UNIVERZITA PARDUBICE

Fakulta chemicko-technologická
Katedra analytické chemie

Licenční studium chemometrie na téma

Statistické zpracování dat

Semestrální práce z 5. soustředění

Předmět: 3.5 Klasifikace analýzou vícerozměrných dat

Vedoucí licenčního studia: Prof. RNDr. Milan Meloun, DrSc.

Vypracoval: Ing. Jiří Souček, Ph.D.

Licenční studium Statistické zpracování experimentálních dat
Předmět 3.5 Klasifikace analýzou vícerozměrných dat
Přednášející: Prof. RNDr. Milan Meloun, DrSc.
Úloha: Korespondenční analýza (CA)

Zadání:
Na pokusné ploše bylo opakovaně po 10 a 20 letech analyzováno postavení stromů v rámci porostu (nadúrovňový, úrovňový a podúrovňový strom). V průběhu let se postavení jednotlivých stromů může vzájemně měnit z důvodu hospodaření a odlišného výškového růstu. Většinou dochází k negativním přesunům, četnost pozitivních přesunů je omezená, ale jsou možné.

Data:
Přesuny stromů v prvním sledovaném období (1974 – 1984)

		1984		
		Předrůstavý	Úrovňový	Podúroveň
1974	Předrůstavý	9	34	6
	Úrovňový	0	17	27
	Podúroveň	0	1	1

Přesuny stromů v druhém sledovaném období (1984 – 2004)

		2004		
		Předrůstavý	Úrovňový	Podúroveň
1984	Předrůstavý	7	11	1
	Úrovňový	7	29	14
	Podúroveň	1	4	1

Řešení pro první období:

Řádkové profily v procentech:

	Předrůstavý	Úrovňový	Podúroveň	Total
Předrůstavý	18,37	69,39	12,24	100
Úrovňový	0	38,64	61,36	100
Podúroveň	0	50,00	50,00	100
Total	9,47	54,74	35,79	100

Sloupcové profily v procentech:

	Předrůstavý	Úrovňový	Podúroveň	Total
Předrůstavý	100,00	65,38	17,65	51,58
Úrovňový	0	32,69	79,41	46,32
Podúroveň	0	1,92	2,94	2,11
Total	100	100	100	100

Hledání počtu projekčních dimenzí

Factor No.	**Eigenvalue**	**Individual Percent**	**Cumulative Percent**	**Bar Chart**
1	0,296505	99,87	99,87	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
2	0,000398	0,13	100,00	|
Total	0,296903			

Již první proměnná pokrývá 99,87 % celkové informace, součet obou dimenzí pokrývá 100 % informací.

Zobrazení řádkového profilu a příspěvek do inercie

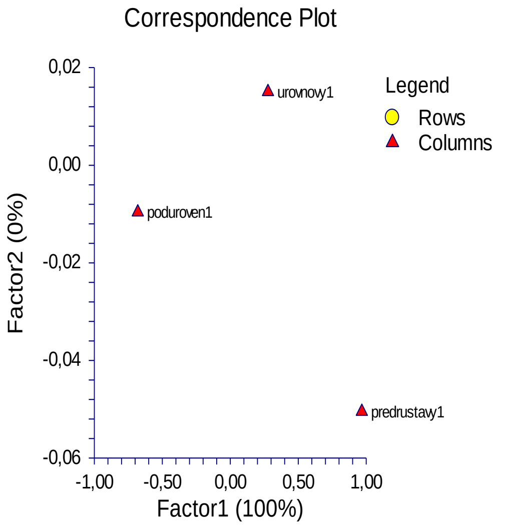
				--------	**Axis1**	--------	--------	**Axis2**	--------
Name	**Quality**	**Mass**	**Inertia**	**Factor**	**COR**	**CTR**	**Factor**	**COR**	**CTR**
predrustavy	1	0,516	0,482	0,527	1	0,483	-0,001	0	0,001
urovnovy	1	0,463	0,507	-0,57	1	0,507	-0,005	0	0,029
poduroven	1	0,021	0,011	-0,37	0,882	0,01	0,135	0,118	0,969

