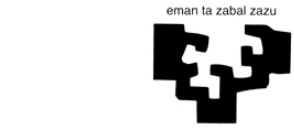


ZTF-FCT
Zientzia eta Teknologia Fakultatea
Facultad de Ciencia y Tecnología

Ekologia-datuak aztertzeko aldagai anitzeko analisiaren aplikazioa

Gradu Amaierako Lana
Matematikako Gradua

Miren López Llona

Arantza Urkaregi Etxepare
Irakasleak zuzendutako lana

Leioa, 2015eko irailaren 3a

Gaien Aurkibidea

Esker onak

Nire eskerrik beroena eman nahi diet modu batean edo bestean lan hau egiten lagundu didaten pertsona guztiei.

Lehenik eta behin, eskerrak eman nahi dizkiot Arantza Urkaregiri, memoria hau zuzendutako irakasleari, eskaini didan laguntza guztiagatik eta beti erakutsi didan prestutasunagatik. Eta nola ez, nirekin izan duen pazientziagatik eta eman dizkidan animoengatik, eskerrik asko.

Irantzu Barrio ere eskertu nahi nuke, edozein zalantzaren aurrean laguntzeko prest egon delako.

Era berean, eskerrak Matematika Graduko irakasle guztiei, lau urte hauetan irakatsitako guztiagatik eta erakutsitako gogo eta inplikazioagatik.

Azkenik, eta batez ere, prozesuaren alde gogorrenak jasan behar izan dituzten pertsonak eskertu nahi ditut, nire gurasoak eta ahizpa, nigan konfiantza izateagatik eta egunero emandako indar eta orekagatik. Eskerrak ematen dizkizuet orain eta beti.

Beste barik, espero dut memoria hau irakurtzen, nik idazten bezainbeste gozatzea.

Laburpena

Gradu Amaierako Lan honetan ekologia-datuak aztertzeko erabiltzen diren aldagai anitzeko hainbat teknika lantzen dira.

Lehenengo kapituluan, aldagai anitzeko analisiko oinarrizko lau metodoak aurkezten ditugu: erregresioa, eskala edo ordenazioa, kluster analisia eta sailkapena. Ondorengo kapituluetan teknika horiek landu ditugu.

Erregresio metodoekin hasten gara bigarren kapituluan. Bertan, erregresio lineala eta horren orokorpenak diren eredu lineal orokortuak azaltzen ditugu. Hirugarren kapituluan, ordenazio metodoak aurkezten ditugu, jatorrizko aldagaiak dimentsio txikiagoen bidez deskribatzeko balio dutenak. Horien artean daude faktore-analisia eta korrespondentzia analisia, eta baita ekologian funtsezkoa den korrespondentzia analisi kanonikoa ere, datu biologikoak eta ingurumen-ezaugarriak erlazionatzen baititu. Laugarren kapituluan kluster analisia eta analisi diskriminatzailea azaltzen ditugu, indibiduoak kategoriatan sailkatzeko balio dutenak. Teknika horiek modu sinplean deskribatzen dituzte aldagaien arteko eta indibiduoen arteko erlazioak. Azken kapituluan, memoriaren inguruko zenbait ondorio biltzen dira.

Lan honetan zenbait adibide praktiko eskaintzen dira, aurkeztutako metodoen argigarri gisa. Metodoak ekologia-datuetan aplikatu ditugu. Erabilitako datu-baseak eta R-ko scriptak A eranskinean eskura daude.

Abstract

In this Final Degree Project we discuss different multivariate analysis techniques applied in ecology and environmental biology.

In the first chapter, we present four basic classes of methods of multivariate analysis: regression, scaling or ordination, clustering and classification. The following chapters are devoted to explain those techniques.

We start with regression methods in Chapter 1. We explain linear regression and several variants of it, gathered together under the collective title of generalized linear models. The third chapter contains ordination methods such us factor analysis, correspondence analysis and canonical corresponence analysis, one of the key methodologies in ecology, which attemps to relate multivariate biological responses to multivariate environmental predictors. Ordination methods simplify multivariate data into low dimensional graphics. The fourth chapter introduces clustering and discriminant analysis, which are used for grouping a set of objects in different categories. Both methods visualize inter-sample and inter-variable relationships in a fairly simple way. In the last chapter, we have drawn several conclusions from this work.

