Agrovoc descriptors: statistical methods, classification, hazelnuts, corylus avellana, weight, height, diameter

Agris category codes: F01, U10

Univerza v Ljubljani
Biotehniška fakulteta
Oddelek za agronomijo

COBISS koda 1.02

Diskriminantna analiza in klasifikacija: osnove in primer

Damijana KASTELEC[1], Katarina KOŠMELJ[2]

Delo je prispelo 15. januar 2008, sprejeto 28. april 2008.
Received January 15, 2008; accepted April 28, 2008.

POVZETEK

V članku so predstavljene osnove diskriminantne analize in klasifikacije. Njuna uporabnost je predstavljena na preprostem primeru analize treh sort leske na podlagi treh morfoloških lastnosti ploda: mase, višine in premera. Izračuni so narejeni s programom SPSS.

Ključne besede: diskriminantna analiza, klasifikacija

ABSTRACT

DISCRIMINANT ANALYSIS AND CLASSIFICATION: THEORY AND ILLUSTRATION

Discriminant analysis and classification are presented in the paper. Their applicability is illustrated on an example of three hazel cultivars described by hazelnut mass, height and diameter. The calculations were made with the SPSS programme.

Key words: discriminant analysis, classification

1 UVOD

Proučujemo k populacij (skupin), iz katerih vzorčimo dovolj velike vzorce; za vsako enoto imamo podatke za večje število številskih spremenljivk. Diskriminantno analizo naredimo, kadar želimo ugotoviti, po katerih spremenljivkah se populacije (skupine) najbolj razlikujejo med seboj. V kmetijstvu so populacije npr. sorte (kultivarji, genotipi), za vsako sorto imamo vzorec enot, za

[1] Asis. Ph. D., Jamnikarjeva 101, SI-1111 Ljubljana, P. O. Box 2995, e-mail: damijana.kastelec@bf.uni-lj.si

[2] Prof. Ph. D., Jamnikarjeva 101, SI-1111 Ljubljana, P. O. Box 2995, e-mail: katarina.kosmelj@bf.uni-lj.si

katere so izmerjene številne lastnosti (npr. morfološke, genetske, kemijske,…). Želimo ugotoviti, po katerih lastnostih se sorte najbolj razlikujejo med seboj.

Na podlagi rezultatov diskriminantne analize lahko nadaljujemo s t. i. klasifikacijo (uvrščanje enot) v populacije (skupine). Klasificiramo enote, za katere ne vemo, v katero populacijo sodijo, imamo pa vrednosti istih osnovnih spremenljivk kot za enote, ki so bile vključene v diskriminantno analizo. Ta postopek ima vlogo napovedovanja. V literaturi je ponavadi diskriminacija in uvrščanje v skupine v istih poglavjih (Johnson in Wichern, 2002, Huberty 1994, Klecka, 1980), včasih je ločeno (Rencher, 1995).

Diskriminacija je iskanje linearnih kombinacij osnovnih p spremenljivk, ki najbolje pojasnijo razlike med k skupinami. Dobljenim linearnim kombinacijam rečemo diskriminantne spremenljivke ali diskriminantne funkcije (angl. discriminant functions, discriminant coordinates, canonical variates). Prva diskriminantna spremenljivka določa, po katerih osnovnih spremenljivkah se populacije najbolj razlikujejo, v drugi diskriminantni spremenljivki so kot pomembnejše zastopane osnovne spremenljivke, ki sledijo po pomembnosti tistim v prvi diskriminantni spremenljivki, itd. Pomembnost posameznih spremenljivk pri razlikovanju skupin ugotavljamo na podlagi velikosti uteži diskriminantnih spremenljivk.

V praksi si želimo, da je pomembnih diskriminantnih spremenljivk čim manj, kar pomeni, da lahko razlike med skupinami razložimo z eno, dvema ali kvečjemu s tremi diskriminantnimi spremenljivkami. Tedaj vrednosti diskriminantnih spremenljivk omogočajo grafične prikaze, ki vizualno predstavijo razlike med skupinami: npr. razsevni grafikon enot v prostoru prvih dveh diskriminantnih spremenljivk predstavlja najboljšo možno dvodimenzionalno predstavitev v smislu razločevanja med skupinami (slika 3).

Za uporabo diskriminantne analize je potrebno, da imamo v posamezni skupini dovolj enot; število enot v posamezni skupini mora biti večje od števila spremenljivk.

V članku bomo predstavili osnove diskriminantne analize in osnove klasifikacije ter njuno praktično uporabo na enostavnem primeru razločevanja treh sort leske, ki so opisane z maso, višino in premerom ploda. Uporabili bomo program SPSS.

2 OSNOVE MATEMATIČNE TEORIJE

2.1 ANOVA in MANOVA

Idejo diskriminantne analize bomo razložili z analogijo z enosmerno analizo variance (ANOVA) in enosmerno multivariatno analizo variance (MANOVA).

2.1.1 Variabilnost med skupinami in znotraj skupin

V analizi variance (ANOVA) analiziramo številsko spremenljivko X na k populacijah. Zanima nas, ali se povprečja k populacij razlikujejo med sabo. Ob

predpostavki o enaki varianci po skupinah postavimo ničelno domnevo $H_0 : \mu_1 = \mu_2 = \Lambda = \mu_k$ in alternativno domnevo H_1, da vsaj dve povprečji nista enaki. Populacije predstavljajo vzorci velikosti n_i, $i = 1,\text{K},k$. Na osnovi vzorčnih vrednosti izračunamo vzorčna povprečja $\bar{x}_i$, $i = 1,\text{K},k$, in skupno povprečje $\bar{x}$. Vsota kvadriranih odklonov B (angl. between groups) in vsota kvadriranih odklonov W (angl. within groups) vrednotita variabilnost med skupinami in variabilnost znotraj skupin:

$$B = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2 \qquad W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 .$$

Njuno razmerje je vsebovano v F-statistiki:

$$F = \frac{B/SP_1}{W/SP_2},$$

ki je v primeru, da ničelna domneva velja, porazdeljena po F-porazdelitvi s stopinjami prostosti $SP_1 = k - 1$ in $SP_2 = n - k$, $n = \sum_{i=1}^{k} n_i$. Večja vrednost F-statistike odraža večje razlike med skupinami.

Če proučujemo več kot eno številsko spremenljivko naenkrat, npr. p spremenljivk X_1, X_2, K, X_p, multivariatna analiza variance (MANOVA) nadgradi analizo variance. Predpostavimo, da ima k populacij enako variančno-kovariančno matriko $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \Lambda = \mathbf{\Sigma}_k = \mathbf{\Sigma}$. Zanima nas, ali se populacije razlikujejo v p-dimenzionalnih vektorjih povprečij $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ do $\boldsymbol{\mu}_k$. Vzorčne vrednosti p spremenljivk v i-ti skupini zapišemo z vektorjem $\mathbf{x}_{ij}$, $j = 1,\text{K},n_i$, in izračunamo vektorje vzorčnih povprečij $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \text{K}, \bar{\mathbf{x}}_k$, $\bar{\mathbf{x}}_i = (1/n_i)\sum_{j=1}^{n_i} \mathbf{x}_{ij}$ ter njihovo skupno povprečje $\bar{\mathbf{x}} = 1/k \sum_{i=1}^{k} \bar{\mathbf{x}}_i$. Podobno kot v univariatnem primeru ocenimo variabilnost med skupinami z matriko $\mathbf{B}$:

$$\mathbf{B} = \sum_{i=1}^{k} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T$$

in variabilnost znotraj skupin z matriko $\mathbf{W}$:

$$\mathbf{W} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T .$$

Variabilnost znotraj skupin izrazimo z vzorčnimi variančno-kovariančnimi matrikami po skupinah $\mathbf{S}_i$, $\mathbf{W} = \sum_{i=1}^{k} (n_i - 1)\mathbf{S}_i$. Nepristrana ocena za $\mathbf{\Sigma}$ je $\mathbf{S}_{skupna} = \mathbf{W}/(n-k)$.