Principal Coordinate Section for Rows - Axis 1

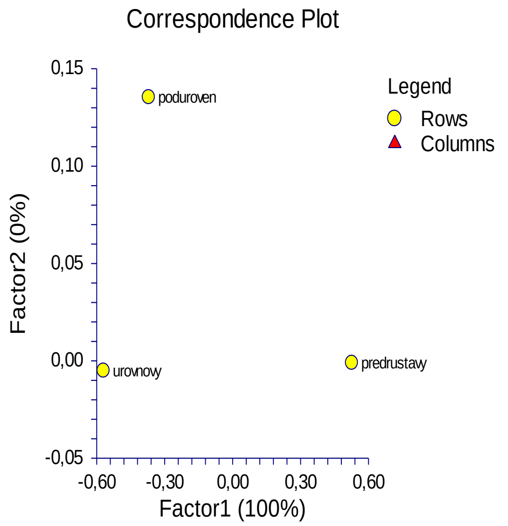
Name	Mass	Inertia	Distance	Factor	COR	CTR	Angle	Eigenvalue
1 predrustavy	0,516	0,482	0,278	0,527	1	0,483	0,1	0,143181
2 urovnovy	0,463	0,507	0,325	-0,57	1	0,507	0,5	0,150441
3 poduroven	0,021	0,011	0,155	-0,37	0,882	0,01	20,1	0,002883

Principal Coordinate Section for Rows - Axis 2

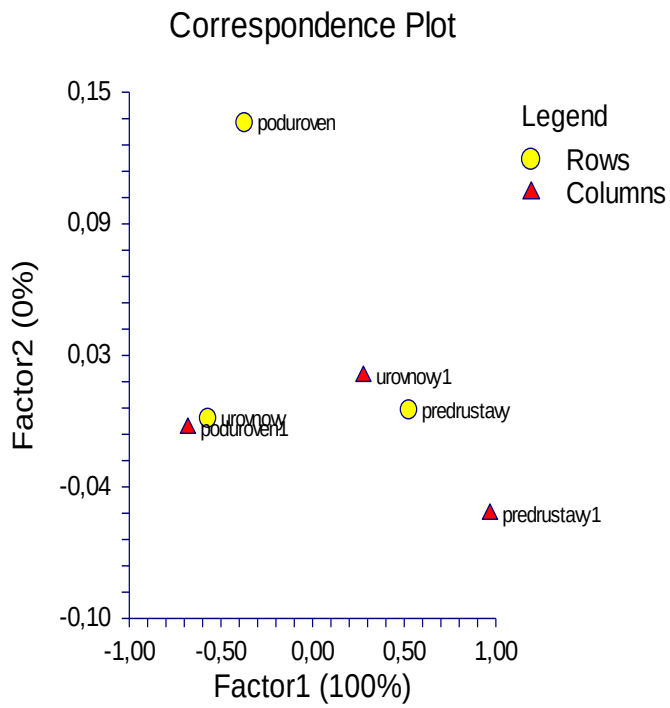
Name	Mass	Inertia	Distance	Factor	COR	CTR	Angle	Eigenvalue
1 predrustavy	0,516	0,482	0,278	-0,001	0	0,001	89,9	0,000001
2 urovnovy	0,463	0,507	0,325	-0,005	0	0,029	89,5	0,000012
3 poduroven	0,021	0,011	0,155	0,135	0,118	0,969	69,9	0,000386



Graf sloupcových profilů



Graf řádkových profilů



Spojený graf řádkových a sloupcových profilů

Závěr:

Graf řádkových profilů zobrazuje roztřídění kategorií stromů podle jejich výchozího postavení v porostu, graf sloupcových profilů podle postavení po 10 letech. Spojený graf sloupcových a řádkových profilů naznačuje přesun většiny původně předrůstavých stromů do stromů úrovňových a původně úrovňových do podúrovňových. Stromy podúrovňové na počátku sledování a stromy předrůstavé v roce 1984 se od ostatních skupin výrazně liší.