In order to illustrate the statistical methods discussed throughout the proyect, we carry out some case studies using data sets on ecology. All the analyses presented in the project can be replicated using the R scripts and data sets that are avaiable on appendix A.

Sarrera

Aldagai anitzeko analisia estatistikaren adarra da, eta aldagai bat baino gehiago behatzetik lorturiko datuak aztertzea, deskribatzea eta interpretatzea du helburu.

Aldagai anitzeko analisiak metodo estatistiko eta matematiko ugari barneratzen ditu. Azken hamarkadetan informatika arloan izandako aurrerapenek ahalbidetu dute aldagai anitzeko analisiko tekniken erabilpena.

Teknika horiek metodo eta prozedura ezberdinak erabiltzen dituzten arren, helburu berdina dute finean: fenomeno bat laburbiltzea, sintetizatzea eta bera ulertzeko erraztasunak ematea. Teknika horiekin, n indibiduorengan neurtutako p aldagaien balioak jasotzen dituen datu-taulatik abiatuz, bertan bildutako informazioa laburbiltzen da parametro kopuru txikiago baten bidez, eta elementuak modu sinplean irudikatzen dira, beraien arteko konparazioak errazago egiteko; beste batzuetan, berriz, interesezko aldagai bat, besteen arabera azaltzea bilatzen da.

Aldagai anitzeko analisiak aplikazio ugari ditu zientzia esperimentaletan, hala nola biologian eta ekologian. Izan ere, biologia-aniztasuna espezie askoren interakzioen emaitza da, baita espezieen ingurunea zehazten duten faktore mugatzaileen ondorio ere. Faktore horien adibide dira tenperatura edo euria bezalako parametro meteorologikoak, lurraren egitura edo sakontasuna gisako parametro fisikoak; karbono dioxido-maila edo kutsadura-maila bezalako parametro kimikoak, eta abar.

Hala ere, ekologia-datuek zenbait zailtasun aurkezten dituzte estatistika-teknikak aplikatu ahal izateko.

Batetik, ekologia-datuetan ohikoa da laginak txikiak izatea eta laginako indibiduoak multzokatuta egotea. Horrek arazoak eragin ditzake inferentziak egiteko. Era berean, datu ekologikoekin lan egitean, mota guztietako aldagaiak aurki ditzakegu: jarraituak, osoak, dikotomikoak, kategorikoak... Zoritzarrez, metodo estatistikoek ez dute balio edozein aldagai analizatzeko; beraz, aztertzen ari garen aldagaiaren arabera,

metodorik egokiena aukeratu beharko dugu.

Ekologia-datuak lantzeko aldagai anitzeko metodoak aplikatzerakoan aurki ditzakeegun beste arazo batzuk dira: *outlierrak* edo muturreko datuak –azken emaitzak eta ondorioak baldintza ditzaketenak–, aldagaien arteko kolinealtasuna, normalitate falta, edota datuetan zero ugari egotea, beste askoren artean.

Gradu Amaierako Lan honetan, aldagai anitzeko analisiko teknika ezberdinetan sakonduko dugu eta ekologia-datuetan aplikatuko ditugu. Kasu bakoitzean dauden zailtasunak nabarmenduko ditugu eta estatistika-teknika egokiena aukeratzen saiatuko gara.

Lan honen **helburu nagusiak** honako hauek dira:

1. Analisi estatistikoa burutzeko metodo egokiena aukeratzen jakitea.
2. Metodoa zuzentasunez inplementatzea programa estatistiko baten laguntzaz (lan honetan R software librea [12] erabiliko dugu).
3. Emaitzak zuzen interpretatzea.
4. Adierazpen grafikoak erabiltzea interpretazioak errazteko.
5. Metodo ezberdinen osagarritasuna nabarmentzea.
6. Ekologia-adituen eta matematikarien arteko zubia eraikitzeko beharra azpimarratzea.

1. Kapitulua

Aldagai anitzeko analisiko oinarrizko lau metodoak

Kapitulu honetan, ekologia-datuak aztertzeko erabiltzen diren aldagai anitzeko metodoak sailkatzen dira, Greenacrek eta Primiceriok [1] proposatzen duten bezala.