Ničelno domnevo zapišemo v vektorski obliki – delamo primerjavo vektorjev povprečij za posamezne skupine, $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \Lambda = \boldsymbol{\mu}_k$, nasproti alternativni domnevi H_1, da vsaj dva vektorja povprečij nista enaka.

Za preizkus ničelne domneve lahko uporabimo več različnih preizkusov: Wilksov, Lawley-Hottelingov, Royev, Pillaijev. Tu omenimo najpogosteje uporabljen Wilksov preizkus, v katerem je testna statistika Wilksova lambda Λ:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}.$$

Njeno vrednost izračunamo kot razmerje determinante matrike $\mathbf{W}$, $|\mathbf{W}|$, in determinante matrike $\mathbf{B}+\mathbf{W}$, $|\mathbf{B}+\mathbf{W}|$; njeno ničelno porazdelitev najdemo v literaturi (Rencher, 1995, str. 181). Če ničelno domnevo zavrnemo, nadaljujemo ali z analizo kontrastov (načrtovanih primerjav) ali s preizkusi mnogoterih primerjav, oboje postane zapleteno zaradi večjega števila spremenljivk. Če imamo dovolj velike vzorce, lahko kot lažje nadaljevanje MANOVA naredimo diskriminantno analizo.

2.1.2 Enakosti varianc oz. variančno-kovariančnih matrik

Predpostavko o enakosti varianc po skupinah pri ANOVA preverjamo z različnimi preizkusi (Leveneov[3], F-max preizkus,…), ki jih na tem mestu ne bomo podrobneje opisovali (Kuehl, 2000). Nekaj več povejmo o preizkusu domneve o enakosti variančno-kovariančnih matrik v primeru MANOVA. Program SPSS za ta primer uporablja Boxov M-preizkus (Box, 1949). Preverjamo ničelno domnevo $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \Lambda = \mathbf{\Sigma}_k$. Boxova M-statistika je:

$$M = (n-k)\log|\mathbf{S}_{skupna}| - \sum_{i=1}^{k}(n_i - 1)\log|\mathbf{S}_i|.$$

Aproksimacija njene ničelne porazdelitve je F-porazdelitev (Bryan, 2004, str. 49). Ta preizkus je zelo občutljiv na prisotnost že manjših odstopanj od multivariatne normalne porazdelitve, zato je njegove rezultate treba vzeti z rezervo.

2.2 Diskriminantna analiza

V diskriminantni analizi iščemo take linearne kombinacije spremenljivk $X_1, X_2, \mathrm{K}, X_p$, ki kar najbolje razločujejo k populacij. Model diskriminantne analize zapišemo takole:

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$
$$Y_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$
$$\ldots$$
$$Y_s = a_{s1}X_1 + a_{s2}X_2 + \ldots + a_{sp}X_p.$$

Spremenljivki Y_r, $r = 1, \mathrm{K}, s$ rečemo **diskriminantna spremenljivka**, koeficientom linearne kombinacije $\mathbf{a}_r = (a_{r1}, a_{r2}, \ldots, a_{rp})^T$ pa **uteži diskriminantne**

[3] Program SPSS uporablja Leveneov preizkus.

spremenljivke. Veljati mora predpostavka, da so variančno-kovariančne matrike po populacijah enake, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \Lambda = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$.

Uteži diskriminantnih spremenljivk določimo tako, da so razdalje[4] med povprečji diskriminantnih spremenljivk po skupinah maksimalne (slika 3, razdalje med težišči skupin). Kriterij, ki ga uporabimo, je:

$$\max\left(\frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}}\right),$$

kjer je $\mathbf{a}$ matrika uteži. Matematično maksimiranje izraza naredimo z odvajanjem zgornjega izraza po $\mathbf{a}$ in izenačitvi izraza z 0. To privede do posplošenega problema lastnih vrednosti in lastnih vektorjev matrike $\mathbf{W}^{-1}\mathbf{B}$:

$$\left(\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I}\right)\mathbf{a} = 0.$$

Rešitev so lastne vrednosti $\lambda_1 \geq \lambda_2 \geq \mathrm{K} \geq \lambda_s \geq 0$ in pripadajoči lastni vektorji $\mathbf{a}_1, \mathbf{a}_2, \mathrm{K}, \mathbf{a}_s$ matrike $\mathbf{W}^{-1}\mathbf{B}$**.** Lastne vektorje normiramo, sicer pa velja, da so med sabo nekorelirani:

$$\mathbf{a}_i^T\mathbf{S}_{skupna}\mathbf{a}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

Število neničelnih lastnih vrednosti je enako rangu matrike $\mathbf{B}$, $s = \min(k-1, p)$. Tako dobimo s diskriminantnih spremenljivk, ki maksimalno razločujejo (diskriminirajo) skupine. Prva diskriminantna spremenljivka ima smer, vzdolž katere je »razmerje $\mathbf{B}$ proti $\mathbf{W}$« največje; druga diskriminantna spremenljivka ji po tem razmerju sledi, itd. »Razmerje $\mathbf{B}$ proti $\mathbf{W}$« je mera za razločevanje skupin glede na osnovne spremenljivke; poimenovali ga bomo **različnost skupin**. Relativna pomembnost posamezne diskriminantne spremenljivke je razvidna iz razmerja $\lambda_i / \sum_{j=1}^{s} \lambda_j$. Glede na lastnost lastnih vrednosti, $\lambda_1 \geq \lambda_2 \geq \Lambda \geq \lambda_s \geq 0$, to razmerje od prve do zadnje lastne vrednosti pada. Potem razmerje $\lambda_i / \sum_{j=1}^{s} \lambda_j$ obrazložimo kot delež različnosti skupin, ki je pojasnjen z i-to diskriminantno spremenljivko.

Ponovno naj poudarimo, da moramo imeti za diskriminantno analizo dovolj velike vzorce: $n_i > p$. Matematični izračuni se sicer izvedejo tudi ob manj strogem pogoju $n - 2 > p$ in hkrati $n_i \geq 2$, kjer je $n = \sum_{i=1}^{k} n_i$. Izračuni se matematično ne morejo izvesti, če je ena izmed spremenljivk linearna kombinacija ostalih spremenljivk (kolinearnost).

[4] Uporablja se Mahalanobisova razdalja.

2.3 Obrazložitev rezultatov diskriminantne analize

2.3.1 Uteži diskriminantnih spremenljivk

Uteži diskriminantnih spremenljivk odražajo velikost parcialne korelacije med posamezno diskriminantno spremenljivko in posamezno osnovno spremenljivko, torej njuno povezanost ob hkratnem upoštevanju vseh ostalih osnovnih spremenljivk.

Če imajo osnovne spremenljivke isto mersko lestvico in je njihova variabilnost približno enaka, absolutne vrednosti uteži diskriminantnih spremenljivk izražajo relativno pomembnost pripadajoče spremenljivke pri razlikovanju skupin. V praksi so osnovne spremenljivke pogosto po variabilnosti in merski lestvici različne; v takem primeru za obrazložitev uporabimo **standardizirane uteži diskriminantnih spremenljivk**, ki bi jih dobili na standardiziranih osnovnih spremenljivkah[5].

Absolutna velikost standardiziranih uteži diskriminantnih spremenljivk izraža pomembnost pripadajočih osnovnih spremenljivk za razločevanje skupin. Če želimo posamezno diskriminantno spremenljivko vsebinsko poimenovati, je poleg absolutne vrednosti uteži pomemben tudi njen predznak.