Řešení pro druhé období (1984 – 2004)

Řádkové profily v procentech

	Předrůstavý	Úrovňový	Podúroveň	Total
Předrůstavý	36,84	57,89	5,26	100
Úrovňový	14,00	58,00	28,00	100
Podúroveň	16,67	66,67	16,67	100
Total	20,00	58,67	21,33	100

Sloupcové profily v procentech:

	Předrůstavý	Úrovňový	Podúroveň	Total
Předrůstavý	46,67	25,00	6,25	25,33
Úrovňový	46,67	65,91	87,50	66,67
Podúroveň	6,67	9,09	6,25	8,00
Total	100	100	100	100

Factor No.	**Eigenvalue**	**Individual Percent**	**Cumulative Percent**	**Bar Chart**
1	0,092400	97,58	97,58	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
2	0,002296	2,42	100,00	|II
Total	0,094696			

První dimenze pokrývá 97,58 % celkové informace, součet obou dimenzí pokrývá 100 % informací.

Zobrazení řádkového profilu a příspěvek do inercie

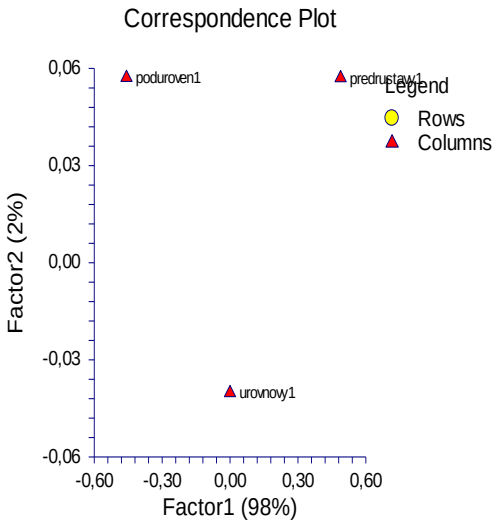
				--------	**Axis1**	--------	--------	**Axis2**	--------
Name	**Quality**	**Mass**	**Inertia**	**Factor**	**COR**	**CTR**	**Factor**	**COR**	**CTR**
predrustavy	1	0,253	0,704	0,513	0,999	0,720	0,015	0,001	0,026
urovnovy	1	0,667	0,274	-0,197	0,995	0,279	0,014	0,005	0,054
poduroven	1	0,080	0,023	0,017	0,010	0,000	-0,162	0,990	0,920

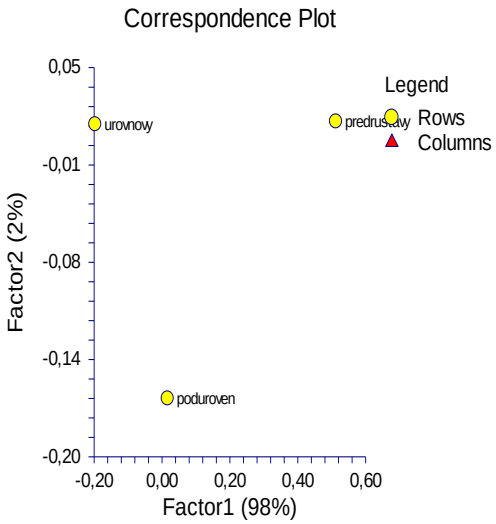
Principal Coordinate Section for Rows - Axis 1

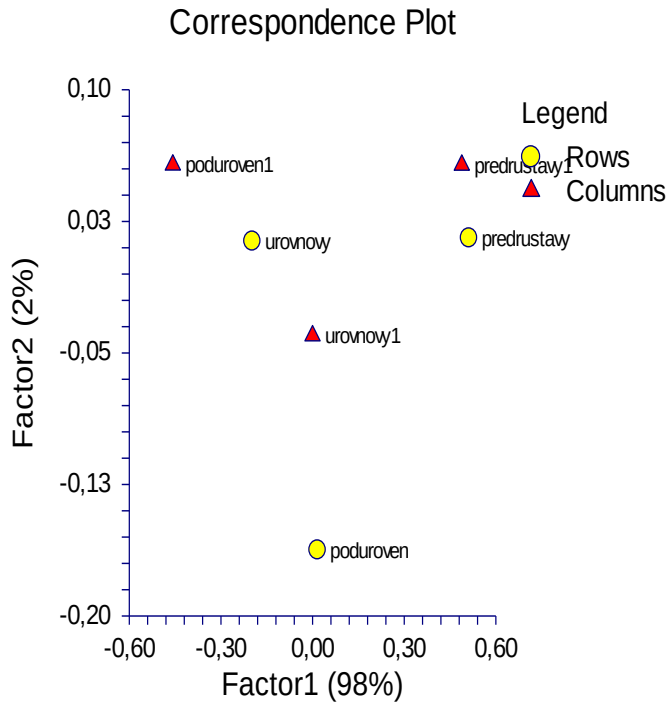
Name	Mass	Inertia	Distance	Factor	COR	CTR	Angle	Eigenvalue
1 predrustavy	0,253	0,704	0,263	0,513	0,999	0,720	1,7	0,066562
2 urovnovy	0,667	0,274	0,039	-0,197	0,995	0,279	4,0	0,025816
3 poduroven	0,08	0,023	0,027	0,017	0,010	0,000	84,1	0,000022