1.1 Datu-matrizeak

Oro har, aldagai anitzeko analisian erabiltzen diren datu-baseak laukizuzen itxurako matrizeak dira. Errenkadetan indibiduoak edo lagineko unitateak ageri dira, adibidez, laborategiko laginak, animalia-, landare- edo mineral-unitateak, eta abar. Zutabeetan, aldiz, aldagaiak azaltzen dira; esaterako, animalia- edo landare-espezieak, konposatu kimikoak edota ingurumen-ezaugarriak (kutsadura, altuera, sakonera, besteak beste).

Hala ere, indibiduo baino askoz aldagai gehiago daudenean aldagaiak errenkada bezala defini daitezke.

1.2 Metodo funtzionalak eta egiturazko metodoak

Jarraian bi datu-matrize bereiziko ditugu, metodo bakoitzaren helburua kontuan izanda.

(i) Behatutako aldagaietako bat besteengandik bereizten bada azterlanean eginkizun berezia duelako, aldagai horri *erantzun aldagaia* edo *menpeko aldagaia* deritzo eta $\mathbf{y}$ bektorearen bidez adieraziko dugu.

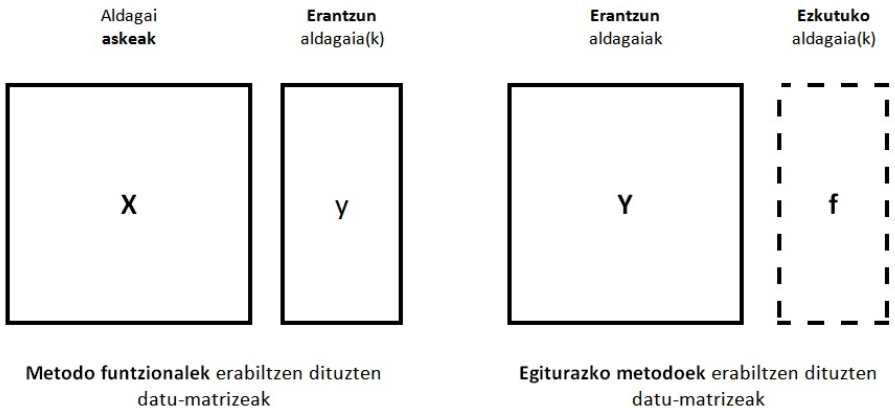
Gainerako aldagaiei *aldagai askeak* edo *aldagai azaltzaileak* deritze eta **X** matrizean biltzen dira. Zenbait erantzun aldagai ere izan genezake, **Y** matrizean bilduta.

Itxura horretako datu-matrizeak erabiltzen dituzten aldagai anitzeko metodoak **metodo funtzionalak** deitzen dira. Metodo funtzionalen helburua **X** aldagai askeen eta **y** menpeko aldagaiaren arteko erlazioa aztertzea da, **y** **X** aldagaien funtzio moduan adieraztea, alegia.

(ii) Bestalde, gerta liteke zenbait erantzun aldagai izatea, menpeko aldagairik gabe, hau da, aldagai guztiak maila berean egotea. Egoera honetan, badago oharkabeko aldagai bat, **f**, *ezkutuko aldagaia* deritzona. **f** aldagaiak guk behatutako **Y** datuak (edo zati bat) eragin dituela pentsatuko dugu. Oharkabeko aldagai bat baino gehiago badago, **F** matrizea izango dugu, ezkutuko aldagaiez osatuta.

Itxura horretako datu-matrizeak erabiltzen dituzten aldagai anitzeko metodoak **egiturazko metodoak** deitzen dira. Egiturazko metodoek **Y** datu matrizearen azpian dagoen egitura aurkitzea dute helburu. Ezkutuko egitura hori forma askotakoa izan daiteke, esate baterako, gradienteak edo topologiak.

1.1. Irudia: Datu-matrizeen formatuak [1]

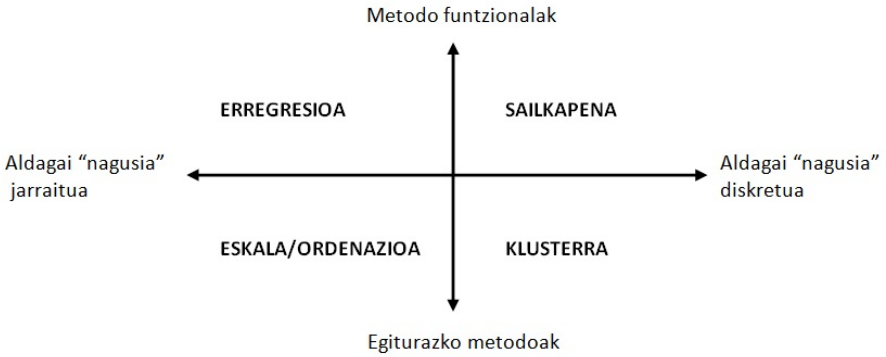


1.3 Aldagai anitzeko analisiaren oinarrizko lau metodoak

Era berean, erantzun aldagaia (metodo funtzionalen kasuan) eta ezkutuko aldagaia (egiturazko metodoen kasuan) jarraitua edo kategorikoa den arabera, beste bi talde bereizten dira.