V izpisih računalniških programov je tudi t. i. **strukturna matrika**[6], ki vsebuje korelacijske koeficiente med diskriminantnimi in osnovnimi spremenljivkami in ima podobno vlogo kot pri faktorski analizi (Johnson in Wichern, 2002). Ti koeficienti so manj primerni za obrazložitev rezultatov, saj želimo osnovne spremenljivke obravnavati multivariatno.

2.3.2 Kanonična korelacija

Royeva statistika θ_1 meri, kako uspešno prva diskriminantna spremenljivka razločuje skupine. Izračunamo jo kot razmerje »variabilnosti med skupinami B_{Y_1} proti skupni variabilnosti $B_{Y_1} + W_{Y_1}$ « za prvo diskriminantno spremenljivko Y_1. To razmerje lahko izrazimo s prvo lastno vrednostjo matrike $\mathbf{W}^{-1}\mathbf{B}$:

$$\theta_1 = \frac{B_{y_1}}{B_{y_1} + W_{y_1}} = \frac{\lambda_1}{1 + \lambda_1}.$$

Teorija pokaže, da je θ_1 kvadrat koeficienta kanonične korelacije $r_{K,1}$, ki meri povezavo med prvo diskriminantno spremenljivko in linearno kombinacijo $k-1$ nemih spremenljivk; le te imajo vrednosti 0 in 1 in predstavljajo pripadnost enote posameznem vzorcu. Na osnovi lastnih vrednosti matrike $\mathbf{W}^{-1}\mathbf{B}$ izračunamo koeficiente kanonične korelacije za vsako diskriminantno funkcijo[7]:

[5] Standardized canonical discriminant function coefficients (SPSS)
[6] Structure matrix (SPSS)
[7] SPSS poda tudi koeficiente kanonične korelacije.

$$r_{K,i} = \sqrt{\frac{\lambda_i}{1+\lambda_i}}\,.$$

Koeficienti kanonične korelacije so med 0 in 1, njihova velikost pada z zaporednimi diskriminantnimi spremenljivkami. Vrednosti blizu 1 kažejo na veliko uspešnost diskriminantne analize. Če imamo samo dve skupini ($k = 2$), dobimo največ eno diskriminantno spremenljivko in je koeficient kanonične korelacije kar Pearsonov koeficient korelacije med diskriminantno spremenljivko in nemo spremenljivko, ki izraža pripadnost posamezni skupini.

2.3.3 Preizkusi o statistični značilnosti diskriminantnih spremenljivk

V nadaljevanju bomo pogledali, ali diskriminantne spremenljivke razločujejo med populacijami. Za statistično sklepanje je poleg zahteve o enakih variančno-kovariančnih matrikah potrebna tudi predpostavka o večrazsežni normalni porazdelitvi p-spremenljivk. Za preverjanje domneve, ali vzorčni podatki kažejo, da se k-populacij razlikuje po povprečjih diskriminantne spremenljivke, uporabimo Wilksovo lambdo. V prvem koraku preverjamo ničelno domnevo, da se vrednosti dobljenih diskriminantnih spremenljivk med populacijami ne razlikujejo, kar pomeni, da so vse lastne vrednosti matrike $\mathbf{W}^{-1}\mathbf{B}$ enake 0:

$H_0 : \lambda_1 = \lambda_2 = \mathrm{K} = \lambda_s = 0$

proti alternativni domnevi, da se populacije razlikujejo vsaj po prvi diskriminantni spremenljivki. Za preverjanje te domneve uporabimo Wilksovo lambdo:

$$\Lambda_1 = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}\,,$$

ki se aproksimativno porazdeljuje po χ^2-porazdelitvi s $SP = p(k-1)$.

Statistično značilnost naslednjih diskriminantnih spremenljivk preverjamo na enak način z dodatnimi preizkusi istega tipa:

$\Lambda_m = \prod_{i=m}^{s} \frac{1}{1+\lambda_i}$, ki se aproksimativno porazdeljuje po χ^2-porazdelitvi s $SP = (p-m+1)(k-m)$.

V vsebinsko obrazložitev rezultatov diskriminantne analize običajno vključimo le statistično značilne diskriminantne spremenljivke.

2.4 Klasifikacija

2.4.1 Osnove klasifikacije

Klasifikacija je postopek, pri katerem uvrščamo novo enoto, za katero imamo podatke za p spremenljivk ne pa pripadnosti skupini, v eno izmed vnaprej poznanih k populacij (skupin). Populacije se lahko glede na vrednosti p

spremenljivk strogo razločujejo med seboj, lahko pa se bolj ali manj prekrivajo, kar klasifikacijo otežuje. Nove enote uvrščamo v skupine na podlagi t. i. **klasifikacijskega pravila,** ki pravi, naj bo verjetnost uvrstitve nove enote v napačno populacijo čim manjša. **Skupno verjetnost za napačno uvrstitev** nove enote *TPM* (angl. total probability of misclassification) izrazimo z vsoto (Johnson in Wichern, 2002, str. 613):

$$TPM = \sum_{i=1}^{k} p_i \left(\sum_{\substack{j=1 \\ j \neq i}}^{k} P(j|i) c(j|i) \right),$$

kjer je p_i začetna verjetnost (angl. prior probability), da nova enota pripada i-ti populaciji, $P(j|i)$ je verjetnost, da novo enoto, ki dejansko pripada i-ti populaciji, napačno razvrstimo v j-to populacijo, $c(j|i)$ pa stroški napačne klasifikacije. Začetna verjetnost p_i temelji na velikosti populacij: če so vse populacije enako velike, je p_i po populacijah enaka, sicer odraža razmerje velikosti populacij.

Optimalno klasifikacijsko pravilo dobimo tako, da ob minimalni vrednosti *TPM* poiščemo k izključujočih se območij uvrščanja, kar omogoča, da novo enoto uvrstimo v natanko eno populacijo. Za izračun verjetnosti $P(j|i)$ moramo za vsako populacijo poznati p-razsežnostno porazdelitev spremenljivk.

V nadaljevanju si oglejmo primer za p-razsežnostno normalno porazdelitev z enakimi variančno-kovariančnimi matrikami za vseh k populacij. Stroški napačne klasifikacije naj bodo za vse skupine enaki 1. Optimalno klasifikacijsko pravilo, ki določa, v katero izmed k-populacij bo razvrščena enota $\mathbf{x}_0$, dobimo na osnovi vrednosti t. i. **linearnih klasifikacijskih funkcij** $d_i(\mathbf{x}_0)$, $i = 1,\text{K},k$ (angl. linear classification function, linear discriminant scores). Za opisani primer se le-te izražajo takole:

$$d_i(\mathbf{x}_0) = -\frac{1}{2} D_i^2(\mathbf{x}_0) + \ln p_i,$$

kjer je $D_i^2(\mathbf{x}_0)$ kvadrat Mahalanobisove razdalje:

$$D_i^2(\mathbf{x}_0) = (\mathbf{x}_0 - \overline{\mathbf{x}}_i)^T \mathbf{S}_{skupna}^{-1} (\mathbf{x}_0 - \overline{\mathbf{x}}_i).$$

To je razdalja med vektorjema $\mathbf{x}_0$ in $\overline{\mathbf{x}}_i$, ki upošteva korelacijo med spremenljivkami. V postopku klasifikacije novo enoto $\mathbf{x}_0$ uvrstimo v tisto populacijo, kjer je $d_i(\mathbf{x}_0)$ največja.

