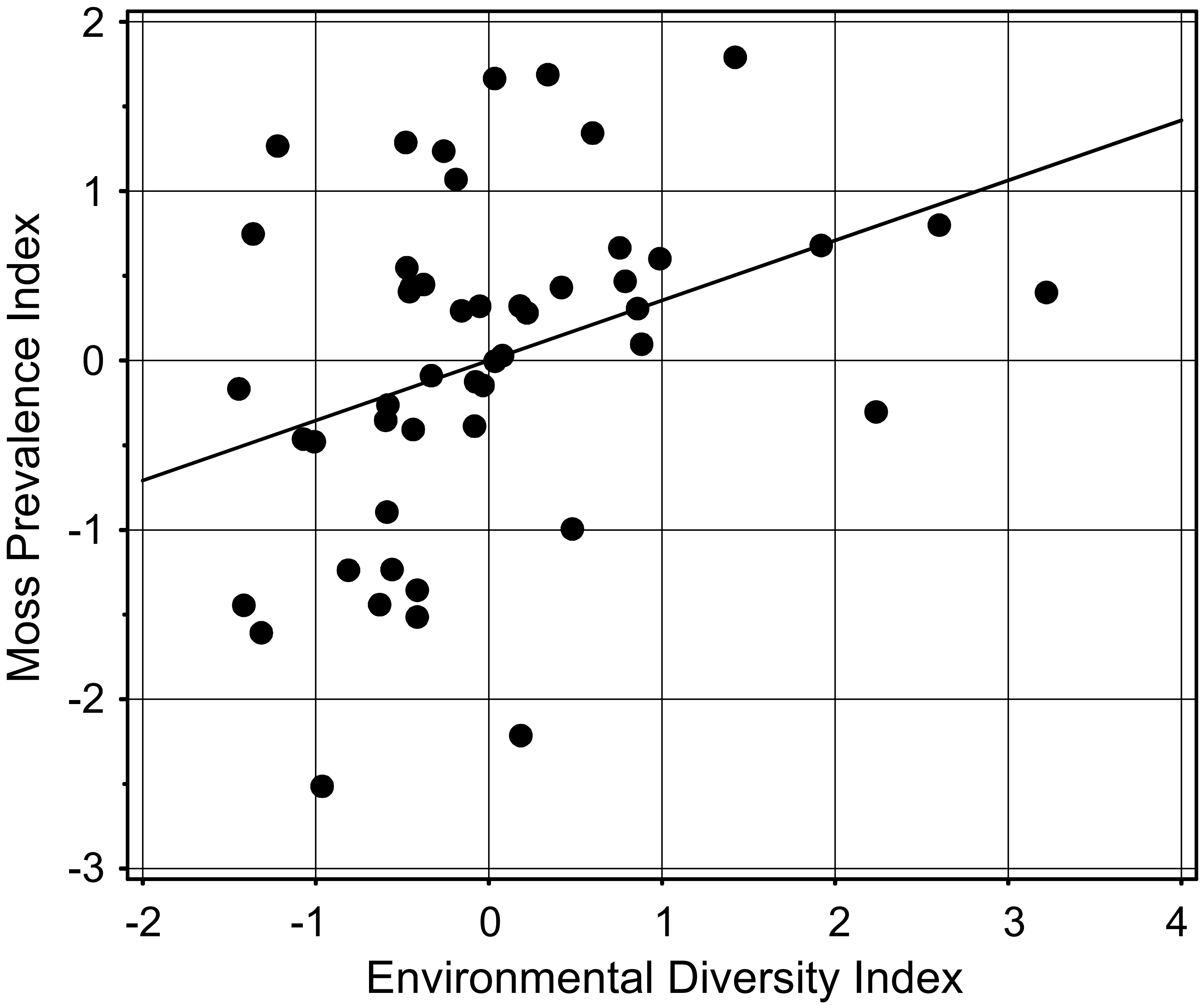
	SNI	MPI
LMosses	0.2441	1.0513
LVasculars	0.8823	-0.6216

Standardized Canonical Coefficients for the geographical conditions

	ASI	EDI
Larea	0.2819	-0.4589
QCoastline	0.0321	1.0228
Latrange	0.0211	-0.2296
Lpopulatio	0.3006	-0.0963
Elev.range	0.1298	0.2157
Precrange	0.5848	0.0796



Species Number Index
3
2
1
0
-1
-2
-2
-1
0
1
2
3
4
Area Size Index



Moss Prevalence Index
2
1
0
-1
-2
-3
-2
-1
0
1
2
3
4
Environmental Diversity Index