Hortaz, guztira Greenacrek eta Primiceriok proposaturiko lau multzoak ditugu: aldagai jarraituak azaltzen dituzten metodo funtzionalak (**erregresioa** edo antzeko metodoak), aldagai diskretuak azaltzen dituzten metodo funtzionalak (**sailkapena**), ezkutuko egitura jarraitua duten egiturazko metodoak (**eskala/ordenazioa**) eta ezkutuko egitura kategorikoa duten egiturazko metodoak (**klusterra**).

1.2. Irudia: Aldagai anitzeko analisiko metodoen sailkapena [1]



2. Kapitulua

Erregresioa eta Eredu lineal orokortuak

2.1 Sarrera

Aurreko kapituluan aipatu dugun bezala, metodo funtzionalen helburua $\mathbf{y}$ erantzun aldagaia $\mathbf{X}$ aldagai askeen funtzio moduan azaltzea da.

Kapitulu honetan, metodo funtzionalak azalduko ditugu. Erregresio lineala eta honen zenbait aldaera aurkeztuko ditugu, eredu lineal orokortuak izenez ezagutzen direnak. Metodo horiekin guztiekin ekuazio matematiko bat lortuko dugu, non $\mathbf{y}$ erantzun aldagaia $\mathbf{X}$ aldagai askeen funtzio lineal batekin erlazionatzen den.

Erregresio metodoek erantzun aldagaiaren eta aldagai askeen arteko erlazioa aztertzeaz gain, aldagai askeen behaketa berrien erantzuna aurresateko balio dute.

Gure testuinguruan, erantzun aldagai baten adibidea landare-espezie baten oparotasuna izan daiteke, eta aldagai askeak hurrengo ingurumen-ezaugarriak izan daitezke: lurraren egitura, pH maila, altitudea, landarea eguzkitan egon den ala ez, etab. (landarearen oparotasunean eragina izan dezaketen aldagaiak). Alde batetik, botanikari batek erantzun aldagaiaren eta aldagai askeen arteko erlazioa aztertuko du; eta bestalde, gisa honetako galderak erantzuten ahaleginduko da: "euria egin ezean, zein izango litzateke landarearen oparotasuna? eta zenbateko zehaztasunarekin aurresan daiteke hori?"

Hala ere, eredu estatistikoak sortzerakoan hainbat arlo izan behar ditugu kontuan:

- Eredu estatistikoek errealitatea imitatzen dute modu sinplean.
- Ereduak errealitatea aurresaten laguntzeko sortzen dira.
- Eredu estatistikoek lagin bateko datuetatik abiatuta populazioaren gaineko inferentzia egitea baimentzen dute.
- Egindako inferentziaren kalitatea, erabilitako datuek, ereduek eta metodo estatistikoek zehaztuko dute.
- Datuak antzera doitzen dituzten bi ereduen artean bietatik sinpleena aukeratu behar da.
- Eredua oso ona izan arren, errealitatea hurbildu baino ez dute egiten; eta, hortaz, ezin dira erabili benetako errealitatea adieraziko balute bezala.

Edozein analisi estatistiko egin aurretik, ikerketaren helburuek eta hipotesiek finkatuta egon behar dute. Bestela, estatistikoki adierazgarriak baina zientifikoki zentzugabeak diren erlazioak aurkitzera eraman gaitezke.

2.2 Erregresio lineala

Erregresio lineal anizkoitzean $\mathbf{y}$ erantzun aldagaiaren eta $\mathbf{X}$ aldagai askeen arteko erlazio lineala aztertzen da. $\mathbf{y}$ erantzun aldagaiak jarraitua izan behar du eta banaketa normalari jarraitu behar dio. Bestalde, eragina izan dezaketen $\mathbf{X}$ aldagai askeek ere jarraituak izan behar dute.