Če izraz za $d_i(\mathbf{x}_0)$ razvijemo in zanemarimo člen $1/2\,\mathbf{x}_0^T \mathbf{S}_{skupna}^{-1} \mathbf{x}_0$, ki je za vse skupine enak, dobimo enačbo (Johnson in Wichern, 2002, str. 613):

$$d_i(\mathbf{x}_0) = \bar{\mathbf{x}}_i^T \mathbf{S}_{skupna}^{-1} \mathbf{x}_0 - \frac{1}{2} \bar{\mathbf{x}}_i^T \mathbf{S}_{skupna}^{-1} \bar{\mathbf{x}}_i + \ln p_i ,$$

kjer prvi člen predstavlja linearno kombinacijo osnovnih p spremenljivk, druga dva člena pa konstanto za i-to skupino. Linearno klasifikacijsko funkcijo zato lahko zapišemo tudi takole:

$$d_i(\mathbf{x}_0) = c_{i0} + c_{i1} x_{01} + c_{i2} x_{02} + \Lambda + c_{ip} x_{0p} ,$$

c_{i0} je konstanta, c_{ij}, $j = 1,\mathrm{K}, p$ so koeficienti linearne kombinacije, x_{0j} so vrednosti j-te osnovne spremenljivke na novi enoti.[8] Za novo enoto $\mathbf{x}_0$ torej v postopku klasifikacije izračunamo vrednosti k linearnih klasifikacijskih funkcij $d_i(\mathbf{x}_0)$ in jo uvrstimo v tisto skupino, za katero je vrednost klasifikacijske funkcije največja.

2.4.2 Klasifikacija in diskriminantna analiza

Fisherjevo klasifikacijsko pravilo (1936) je direktno povezano z diskriminantno analizo. Za novo enoto $\mathbf{x}_0 = (x_{01},\mathrm{K}, x_{0p})^T$ izračunamo njen položaj v prostoru diskriminantnih spremenljivk $\mathbf{y}_0 = (y_{01},\mathrm{K}, y_{0s})^T$. Za vsako skupino izračunamo oddaljenost $\mathbf{y}_0$ od težišča skupine v prostoru diskriminantnih spremenljivk. Izračuna se kvadrat Mahalanobisove razdalje, ki je v prostoru diskriminantnih spremenljivk enak kvadratu Evklidske razdalje:

$$D_i^2(\mathbf{x}_0) = \sum_{j=1}^{s} (y_{0j} - \bar{y}_{ij})^2 = \sum_{j=1}^{s} (\mathbf{a}_j^T (\mathbf{x}_0 - \bar{\mathbf{x}}_i))^2 ,$$

y_{0j} je vrednost j-te diskriminantne spremenljivke za novo enoto, $\mathbf{a}_j$ je vektor uteži j-te diskriminantne spremenljivke. Novo enoto uvrstimo v skupino, za katero je vrednost Mahalanobisove razdalje v prostoru diskriminantni spremenljivk najmanjša.[9]

Fisherjevo klasifikacijsko pravilo je enakovredno optimalnemu klasifikacijskemu pravilu, če v slednjem verjetnosti p_i ocenimo s $p_1 = p_2 = \mathrm{K} = p_k = 1/k$. Opozorimo naj, da pri klasifikaciji upoštevamo vseh s diskriminantnih spremenljivk ne glede na njihovo pomembnost pri razločevanju skupin (Johnson in Wichern, 2002, str. 638).

2.4.3 Klasifikacija v vlogi verifikacije modela diskriminantne analize

Postopek klasifikacije lahko uporabimo tudi za neke vrste oceno ustreznosti dobljenih diskriminantnih spremenljivk. V ta namen z dobljenimi utežmi

[8] SPSS, *Classification Function Coefficients*
[9] SPSS, *Discriminant Analysis:Classification*, možnost *Casewise results* izračuna kvadrate Mahalanobisove razdalje za vse enote vključene v diskriminantno analizo.

disriminantnih spremenljivk povratno izračunamo vrednosti linearne klasifikacijske funkcije za vsako enoto vključeno v diskriminantno analizo ter jo po opisanem postopku klasificiramo (uvrstimo) v skupino. Nato naredimo t. i. klasifikacijsko tabelo, v kateri je razvidno, koliko enot je bilo pravilno in koliko napačno uvrščenih (preglednica 10). Rezultati tega postopka so zgolj informativni, običajno preoptimistični, ker delamo model diskriminantne analize in njegovo verifikacijo na podlagi istih podatkov.

Boljši način ocene ustreznosti dobljenih diskriminantnih spremenljivk je t. i. **navzkrižno preverjanje** (angl. cross-validation, leaving one-out method). Ta postopek je računsko zahtevnejši, saj naredimo izračune uteži diskriminantnih funkcij n-krat: pri i-tem izračunu izpustimo i-ti podatek in ga nato po klasifikacijskem pravilu uvrstimo v posamezno skupino. Tudi v tem primeru naredimo klasifikacijsko tabelo, na podlagi katere izračunamo delež pravilno uvrščenih enot (preglednica 10). Ta način verifikacije je bolj smiseln in verodostojen.

3 PRIMER UPORABE DISKRIMINANTNE ANALIZE S PROGRAMOM SPSS

Uporabo diskriminantne analize ilustriramo na primeru treh sort leske ('Istrske dolgoplodne leske', 'Tonda gentile dele langhe', 'Fertile de coutard'), ki so opisane s tremi morfološkimi lastnostmi (masa, višina, premer plodu). Za vsako sorto imamo vzorec velikosti 30. Empirične porazdelitve spremenljivk so prikazane na sliki 1, ki nakazuje, da obstajajo razlike v omenjenih morfoloških lastnostih lešnika med tremi sortami.

3.1 Univariatna analiza variance (ANOVA)

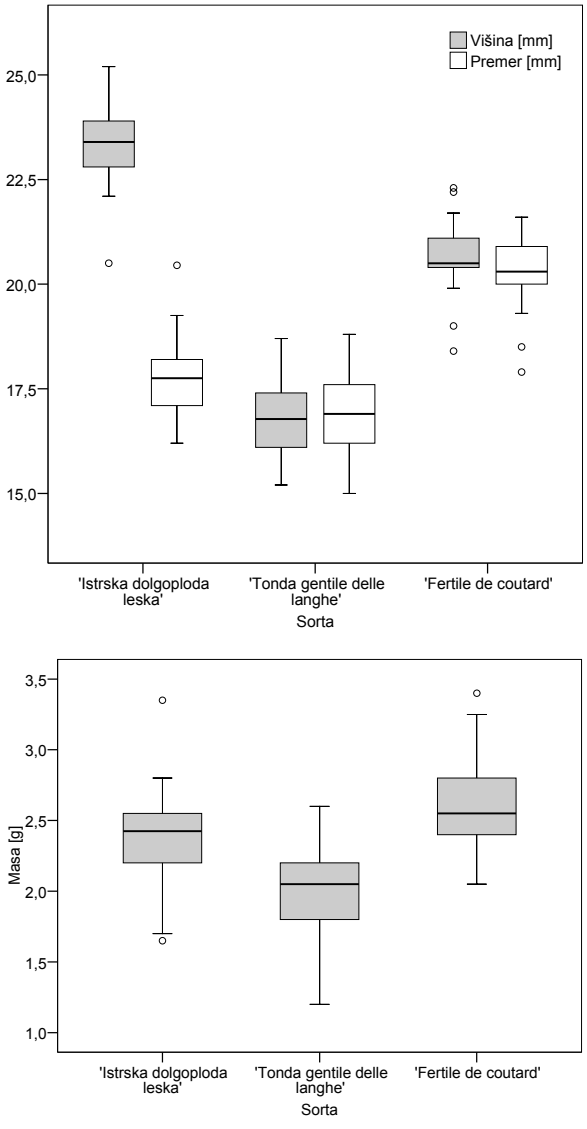
Zaradi kasnejše primerjave rezultatov bomo najprej naredili ANOVA za vsako obravnavano spremenljivko. Iz preglednice 1 je razvidna enakost varianc po sortah, iz preglednic 2 in 3 pa, da se tri izbrane sorte leske statistično značilno razlikujejo med sabo v vseh treh morfoloških lastnostih (ANOVA in Duncanov preizkus).