Principal Coordinate Section for Rows - Axis 2

Name	Mass	Inertia	Distance	Factor	COR	CTR	Angle	Eigenvalue
1 predrustavy	0,253	0,704	0,263	0,015	0,001	0,026	88,3	0,000060
2 urovnovy	0,667	0,274	0,039	0,014	0,005	0,054	86,0	0,000124
3 poduroven	0,080	0,023	0,027	-0,162	0,990	0,920	5,9	0,002112







Závěr:

Grafy řádkových a sloupcových profilů mají odlišné rozložení bodů ve srovnání s grafy z předchozího období. Spojený graf sloupcových a řádkových profilů naznačuje již poměrnou stálost předrůstavých stromů a možnost přesunů původně úrovňových stromů do obou sousedních tříd. Pravděpodobnost přesunu původně podúrovňových stromů do vyšších tříd je omezená.

Licenční studium Statistické zpracování experimentálních dat
Předmět 3.5 Klasifikace analýzou vícerozměrných dat
Přednášející: Prof. RNDr. Milan Meloun, DrSc.
Úloha: mapování objektů vícerozměrným škálováním (MDS)

Data
V Krkonoších byly odebrány vzorky jehličí kleče na 18 lokalitách pro stanovení výživy těchto porostů. V jehličí byly analyzovány základní prvky, N, P, K, Ca, Mg a dále jejich vzájemné poměry.

Řešení:
Indexový graf úpatí vlastních čísel

číslo	**Vl. číslo**	**Individ. %**	**Kumulat. %**	**Čárový diagram**
1	0,12	70,00	70,00	|IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
2 (použito)	0,03	18,37	88,37	|IIIIIIIII
3	0,01	8,26	96,63	|IIII
4	0,01	2,92	99,54	|I
5	0,00	0,26	99,81	|
6	0,00	0,13	99,93	|
7	0,00	0,05	99,98	|
8	0,00	0,01	99,99	|
9	0,00	0,01	100,00	|
10	0,00	0,00	100,00	|
11	0,00	0,00	100,00	|
12	0,00	0,00	100,00	|
13	0,00	0,00	100,00	|
14	0,00	0,00	100,00	|
15	0,00	0,00	100,00	|
16	0,00	0,00	100,00	|
17	0,00	0,00	100,00	|
18	0,00	0,00	100,00	|
Celkem	0,18			