Example

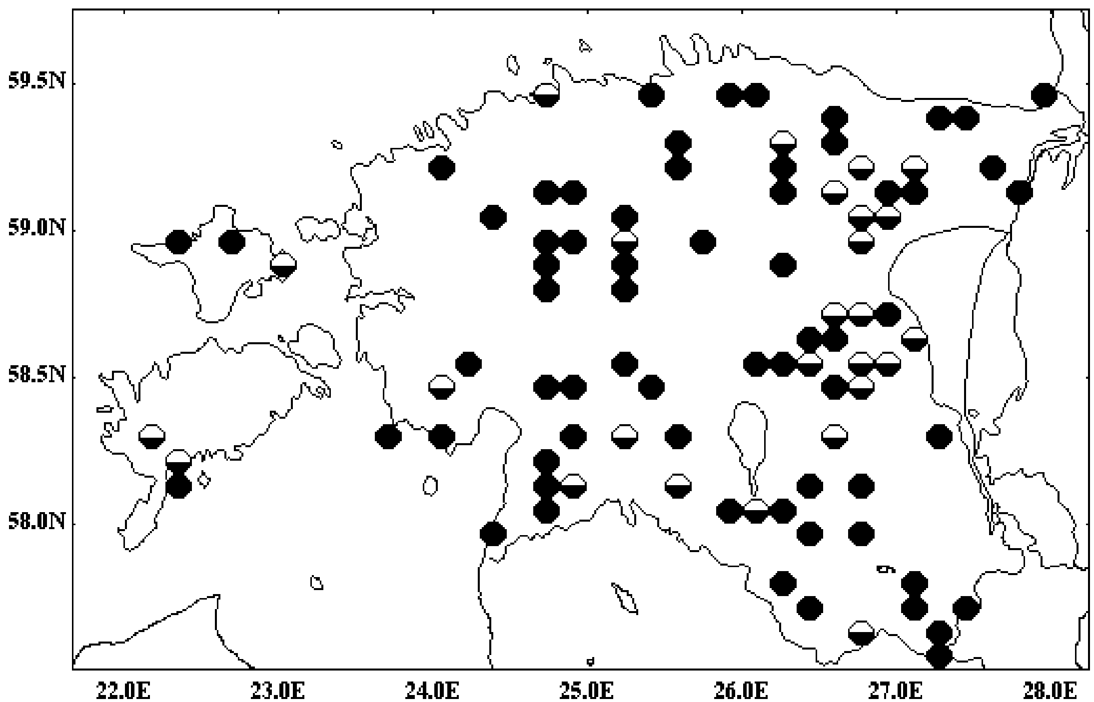
(One dependent variable only)

?

Shoot Increment =

lin(SUVENIISK, POSNIIS, SUVEtemp, POStemp, POSSADE, SUVESAD)

Example 1: Nele Ingerpuu, Kai Vellak, Tõnu Möls. Growth of *Neckera pennata*, an epiphytic moss of old-growth forests–The Bryologist, 110 (2), 309-318.



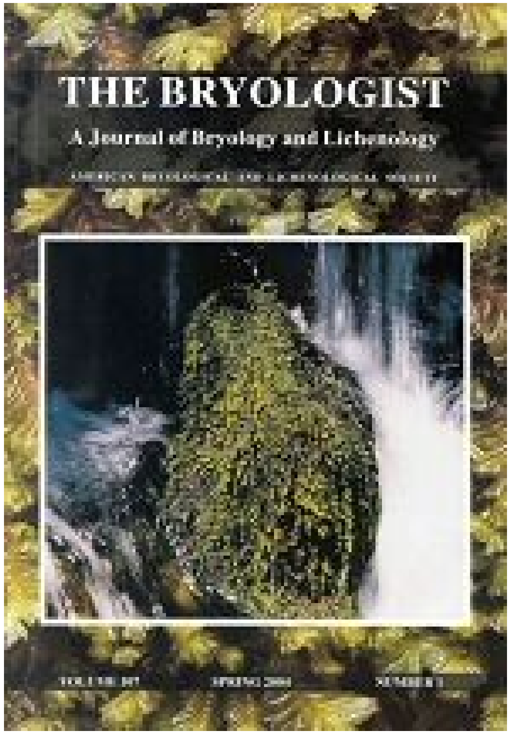


Figure 1. Distribution map of *Neckera pennata* in Estonia. Dots are marking the centres of the UTM-grid.