Preglednica 1: Rezultati Leveneovega preizkusa o enakih variancah po sortah leske za maso, višino in premer plodu. (SPSS, *Analyse/General linear models/Multivariate*, izbira možnosti *Homogeneity tests* v pogovornem oknu *Multivariate:Options*).

Table 1: Levene's test of equality of variances for hazelnut mass, height and diameter for three cultivars.

Leneneov preizkus o enakih variancah
Levene's Test of Equality of Error Variances

	F *F*	SP_1 *df*1	SP_2 *df*2	*p*-vrednost *Sig.*
Masa [g] *Mass* [g]	,044	2	87	,957
Višina [mm] *Height* [mmg	,798	2	87	,453
Premer [mm] *Diameter* [mm]	,522	2	87	,595



Slika 1: Okvirji z ročaji za višino, premer (zgoraj) in maso (spodaj) lešnikov treh sort leske, velikosti vzorcev 30.

Figure 1: Box plot for hazelnut height and diameter (above) and hazelnut mass (below) for three cultivars, sample size 30.

Preglednica 2: ANOVA za maso, višino in premer lešnikov treh sort leske.

Table 2: ANOVA for hazelnut mass, height and diameter for three cultivars.

Spremenljivka *Variables*	Vir *Source*	*VKO* *SS*	*SP* *df*	*SKO* *MS*	*F* *F*	*p*-vrednost *Sig.*
Masa [g] *Mass* [g]	Med sortami *Between cultivars*	5,898	2	2,949	30,3	,000
	Znotraj sort *Within cultivars*	8,473	87	,097		
	Skupaj *Total*	14,371	89			
Višina [mm] *Height* [mm]	Med sortami *Between cultivars*	654,463	2	327,232	426,9	,000
	Znotraj sort *Within cultivars*	66,695	87	,767		
	Skupaj *Total*	721,158	89			
Premer [mm] *Diameter* [mm]	Med sortami *Between cultivars*	192,755	2	96,378	119,2	,000
	Znotraj sort *Within cultivars*	70,379	87	,809		
	Skupaj *Total*	263,134	89			

Preglednica 3: Povprečja za maso, višino in premer lešnikov po sortah in rezultati Duncanovega preizkusa mnogoterih primerjav, $p < 0{,}05$. (SPSS, Analyse/General linear models/Multivariate, v pogovornem oknu Multivariate: Post Hoc izberemo možnost *Duncan*).

Table 3: Averages for hazelnut mass, height and diameter for three cultivars and results of Duncan's test.

Masa [g] *Mass* [g]

	Sorta *Cultivar*	*N*	Podskupina *Subset*		
			1	2	3
Duncan	'Tonda gentile delle langhe'	30	1,992		
	'Istrska dolgoploda leska'	30		2,398	
	'Fertile de coutard'	30			2,608

Višina [mm] *Height* [mm]

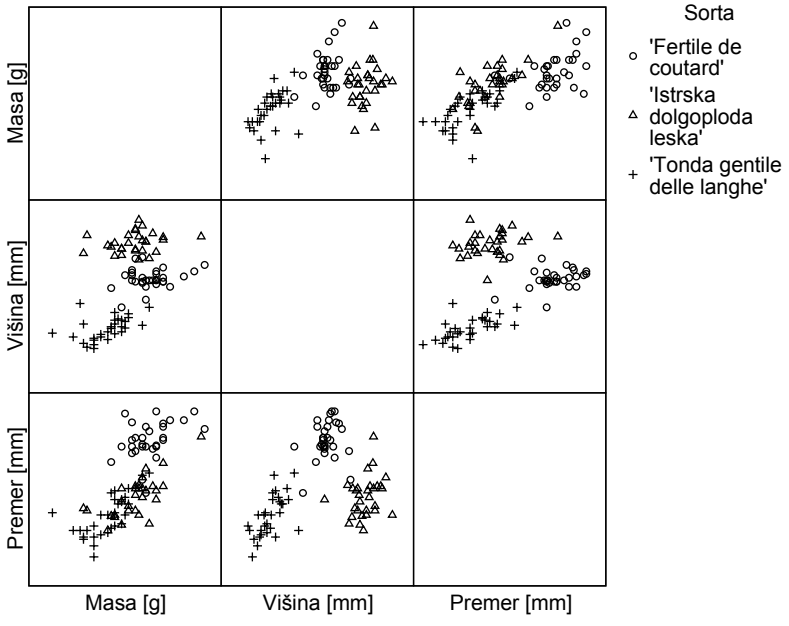
	Sorta *Cultivar*	*N*	Podskupina *Subset*		
			1	2	3
Duncan	'Tonda gentile delle langhe'	30	16,745		
	'Istrska dolgoploda leska'	30		20,647	
	'Fertile de coutard'	30			23,312

Premer [mm] *Diameter* [mm]

	Sorta *Cultivar*	*N*	Podskupina *Subset*		
			1	2	3
Duncan	'Tonda gentile delle langhe'	30	16,860		
	'Istrska dolgoploda leska'	30		17,718	
	'Fertile de coutard'	30			20,303

3.2 Multivariatna analiza variance (MANOVA)

Rezultati univariatne analize variance zadoščajo, če spremenljivke niso povezane med sabo. Morebitno povezanost spremenljivk najlepše vidimo v matriki razsevnih grafikonov (slika 2). Opazimo, da med izbranimi tremi spremenljivkami obstaja rahla linearna povezanost, največja je med maso in premerom ($r = 0{,}609$, preglednica 4).



Slika 2: Matrika razsevnih grafikonov za maso, višino in premer lešnika za tri sorte.

Figure 2: Scatterplot matrix for hazelnut mass, height and diameter for three cultivars.

Preden se lotimo MANOVA, moramo preveriti, ali je za izbrane podatke izpolnjena predpostavka o enakosti variančno-kovariančnih matrik po sortah. V preglednici 4 je podana matrika $\mathbf{S}_{skupna}$ in vzorčne variančno-kovariančne matrike po sortah ($\mathbf{S}_i$, $i = 1,...,3$). Vidimo, da se matrika $\mathbf{S}_{skupna}$ na različne načine razlikuje od matrik $\mathbf{S}_i$, vendar razlike niso dovolj velike, da bi jih Boxov M-preizkus (preglednica 5) odkril kot statistično značilne ($p = 0{,}484$).

Preglednica 4: Vzorčna variančno-kovariančna matrika $\mathbf{S}_{skupna}$ in pripadajoča korelacijska matrika (zgoraj) ter vzorčne variančno-kovariančne matrike za vsako sorto posebej. (SPSS, *Analyze/Classify/Discriminant*, možnost *Within groups covariances* in *Separate groups covariances* v pogovornem oknu *Discriminat Analysis: Statistics*).

Table 4: Sample variance-covariance matrix and its corresponding correlation matrix (upper table) and sample variance-covariance matrices for each cultivar.

Skupna variančno-kovariančna matrika *Pooled Within-Groups Matrices*

		Masa *Mass* [g]	Višina *Height* [mm]	Premer *Diameter* [mm]
Kovarianca *Covariance*	Masa *Mass* [g]	,097	,074	,171
	Višina *Height* [mm]	,074	,767	,291
	Premer *Diameter* [mm]	,171	,291	,809
Korelacija *Correlation*	Masa *Mass* [g]	1,000	,271	,609
	Višina *Height* [mm]	,271	1,000	,369
	Premer *Diameter* [mm]	,609	,369	1,000

Variančno-kovariančne matrike *Covariance Matrices*

Sorta		Masa *Mass* [g]	Višina *Height* [mm]	Premer *Diameter* [mm]
'Istrska dolgoploda leska'	Masa *Mass* [g]	,112	,026	,216
	Višina *Height* [mm]	,026	,870	,210
	Premer *Diameter* [mm]	,216	,210	,826
'Tonda gentile delle langhe'	Masa *Mass* [g]	,092	,132	,197
	Višina *Height* [mm]	,132	,798	,531
	Premer *Diameter* [mm]	,197	,531	,885
'Fertile de Coutard'	Masa *Mass* [g]	,088	,064	,100
	Višina *Height* [mm]	,064	,632	,132
	Premer *Diameter* [mm]	,100	,132	,716

Preglednica 5: Boxov *M*-preizkus o enakosti variančno-kovariančnih matrik po sortah (SPSS, *Analyse/Calssify/Discriminant*, možnost *Box's M* v pogovornem oknu *Discriminant Analysis: Statistics*).