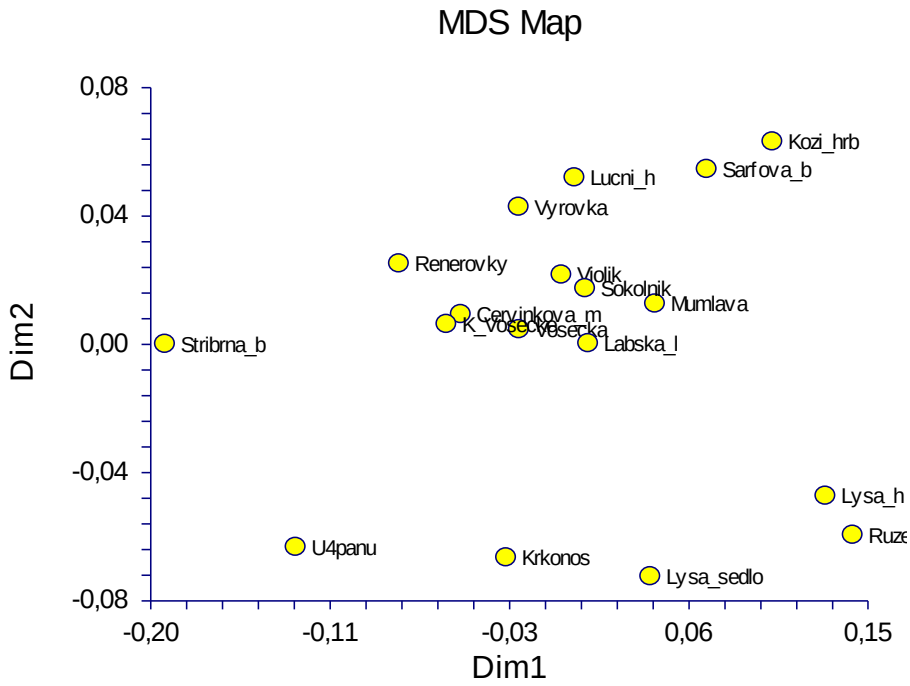
Index	**Čtverec rozdílů**	**Stress**	**Pseudo R^2**
1	0,275311	0,293336	52,71
2	0,062183	0,139409	89,32
3	0,004874	0,039030	99,16
4	0,000037	0,003397	99,99

Hodnoty stresu a Pseudo R^2 naznačují, že čtyřrozměrná škálovací mapa by byla nejvhodnější (minimální hodnota stresu, maximální Pseudo R^2), i použití pouze 2 rozměrné mapy však vysvětluje 88 % variability. Pro další hodnocení bude použita pouze 2 rozměrná škálovací mapa.

Solution Section

Znak	**Dim1**	**Dim2**	**Dim3**	**Dim4**
Renerovky	-0,0787	0,0251	0,0098	0,0003
Lucni_h	0,0070	0,0519	0,0286	-0,0123
Kozi_hrb	0,1036	0,0631	0,0000	0,0515
Sarfova_b	0,0715	0,0545	-0,0019	-0,0153

Cervinkova_m	-0,0485	0,0093	-0,0695	-0,0175
Stribrna_b	-0,1928	0,0000	0,0032	0,0060
Violik	0,0006	0,0216	0,0309	-0,0147
Vyrovka	-0,0202	0,0427	-0,0306	-0,0199
Krkonos	-0,0264	-0,0666	0,0094	-0,0039
Mumlava	0,0463	0,0126	-0,0159	0,0121
Vosecka	-0,0201	0,0046	0,0276	-0,0049
Sokolnik	0,0122	0,0174	-0,0349	0,0132
Labska_l	0,0137	0,0002	0,0279	-0,0046
Ruzencina_z	0,1428	-0,0595	-0,0413	-0,0102
Lysa_h	0,1295	-0,0473	0,0425	-0,0090
Lysa_sedlo	0,0440	-0,0725	-0,0020	0,0088
U4panu	-0,1290	-0,0633	-0,0038	0,0231
K_Vosecke	-0,0555	0,0062	0,0201	-0,0027



Obr. MDS mapa výživy kleče v Krkonoších

Dimense	2
Suma čtverců vzdáleností	3,1996
Suma čtverců reziduí	0,0622
Stress	0,0195
Pseudo R^2	89,3192

Metoda MDS rozdělila lokality s výskytem kleče do několika shluků. Hodnota stresu (0,02) je považována za velice dobrou, stejně tak jako hodnota PseudoR^2 (89 %).

Závěr:

Výrazně odlišnou lokalitou je Stříbrná bystřina, tato lokalita je jako jedna z mála sledovaných lokalit na rašeliništi. Další skupinu odlišnou od většiny lokalit jsou lokality v západní části Krkonoš (U čtyř pánů, Krkonoš, Lysá hora, Lysá sedlo a Růženčina vyhlídka). Všechny tyto

lokality leží v západní části Krkonoš na Českém hřebenu. Další nejbližší lokalita Mumlava je již v zákrytu z JZ směru, tato lokalita je již ve hlavním shluku lokalit. Z tohoto shluku se odlišuje lokalita Kozí hřbety. První dimenze rozděluje lokality podle obsahu dusíku, druhá dimenze zejména podle obsahu fosforu.