Demagun $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ aldagai aske ditugula. Erregresio lineal anizkoitzeko eredua hau da [4]:

$$\mathbf{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

non $E(\mathbf{y}) = \mu$ eta $Var(\mathbf{y}) = \sigma^2$.

Eredua baliozkoa izan dadin honakoak bete behar dira:

- Lagin tamainak ereduaren parametro kopurua baino handiagoa izan behar du.
- X_j aldagaiek linealki askeak izan behar dute.

Izan bitez $y_1, y_2, \ldots, y_n$ n tamainko zorizko lagin bakunean behatutako $\mathbf{y}$ banaketa normala duen zorizko aldagaiaren balioak eta $X_1, X_2, \ldots, X_p$

aldagai askeak (denak jarraituak). $\beta_0, \beta_1, \ldots, \beta_p$ parametroak estimatu behar ditugu.

Parametroen estimazioa

Parametroen estimazioa egiteko bi metodo erabil daitezke: *Karratu Txikien Metodoa* eta *Egiantz Handieneko Metodoa*.

Karratu txikien metodoaren helburua erroreen karratuen batura minimizatzea da, hau da, $\sum_{i=1}^{n}(y_i - \hat{y_i})^2 = \sum_{i=1}^{n}(y_i - \sum_{j=0}^{p} x_{ij}\beta_j)^2$ funtzioa minimizatzea non $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ i. indibiduoari dagozkion behatutako balioak diren eta $\hat{y_i}$ itxarotako balioak diren $\forall j = 1, \ldots, p$.

Egiantz handieneko metodoaren helburua, berriz, $\mathbf{y}$: $N\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$ zorizko bektorearen egiantz funtzioa maximizatzea da, hots,

$$L(\beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\sum_{j=1}^{p} x_{ij}\beta_j)^2}$$

funtzioa maximizatzea [3]. Ekuazioan logaritmoak aplikatuz, ondokoa lortzen da:

$$\ln(L(\beta, \sigma^2)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

Aurreko ekuazioa maximizatzeko $\beta_0, \beta_1, \ldots, \beta_p$ eta σ^2 parametroekiko deribatu eta zerora berdinduko dugu. Horrela, egiantz handieneko estimatzaileak $\hat{\beta_0}, \hat{\beta_1}, \ldots, \hat{\beta_p}$ eta $\hat{\sigma}^2$ lortzen dira.

Oharra.- Bi metodoekin lortzen diren estimatzaileak berdinak dira.

Parametroei buruzko inferentzia

Koefizienteak estimatu ondoren $X_1, \ldots X_p$ aldagai askeak adierazgarriak diren aztertzen da, hau da, X_j aldagia barnean duen eredua X_j barnean ez duen eredua baino hobea den, $\forall j = 1, \ldots, p$. Horretarako, $\hat{\beta_j}$-rentzako Wald-en testa erabiltzen da.

$\hat{\beta_j}$-rentzako $\forall j = 1, \ldots, p$ hipotesi kontrastea honakoa da:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

Pibot estatistikoa $T_p = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{q_{jj}}} : t_{n-(p+1)}$ da non $q_{jj} = [(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}$ eta $\hat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ (σ^2-ren estimatzaile alboragabea) diren.

H_0 egia izatekotan estatistikoaren balioa honakoa da: $t_p = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{q_{jj}}}$ eta p-balioa $p = 2P(t_{n-(p+1)} > |t_p|)$.

Halaber, Wald-en testan oinarritzen den konfiantza-tartea ere erabil dezakegu X_j aldagaiaren adierazgarritasuna aztertzeko:

$$I_{\beta_j}^{1-\alpha} = (\hat{\beta}_j - t_{\frac{\alpha}{2};(n-p-1)}\hat{\sigma}\sqrt{q_{jj}}, \hat{\beta}_j + t_{\frac{\alpha}{2};(n-p-1)}\hat{\sigma}\sqrt{q_{jj}}) \qquad \forall j = 1, \ldots, p$$

Ereduaren adierazgarritasuna

Behin eredua eraikita, $\mathbf{X} = (X_1, \ldots, X_p)$ aldagai askeen adierazgarritasuna aztertu behar da. Kontraste orokorra F kontrastearen bidez egin dezakegu, non p aldagaiez osatutako eredua (Ω) eta eredu nulua (ω) alderatzen diren. Egin beharreko hipotesi kontrastea hurrengoa da:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 & (\omega) \\ H_1 : \exists j \in 1, 2, \ldots, p | \beta_j \neq 0 & (\Omega) \end{cases}$$

Test estatistikoa $F = \frac{(EKB_\omega - EKB_\Omega)/p}{EKB_\Omega/(n-(p+1))}$ da non EKB erorreen karratuen batura den (EKB=$\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$).