Table 5: Box's *M*-test for equality of variance-covariance matrices for cultivars.

Rezultati preizkusa *Test Results*

Boxova *M*-statistika *Box's M*	12,146
Aproksimativna *F* statistika *F Approx.*	,961
SP_1 *df*1	12
SP_2 *df*2	36680,5
p-vrednost *Sig.*	,484

Torej lahko nadaljujemo z MANOVA in preizkusimo ničelno domnevo, da so vektorji povprečij (masa, višina, premer) za vse tri sorte enaki. Rezultati MANOVA (preglednica 6) pokažejo, da ničelno domnevo zavrnemo, štirje različni testi dajo enako statistično značilnost.

Preglednica 6: MANOVA za maso, višino in premer lešnikov treh sort leske (SPSS, *Analyse/General linear models/Multivariate*).

Table 6: MANOVA for hazelnut mass, height and diameter for three cultivars.

Multivariatni preizkusi *Multivariate Tests*

Dejavnik*Effect*		Vrednost *Value*	*F* *F*	SP_{sorta} *Hypothesis df*	$SP_{ostanka}$ *Error df*	*p*-vrednost *Sig.*
Sorta *Cultivar*	Pillai's Trace	1,650	135,3	6	172	,000
	Wilks' Lambda	,023	158,6	6	170	,000
	Hotelling's Trace	13,220	185,1	6	168	,000
	Roy's Largest Root	10,402	298,2	3	86	,000

Z diskriminantno analizo bomo ugotavljali, katera morfološka lastnost sorte najbolj razločuje.

3.3 Diskriminantna analiza

V diskriminantno analizo vključimo vse tri spremenljivke, izmerjene na treh vzorcih sort leske. Že pri MANOVA smo preverili predpostavko o enakosti variančno-kovariančne matrike za tri sorte leske (preglednica 5), kar je pogoj za uporabo diskriminantne analize.

Diskriminantna analiza pokaže (preglednica 7), da prva diskriminantna spremenljivka pojasni 78,7 % različnosti skupin, druga pa preostalih 21,3 %. Tudi njuna koeficienta kanonične korelacije sta velika (0,96; 0,86). Obe diskriminantni

spremenljivki sta statistično značilni, Wilks-ovi lambdi sta dovolj majhni, da je $p = 0{,}000$.

Največjo standardizirano utež prve diskriminantne spremenljivke (preglednica 8) ima višina (1,064), sledi premer (-0,338), kar pomeni, da se sorte leske med seboj najbolj razločujejo po višini plodov. Druga diskriminantna spremenljivka ima največjo standardizirano utež pri premeru (1,181), sledi masa (-0,308), kar pomeni, da se sorte v manjši meri razločujejo po premeru. Ob upoštevanju višine in premera postane masa plodu nepomembna za razlikovanje treh sort lešnika.

Če pogledamo strukturno matriko - korelacijske koeficiente (preglednica 8), ki merijo povezanost med posamezno diskriminantno spremenljivko in posamezno osnovno spremenljivko, dobimo enake rezultate.

Preglednica 7: Lastne vrednosti (Eigenvalues) diskriminantnih spremenljivk, njihova relativna pomembnost (% of Variance) in pripadajoča kumulativa (Cumulative %) ter koeficient kanonične korelacije (zgoraj). Rezultati Wilksovega preizkusa (spodaj).

Table 7: Eigenvalues, % of variance, cumulative % and canonical correlation (above), Wilks' test (below).

Lastne vrednosti *Eigenvalues*

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	10,402(a)	78,7	78,7	,955
2	2,818(a)	21,3	100,0	,859

Wilksova lambda *Wilks' Lambda*

Preizkus diskriminantnih spremenljivk *Test of Function(s)*	Wilksova lambda *Wilks' Lambda*	Hi-kvadrat *Chi-square*	*SP* *df*	*p*-vrednost *Sig.*
$\lambda_1 = \lambda_2 = 0$ oz. vsaj $\lambda_1 = 0$ *1 through 2*	,023	324,528	6	,000
$\lambda_2 = 0$ *2*	,262	115,222	2	,000

Preglednica 8: Standardizirane uteži diskriminantnih spremenljivk in strukturne uteži (SPSS, *Analyze/Classify/Discriminant*).

Table 8: Stadardized discriminant function coefficients and structure matrix.

Standardizirane uteži diskriminantnih spremenljivk
Standardized Canonical Discriminant Function Coefficients

	Funkcija *Function*	
	1	2
Masa *Mass* [g]	,077	-,308
Višina *Height* [mm]	1,064	-,074
Premer *Diameter* [mm]	-,338	1,181

Strukturna matrika
Structure Matrix

	Diskriminantna spremenljivka *Function*	
	1	2
Višina *Height* [mm]	,960	,279
Premer *Diameter* [mm]	,102	,966
Masa *Mass* [g]	,159	,391

V preglednici 9 so navedene uteži diskriminantnih spremenljivk. Potrebujemo jih za izračun vrednosti diskriminantnih spremenljivk (angl. discriminant score) za enote v vzorcih. Poglejmo si en primer izračuna: za lešnik sorte 'Istrska dolgoploda leska' imamo naslednje vrednosti spremenljivk $masa = 2{,}65$ g , $višina = 24{,}15$ mm in $premer = 18{,}20$ mm ; vrednosti diskriminantnih spremenljivk (y_1 in y_2) izračunamo z enačbama[10]:

$$y_1 = -18{,}299 + 0{,}246 \cdot masa + 1{,}216 \cdot višina - 0{,}376 \cdot premer = 4{,}783$$
$$y_2 = -20{,}006 - 0{,}988 \cdot masa - 0{,}085 \cdot višina + 1{,}314 \cdot premer = -0{,}769 .$$

Ti dve vrednosti določata točko, ki prikazuje izbrani lešnik v ravnini diskriminantnih spremenljivk na sliki 3. Po enakem postopku so izračunane vrednosti diskriminantnih spremenljivk za vse ostale lešnike. Njihovo povprečje za posamezno sorto predstavlja težišče sorte, ki je na sliki 3 prikazano s črnim kvadratkom. Koordinate težišč sort leske v ravnini diskriminantnih spremenljivk so v zadnji tabeli preglednice 9. Slika 3 prikazuje, da prva diskriminantna spremenljivka dobro razločuje vse tri sorte, kar je razvidno iz projekcije težišč na absciso. Druga diskriminantna spremenljivka pa razločuje predvsem sorto 'Fertile de coutard' od ostalih dveh, saj sta projekciji težišč sort 'Istrska dolgoploda leska' in 'Tonda gentile delle langhe' na ordinato zelo blizu skupaj.

[10] Vrednosti diskriminantnih spremenljivk so izračunane na podlagi uteži diskriminantnih spremenljivk z vsaj šestimi decimalkami, v tabelah in v enačbi so le te prikazane samo s tremi decimalnimi mesti.

Preglednica 9: Uteži diskriminantnih spremenljivk in težišča sort v ravnini diskriminantnih spremenljivk (SPSS, *Analyze/Classify/Discriminant*, gumb *Statistics*, možnost *Unstandardized* v razdelku *Function Coefficients*).