Neckera pennata -- sulgjas õhik,

ohustatud samblaliik, Eestis veel suhteliselt tavaline. Millised elupaigad valida liigi säilitusbaasiks?


© Michael Lüth

WEATHER INDEX = SUVENIISK*0.139 - POSNIIS*0.0467 + SUVEtemp*0.2509 - POStemp*0.08951 + POSSADE*0.00546 - SUVESAD*0.00424

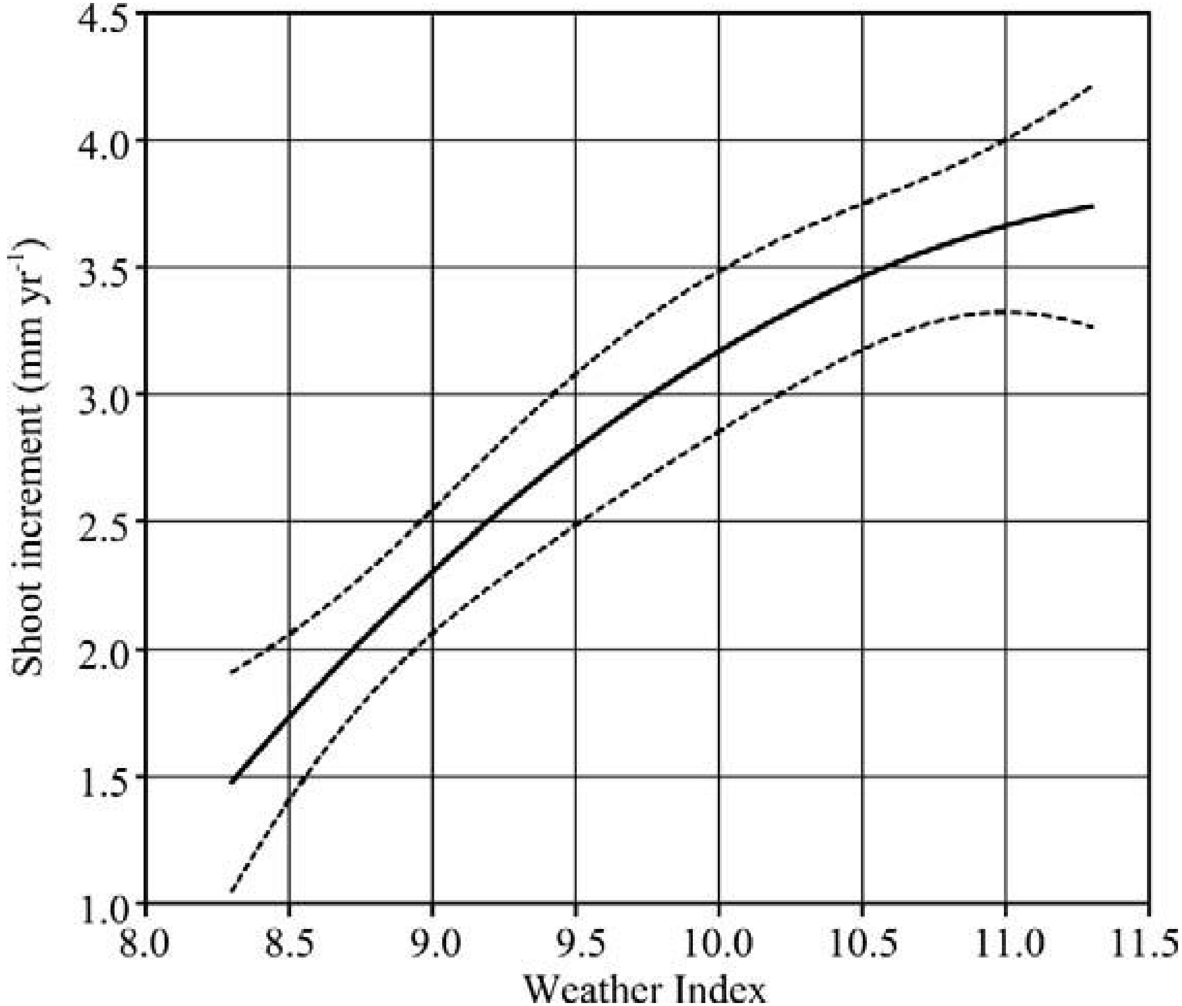


Figure 3. Shoot increment, predicted on the basis of Weather Index value. Dashed lines show the 95% confidence limits for the mean actual shoot increment. N = 480. Dependence between the predicted and observed increment is significant at $p<0.0001$.

The End

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}.$$