H_0 egia dela suposatuz, estatistikoak p eta $n - (p + 1)$ askatasun graduetako Fisherren banaketari darraio, hau da, F:$F_{p;n-(p+1)}$; eta p-balioa $p = P(F_{p;n-(p+1)} > F)$ da.

Aldiz, bi eredu konparatu nahi baditugu ($1 < q < p$), eredu habiaratuentzako kontrastea egin dezakegu. Egin beharreko hipotesi kontrastea hurrengoa da:

$$\begin{cases} H_0 : \beta_{q+1} = \cdots = \beta_p = 0 & (\omega) \\ H_1 : \exists j \in q + 1, \ldots, p | \beta_j \neq 0 & (\Omega) \end{cases}$$

Kasu honetan, test estatistikoa $F = \frac{(EKB_\omega - EKB_\Omega)/(p-q)}{EKB_\Omega/(n-(p+1))}$ da.

H_0 egia dela suposatuz, estatistikoak $p - q$ eta $n - (p + 1)$ askatasun graduetako Fisherren banaketari darraio (F:$F_{p-q;n-(p+1)}$); eta p-balioa $p = P(F_{p-q;n-(p+1)} > F)$ da.

Parametroen interpretazioa

Eredu lineala interpretatzeko, estimatutako $\hat{\beta}_j$ parametroen esanahia ulertu behar da.

- $\hat{\beta}_j$ parametroak adierazten du X_j aldagai askea unitate bat handitzean **y** menpeko aldagaian espero den gehikuntza, gainerako aldagai askeak konstante mantenduz. $\forall j = 1, \ldots, p$

- $\hat{\beta}_0$ parametroak, berriz, X_j aldagai aske guztiak zero direnean **y** menpeko aldagaian itxarotako balioa azaltzen du.

Oharra.- Beti ez da posible X_j aldagai askeak 0 izatea, eta, beraz, β_0-ren interpretazioa zentzugabea da. Kasu horretan, aldagaiak zentratu behar dira eta zentratu ondoren lorturiko konstantea da interpretatu behar dena.

1. Adibidea

bioenv.txt fitxategian itsas hondoko 30 guneko datuak jaso dira (xehetasun gehiagorako ikusi **A.1** eranskina). Datu horiekin, eredu lineala eraiki nahi dugu **d** espeziearen populazioa ahalik eta ondoen deskribatzen duena.

Hortaz, gure erantzun aldagaia (**y**) **d** espeziearen populazioa da, eta hurrengo koaldagaiak kontsideratu ditugu (jarraituak izan behar dute): gunearen sakonera (X_1), kutsadura-indizea (X_2) eta tenperatura (X_3).

Hasteko, erantzun aldagaia normala izatea behar dugu. Normalitatea aztertzeko hipotesi kontrastea egin dugu, *Shapiro-Wilk*-en testa erabiliz. Testaren p-balioa > 0.05 denez, H_0 ez dugu errefusatzen eta gure aldagaia normala da ($\alpha = 0.05$-eko adierazgarritasun mailarekin).

1. Eredu bakunak
Orain, **d** espeziearen eta aldagai askeen arteko erlazioa aztertuko dugu erregresio lineal bakuna erabiliz (Wald-en testa). Emaitzak hurrengo taulan bildu ditugu:

Aldagaiak	**p-balioa**
Sakonera (X_1)	0.0030
Kutsadura-indizea (X_2)	0.0007
Tenperatura (X_3)	0.8417

Eredu lineal anizkoitzean p-balioa $<$ 0.25 duten aldagaiak sartuko ditugu [5]; sakonera eta kutsadura-indizea, hain zuzen ere.

2. Eredu lineal anizkoitza (1 Eredua)
Sartutako aldagaiak: sakonera (X_1) eta kutsadura-indizea (X_2).

Erregresioaren kontraste orokorra eginez (F kontrastea), ereduaren p-balioa $<$ 0.05 dela lortzen dugu. Beraz, H_0 errefusatzen dugu, eta 1 Eredua eredu nulua ($\mathbf{y} = \beta_0$) baino hobea dela ondoriozta dezakegu. Gainera, sakonera eta kutsadura-indizea aldagaiak adierazgarriak dira.