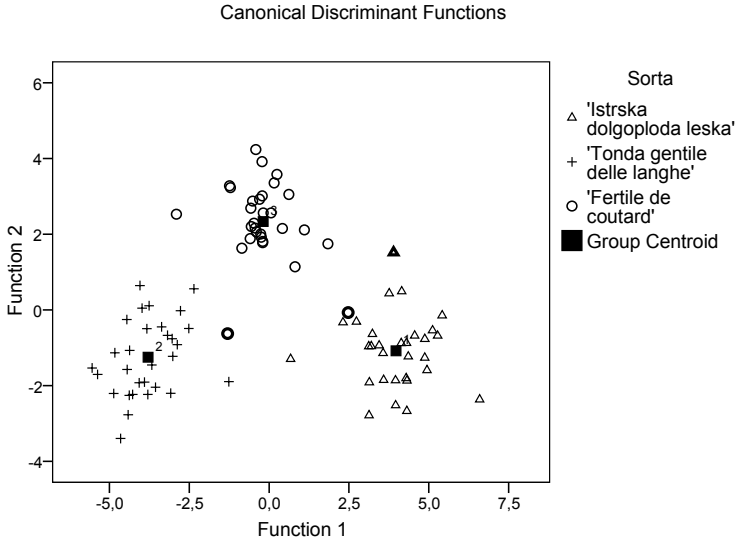
Table 9: Unstandardized canonical discriminant function coefficients and group centroids.

Nestandardizirane uteži diskriminantnih spremenljivk
Canonical Discriminant Function Coefficients

	Diskriminantna spremenljivka *Function*	
	1	2
Masa *Mass* [g]	,246	-,988
Višina *Height* [mm]	1,216	-,085
Premer *Diameter* [mm]	-,376	1,314
(Konstanta) (*Constant*)	-18,299	-20,006

Težiščne vrednosti diskriminantnih spremenljivk
Functions at Group Centroids

Sorta *Cultivar*	Diskriminantna spremenljivka *Function*	
	1	2
'Istrska dolgoploda leska'	3,973	-1,082
'Tonda gentile delle langhe'	-3,787	-1,250
'Fertile de Coutard'	-,186	2,332



Slika 3: Razsevni grafikon v ravnini prvih dveh diskriminantnih spremenljivk dobljenih pri diskriminantni analizi treh morfoloških lastnosti lešnikov treh sort leske. (SPSS, *Analyze/Classify/Discriminant*, gumb *Classification*, razdelek *Plots*, možnost *Combined-groups*).

Figure 3: Scatterplot in the space of the first two discriminant variables.

3.4 Klasifikacija za namen verifikacije modela diskriminantne analize

V preglednici 10 so koeficienti linearnih klasifikacijskih funkcij, ki jih dobimo na podlagi zgoraj predstavljenih diskriminantnih spremenljivk in predpostavk klasifikacije. Poleg enakosti variančno-kovariančnih matrik smo tu predpostavili še več-razsežnostno normalno porazdelitev (Shapiro-Wilkov preizkus za naše podatke pokaže, da je predpostavka upravičena).

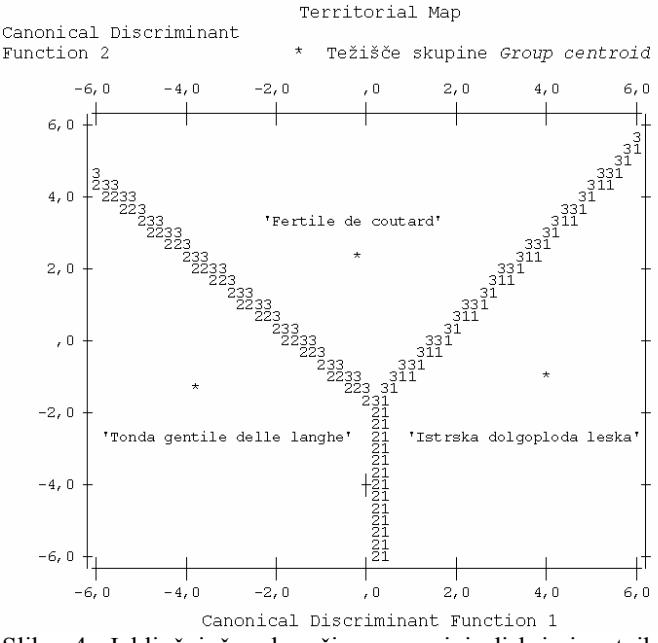
Preglednica 10: Koeficienti treh linearnih klasifikacijskih funkcij: c_{i0} in c_{ij}, $j = (masa, višina, premer)$, $i =$ ('Istrska dolgoploda leska', 'Tonda gentile delle langhe', 'Fertile de coutard'). (SPSS, *Analyze/Classify/Discriminant*, gumb *Statistics*, možnost, v razdelku *Function Coefficients* izberemo možnost *Fisher's*).

Table 10: Coefficients of three classification functions: c_{i0} in, c_{ij}, $j = (mass, height, diameter)$, $i =$ ('Istrska dolgoploda leska', 'Tonda gentile delle langhe', 'Fertile de coutard').

Koeficienti linearnih klasifikacijskih funkcij
Classification Function Coefficients

	Sorta *Cultivar*		
	'Istrska dolgoploda leska'	'Tonda gentile delle langhe'	'Fertile de coutard'
Masa *Mass* [g]	-27,359	-29,101	-31,754
Višina *Height* [mm]	26,109	16,688	20,762
Premer *Diameter* [mm]	18,296	20,991	24,343
(Konstanta) (*Constant)*	-434,697	-288,798	-421,145

Slika 4 prikazuje tri izključujoča si območja v prostoru diskriminantnih spremenljivk, ki so določena na podlagi dobljenega klasifikacijskega pravila (angl. territorial map).



Slika 4: Izključujoča območja v ravnini diskriminantnih spremenljivk, ki jih določajo vrednosti linearnih klasifikacijskih funkcij (SPSS, *Analyze/Classify/Discriminant*, gumb *Classification*, razdelek *Plots*, možnost *Territorial map*).

Figure 4: Territorial map in the space of discriminant variables.

Metodo klasifikacije bomo uporabili za oceno primernosti modela diskriminantne analize. Začetne verjetnosti so v našem primeru za vse tri sorte leske enake, $p_1 = p_2 = p_3 = 1/3$. Za vsak lešnik izračunamo kvadrat Mahalanobisove razdalje D_i^2, $i =$ ('Istrska dolgoploda leska', 'Tonda gentile delle langhe', 'Fertile de coutard'), in ga uvrstimo (klasificiramo) v tisto sorto, za katero je ta razdalja najmanjša. Če za prikaz primera izračuna uporabimo isti lešnik kot zgoraj, dobimo:

$$D^2_{Istrska} = (4{,}873 - 3{,}973)^2 + (-0{,}769 - (-1{,}082))^2 = 0{,}908$$
$$D^2_{Tonda} = (4{,}873 - (-3{,}787))^2 + (-0{,}769 - (-1{,}250))^2 = 75{,}240$$
$$D^2_{Fertile} = (4{,}873 - (-0{,}186))^2 + (-0{,}769 - 2{,}332)^2 = 35{,}219\,.$$

Izbrani lešnik je najbližje težišču 'Istrske dolgoplode leske'. Enak rezultat dobimo, če izračunamo vrednosti treh linearnih klasifikacijskih funkcij (preglednica 10):

$$d_{Istrska} = -434{,}697 - 27{,}359 \cdot masa + 26{,}109 \cdot višina + 18{,}296 \cdot premer = 456{,}31$$
$$d_{Tonda} = -288{,}798 - 29{,}101 \cdot masa + 16{,}688 \cdot višina + 20{,}991 \cdot premer = 419{,}15$$
$$d_{Fertile} = -421{,}145 - 31{,}754 \cdot masa + 20{,}762 \cdot višina + 24{,}343 \cdot premer = 439{,}16\,.$$

Največja vrednost je $d_{Istrska}$, kar pomeni, da po klasifikacijskem pravilu lešnik pravilno uvrstimo k sorti 'Istrska dolgoploda leska'.
Izkaže se, da skupno dobimo dva napačno uvrščena lešnika, ki sicer pripadata sorti 'Fertile de coutard', eden od njiju je uvrščen k sorti ' Istrska dolgoploda leska', drugi pa k sorti 'Tonda gentile delle langhe' (preglednica 11). Ostalih 88 lešnikov (97,8 %) je pravilno uvrščenih. Na sliki 3 sta napačno uvrščena lešnika prikazana kot odebeljena krogca.