Licenční studium Statistické zpracování experimentálních dat.
Předmět: 3.5 Klasifikace analýzou vícerozměrných dat
Přednášející: Prof. RNDr. Milan Meloun, DrSc.
Úloha 2. Logistická regrese (LR)

Na pokusné ploše ve smrkovém porostu byly změřeny dimenze jednotlivých stromů (tloušťka, výška, délka koruny) a jejich souřadnice (X, Y). Ze souřadnic stromů a jejich dimenzí byl vypočítán růstový prostor jednotlivých stromů, tj. teoretická růstová plocha, kterou má strom k dispozici. Po ukončení měření byl na ploše proveden těžební zásah, při kterém byla odstraněna část stromů. Po těžebním zásahu byl opět spočítán růstový prostor ponechaných stromů, odstraněním části stromů se růstový prostor ponechaných stromů zvýšil. V následných 5 letech byl na ploše sledován tloušťkový přírůst.
Příklad testuje velikost tloušťkového přírůstu na velikosti (změně) růstového prostoru po provedeném zásahu. Zvětšil se u uvolněných stromů tloušťkový přírůst ve srovnání se stromy neuvolněnými?

Software: NCSS 2004

Podmínky výpočtu:
Závisle proměnná: ROZDIL
Nezávisle proměnné v regresním modelu: TLOUSTKA, VYSKA, KORUNA, RUST_PLOCHA, PRIRUST
Objektů: 100 stromů (řádků)
Znaků: 6

Categories	Count	Rows	Prior	R-Squared	Classified
0	61	61	0,61000	0,50639	90,164
1	39	39	0,39000	0,50639	71,795
Total	100	100			83,000

Jako závisle proměnná byla zvolena proměnná Rozdil, ta může nabývat hodnoty 0 (strom nebyl uvolněn a jeho růstový prostor se nezměnil) nebo hodnoty 1 (strom byl uvolněn, růstový prostor se zvětšil). Neuvolněných stromů (0) bylo logistickým regresním modelem správně klasifikováno 90 %, stejnému modelu odpovídá 72 % stromů uvolněných (proměnná Rozdil =1). Celkově bylo správně klasifikováno 83 % stromů.

Odhad regresních parametrů
Parameter Significance Tests Section (Reference Group: rozdil = 0)

Parameter	Regression coefficient	Standard Error	Wald Z-value	Wald Prob Level	Lower confidence limit	Upper confidence limit
B0: Intercept	1,03023	4,50277	0,229	0,81902	-7,79503	9,85550
B1: koruna	-0,16269	0,16757	-0,971	0,33162	-0,49112	0,16574
B2: prirust	16,37646	5,00461	3,272	**0,00107**	6,56761	26,18530
B3: rust_plocha	0,61410	0,16386	3,748	**0,00018**	0,29293	0,93527
B4: tloustka	-0,37645	0,16994	-2,215	0,02674	-0,70952	-0,04339
B5: vyska	-0,05352	0,23428	-0,228	0,81930	-0,51271	0,40567

Tučně jsou označeny statisticky významné parametry nalezené Waldovým testem.

Model pro ROZDIL = 1

1.03023472102546 -0.162687603871626*koruna + 16.3764555740948*prirust + 0.614099742357509*rust_plocha -0.376454448415156*tloustka -5.35198263117468E-02*vyska

Klasifikace objektů logickým modelem:

Actual	Estimated		Total
	0	1	
0	55	6	61
1	11	28	39
Total	66	34	100

Percent Correctly classified = 83,0%

Chybně klasifikované objekty

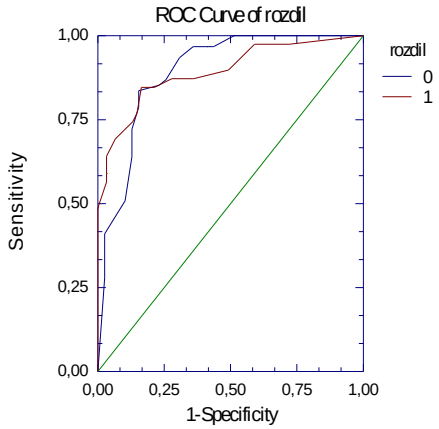
Row	Actual rozdil	Pearson Residual	Deviance Residual	Maximum Hat Diagonal
6*	1	1,26458||.....	1,38217||||||.....	0,22520|||||||||.....
15*	1	1,25499||.....	1,37539||||||.....	0,13442|||||.....
21*	1	1,20612||.....	1,34016||||||.....	0,03575|.....
30*	0	-1,02840||.....	-1,20129|||||.....	0,09996|||.....
44*	1	2,41519||||.....	1,96050|||||||||.....	0,03813|.....
48*	1	2,49738||||.....	1,98957|||||||||.....	0,05377||.....
51*	0	-1,22310||.....	-1,35254||||||.....	0,07403|||.....
56*	1	2,62551||||.....	2,03274|||||||||.....	0,03249|.....
64*	0	-1,04847||.....	-1,21786|||||.....	0,05048||.....
71*	0	-1,12961||.....	-1,28252|||||.....	0,08709|||.....
76*	0	-1,82944|||.....	-1,71432||||||||.....	0,05405||.....
77*	1	1,09509||.....	1,25547|||||.....	0,07306|||.....
80*	1	2,03406||||.....	1,80917||||||||.....	0,03925|.....
85*	1	1,35661||.....	1,44495||||||.....	0,18463|||||||.....
86*	1	7,13000||||||||||||||	2,81002||||||||||||||	0,01398|.....
87*	1	1,71314|||.....	1,65522|||||||.....	0,02806|.....
91*	0	-1,88198|||.....	-1,73973||||||||.....	0,09721||||.....

Sledované znaky chybně klasifikovaných objektů

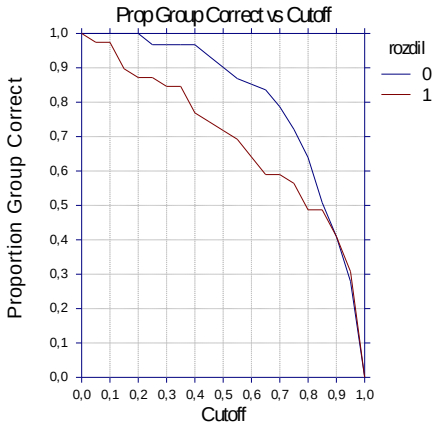
řádek	tloustka	výška	koruna	Růstová plocha	prirust	rozdil
6*	25	20	15	13	0,21	1
15*	14	20	14	5,75	0,22	1
21*	17	20	15	8,25	0,21	1
30*	20	19	17	11,5	0,2	0
44*	19	21	16	6,5	0,25	1
48*	19,9	19	17	6,5	0,27	1
51*	22	21	16	10	0,32	0
56*	15	19	17	6	0,17	1
64*	20	20	17	9,25	0,29	0
71*	17	19	17	10,5	0,18	0
76*	20	20	17	8	0,32	1
77*	19	19	17	8,75	0,19	1
80*	14	17	15	9,5	0,07	1
85*	18	18	16	4,5	0,16	1

86*	18	20	17	8,25	0,21	1
87*	24	22	16	13,25	0,3	0

Metodou byly chybně klasifikovány uvolněné stromy s nízkým přírůstem nebo naopak stromy neuvolněné, přesto mající vyšší tloušťkový přírůst vyšší než střední hodnoty.



Graf ROC pro obě hodnoty ROZDIL



Graf závislosti křivek podílu správně zařazených objektů

Závěr:

Pro soubor dat byl navržen logitový model. Parametry PRIRUST a Rustova_plocha jsou statisticky významné. Správně bylo klasifikováno 83 % objektů. Chybně klasifikovány byly uvolněné stromy s nižším přírůstem nebo naopak stromy bez uvolnění vykazující vyšší přírůst než byla střední hodnota pro celý porost ve sledovaném období.