Dena dela, beste eredu bat eraiki dezakegu, aurreko eredutik p-baliorik handiena duen aldagaia baztertuz, hau da, ereduan sakonera aldagaia sartu gabe (2 Eredua).

Bi ereduak konparatuz gero, p-balioa $<$ 0.05 da; beraz, H_0 errefusatzen dugu eta eredu handienarekin geratzen gara, 1 Ereduarekin, alegia.

Hortaz, lortutako eredu lineala hurrengoa da:

$$\boxed{\mathbf{y} = 6.1352 + 0.1482X_1 - 1.3877X_2}$$

Koefizienteen interpretazioa

$\beta_0 = 6.1352$ balioak adierazten digu **d** espeziearen populazioan itxarotako balioa, itsas-mailan (sakonera=0) eta kutsadura-indizea 0 bada.

$\beta_1 = 0.1482$ balioak adierazten du sakonera metro batean handitzean **d** espeziearen populazioan espero den gehikuntza, kutsadura-indizea konstante mantenduz.

$\beta_2 = -1.3877$ balioak adierazten du kutsadura-indizea unitate batean handitzean **d** espeziearen populazioan espero den txikiagotzea, sakonera konstante mantenduz.

Oharra.- Adibide honetan **d** aldagaia jarraitutzat hartu dugu, eta ez dugu kontuan izan **osoa** dela. Aurrerago ikusiko dugu Poissonen eredua aproposagoa dela datu mota horiek lantzeko.

2.3 Eredu lineal orokortuak

Erregresio lineal anizkoitza **y** erantzun aldagaia jarraitua denean baino ezin daiteke erabili, eta horrek arazoak dakartza ekologia-datuak aztertzerakoan, aldagai asko diskretuak edo kategorikoak direlako. Beraz, beste teknika batzuen beharra dugu.

Erregresio lineal anizkoitzeko eredua orokortu daiteke **y** erantzun aldagaia jarraitua ez den kasurako. Eredu horiei eredu lineal orokortuak deritzegu, eta aldagai askeak jarraituak, ordinalak edo adierazleak izan daitezke. **y** erantzun aldagaiaren arabera, hurrengo eredu lineal orokortuak kontsideratuko ditugu:

- **y** dikotomikoa (2 kategoria besterik ez) bada, datu bitarrak ditugu.
- **y** diskretu kualitatiboa (2 kategoria baino gehiago) bada, datu multinomialak ditugu.
- **y** diskretua eta ordinal finitua bada, datu ordinalak ditugu.
- **y** osoa bada, Poissonen datuak ditugu.

1 Erregresio logistikoa

Atal honetan **y** dikotomikoa deneko kasua aztertuko dugu, **y** aldagaiak bi balio baino ez dituenean hartzen, alegia.

Erregresio logistikoaren helburua **y** erantzun aldagai dikotomikoaren eta **X** aldagai askeen arteko erlazioa aurkitzea da. Kasu honetan, **y** erantzun aldagaiak 1 (arrakasta) edo 0 (porrota) balioak hartuko ditu eta $\mathbf{y} = Bin(1, p)$ non p arrakasta lortzeko probabilitatea den. **X** aldagai askeak, aldiz, jarraituak, diskretuak edo kategorikoak izan daitezke.

Demagun $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ p aldagai aske ditugula, non $p(\mathbf{X}) = E(\mathbf{y}|\mathbf{X})$ den.

Erregresio logistiko anizkoitzerako eredua ondokoa da (*logit* esteka funtzioarekin) [5]:

$$logit(p(\mathbf{X})) = ln\Big(\frac{p(\mathbf{X})}{1-p(\mathbf{X})}\Big) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

$$\Longrightarrow p(\mathbf{X}) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}$$

Izan bitez $y_1, y_2, \ldots, y_n$ n tamainako zorizko lagin bakunean behatutako

$\mathbf{y} = Bin(1, p)$ zorizko aldagaiaren balioak eta $X_1, X_2, \ldots, X_p$ aldagai askeak (jarraituak, diskretuak edo kategorikoak). $\beta_0, \beta_1, \ldots, \beta_p$ parametroak estimatu behar ditugu.