Pri navzkrižnem preverjanju (preglednica 11) dobimo napačno uvrščene tri lešnike, ista dva kot po prejšnjem postopku in še en lešnik iz sorte ' Istrska dolgoploda leska', ki je napačno uvrščen k sorti 'Tonda gentile delle langhe', na sliki 3 je ta lešnik prikazan z odebeljenim trikotnikom. Ti rezultati kažejo, da lahko zaupamo rezultatom diskriminantne analize.

Preglednica 11: Rezultati klasifikacije lešnikov na podlagi linearne klasifikacijske funkcije (SPSS, Analyze/Cassify/Discriminant, klik gumba Cassify, izbira možnosti Summary table in leave-one out classification v razdelku Display).

Table 11: Classification results.

Rezultati klasifikacije *Classification Results*

		Sorta *Cultivar*	Napovedana pripadnost skupini *Predicted Group Membership*			Skupaj *Total*
			'Istrska dolgoploda leska'	'Tonda gentile delle langhe'	'Fertile de Coutard'	
Dejanska pripadnost skupini *Original*	Število *Count*	'Istrska dolgoploda leska'	30	0	0	30
		'Tonda gentile delle langhe'	0	30	0	30
		'Fertile de Coutard'	1	1	28	30
	%	'Istrska dolgoploda leska'	100,0	,0	,0	100,0
		'Tonda gentile delle langhe'	,0	100,0	,0	100,0
		'Fertile de Coutard'	3,3	3,3	93,3	100,0
Navzkrižno preverjanje *Cross-validate*	Število *Count*	'Istrska dolgoploda leska'	29	0	1	30
		'Tonda gentile delle langhe'	0	30	0	30
		'Fertile de Coutard'	1	1	28	30
	%	'Istrska dolgoploda leska'	96,7	,0	3,3	100,0
		'Tonda gentile delle langhe'	,0	100,0	,0	100,0
		'Fertile de Coutard'	3,3	3,3	93,3	100,0

3.5 Razprava

Na preprostem primeru smo prikazali uporabo diskriminantne analize in klasifikacije kot metode za oceno ustreznosti modela diskriminantne analize. Univariatna ANOVA pokaže, da se sorte statistično značilno razlikujejo po masi, premeru in višini lešnikov. Diskriminantna analiza kot multivariatna metoda pa je pokazala, da se plodovi treh sort leske med seboj najbolj razlikujejo po višini, nato

po premeru, masa postane ob upoštevanju višine in premera nepomembna za razlikovanje sort.
Klasifikacijo smo v obravnavanem primeru uporabili le za verifikacijo modela diskriminantne analize ob predpostavkah, ki omogočajo uporabo Fisherjevega klasifikacijskega pravila. Metodo bi lahko uporabili tudi za uvrščanje lešnikov neznane sorte z znanimi vrednostmi za maso, višino in premer plodu, v eno izmed treh obravnavanih sort leske.

4 ZAKLJUČEK

Diskriminantna analiza je multivariatna statistična metoda, ki upošteva linearno povezanost osnovnih spremenljivk, zaradi katere določenih zakonitosti v podatkih ne moremo razbrati ob univariatnih analizah posameznih spremenljivk. Obstajajo primeri podatkov, ko univariatne analize posameznih spremenljivk ne pokažejo statistično značilnih razlik med populacijami, diskriminantna analiza pa pokaže, da lahko populacije razlikujemo na podlagi ene ali več linearnih kombinacij osnovnih spremenljivk (diskriminantnih funkcij).

Rezultati diskriminantne analize so lahko napačni, če korelacija med osnovnimi spremenljivkami ni linearna ali če v podatkih obstaja veliko osamelcev. Slednje ponavadi povzroči, da ne moremo predpostaviti enakih variančno-kovariančnih matrik po populacijah. Zato je potrebno na začetku statistične analize narediti različne pregledovalne grafične predstavitve podatkov, ki pokažejo morebitno nelinearnost in prisotnost osamelcev. V določenih primerih se tako nelinearnosti kot tudi osamelcev znebimo z ustreznimi transformacijami osnovnih spremenljivk.
Pri obrazložitvi rezultatov diskriminantne analize se moramo zavedati, da so rezultati zanesljivi le, kadar je razmerje med številom enot v vzorcih in številom osnovnih spremenljivk ($\sum_{i=1}^{k} n_i / p$) dovolj veliko; nekateri priporočajo vrednost tega razmerja od 4 do 5. Če je to razmerje majhno, so rezultati vezani na izbrane vzorce in jih ne moremo posplošiti na pripadajoče populacije. Včasih se zgodi, da dobimo posamezno diskriminantno spremenljivko statistično značilno, čeprav je njen prispevek k razločevanju skupin $\lambda_i / \sum_{j=1}^{s} \lambda_j$ zelo majhen, v takem primeru ji ne posvečamo posebne pozornosti.

Diskriminantno analizo smo prikazali kot možnost nadaljevanja enosmerne multivariatne analize variance, ki je primerna, kadar obravnavamo podatke pridobljene za slučajne skupine. V primerih, ko osnovnih predpostavk diskriminantne analize ne moremo izpolniti, je primerneje, če podatke analiziramo z logistično regresijo (za dve populaciji) ali pa z multinomsko logistično regresijo (za več populacij).

5 LITERATURA

Box, G. E. P. A general distribution theory for a class of likelihood criteria. Biometrika, 36, 1949, 317-346.

Bryan F. J. Manly. Multivariate Statistical Methods, A primer, Third edition, Chapman and Hall/CRC, London, 2004, 214 str.

Chattfield C./ Collins A. J.. Introduction to multivariate analysis, Chapman and Hall/CRC, London, 1980, 248 str.

Ferligoj, A. http://vlado.fmf.uni-lj.si/vlado/podstat/Mva/DA.pdf, 18. 9. 2007

Fisher, R. A.. The use of multiple measurements in the taxonomic problems. Annals of Eugenics, 7, 1936, 179-188.

Huberty, C. J. Applied Discriminant Analysis, John Wiley & Sons, Inc., New York, 1994, 466 str.

Johnson, R. A./ Wichern, D. W. Applied multivariate Statistical Analysis. Prentice Hall, New Jersey, 2002, 767 str.

Klecka R. William. Discriminant Analysis, Quantitative Applications in the Social Sciences Series, No. 19. Thousand Oaks, CA: Sage Publications, 1980, 71 str.

Kuehl R. O. Design of Experiments, Statistical Principles of Research Design and Analysis, Second Edt., Duxbury Thomson Learning, 2000, 664 str.

Rencher, A. C. Methods of Multivariate Analysis. John Wiley 6 Sons, Inc., New York, 1995, 627 str.

http://www2.chass.ncsu.edu/garson/pa765/discrim.htm, 24. 10. 07

http://www.statsoft.com/textbook/stdiscan.html, 24. 10 2007