Parametroen estimazioa

Erregresio linealean bezala, parametroen estimazioa egiteko *Egiantz Handieneko Metodoa* erabiltzen da. Metodo hori

$$L(\beta) = \prod_{i=1}^{n} p(\mathbf{x}_i)^{y_i}(1 - p(\mathbf{x}_i))^{1-y_i}$$

egiantz funtzioa maximizatzean datza [5]. Egiantz handieneko logaritmoa deribatuz, $p + 1$ ekuazio lortuko ditugu. $\beta_0, \beta_1, \ldots, \beta_p$ parametroen estimatzaileak ekuazio horiek maximizatzen dituzten balioak dira.

Era berean, estimatutako balioen bariantza eta kobariantza estimatzeko egiantz handieneko logaritmoaren bigarren deribatu partzialak erabiliko ditugu:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 p(\mathbf{x_i})(1 - p(\mathbf{x_i}))$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} p(\mathbf{x_i})(1 - p(\boldsymbol{x_i}))$$

non $j, l = 0, 1, \ldots, p$.

Aurreko adierazpenetatik lortutako terminoek $(p + 1) \times (p + 1)$ tamainako matrizea osatzen dute, $I(\boldsymbol{\beta})$ izendatuko duguna. Estimatzaileen bariantza eta kobariantzak matrize horren alderantzizkoaren bidez lortuko ditugu.

Parametroei buruzko inferentzia

X_j aldagaien adierazgarritasuna aztertzen da ondoren. Horretarako, $\hat{\beta}_j$-rentzako Wald-en testa erabiltzen da $\forall j = 1, \ldots, p$.

$\hat{\beta}_j$-rentzako $\forall j = 1, \ldots, p$ hipotesi kontrastea honakoa da:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

Pibot estatistikoa $W_p = \frac{\hat{\beta}_j - \beta_j}{\sqrt{(I^{-1}(\hat{\beta}))_{jj}}} \approx N(0, 1)$ da.

H_0 egia izatekotan estatistikoaren balioa honakoa da: $w_p = \frac{\hat{\beta}_j}{\sqrt{(I^{-1}(\hat{\beta}))_{jj}}}$ eta

p-balioa $p = 2P(Z > |w_p|)$.

Halaber, Wald-en testan oinarritzen den konfiantza-tartea ere erabil dezakegu X_j aldagaiaren adierazgarritasuna aztertzeko:

$$I_{\beta_j}^{1-\alpha} = (\hat{\beta}_j - z_{\frac{\alpha}{2}}\sqrt{(I^{-1}(\hat{\beta}))_{jj}}, \hat{\beta}_j + z_{\frac{\alpha}{2}}\sqrt{(I^{-1}(\hat{\beta}))_{jj}}) \qquad \forall j = 1, \ldots, p$$

Ereduaren adierazgarritasuna

Eredua eraiki ondoren, $\mathbf{X} = (X_1, \ldots, X_p)$ aldagai askeen adierazgarritasuna aztertzen da. Kontraste orokorra egiteko egiantz-arrazoiaren testa erabil dezakegu, non p aldagaiez osatutako eredua (Ω) eta eredu nulua (ω) alderatzen diren. Egin beharreko hipotesi kontrastea hurrengoa da:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 & (\omega) \\ H_1 : \exists j \in 1, 2, \ldots, p | \beta_j \neq 0 & (\Omega) \end{cases}$$

$G = D(\omega) - D(\Omega) = -2ln\frac{\omega\text{-ren egiantza}}{\Omega\text{-ren egiantza}}$.
G-k p askatasun graduko khi karratu banaketari darraio ($G \approx \chi_p^2$); eta p-balioa $p = P(\chi_p^2 > G)$ da.

Bi eredu konparatzen baditugu ($1 < q < p$), egin beharreko hipotesi kontrastea hurrengoa da:

$$\begin{cases} H_0 : \beta_q + 1 = \cdots = \beta_p = 0 & (\omega) \\ H_1 : \exists j \in q + 1, \ldots, p | \beta_j \neq 0 & (\Omega) \end{cases}$$

Kasu honetan, $G = D(\omega) - D(\Omega) = -2ln\frac{\omega\text{-ren egiantza}}{\Omega\text{-ren egiantza}}$.
G-k $p - q$ askatasun graduko khi karratu banaketari darraio ($G \approx \chi_{p-q}^2$); eta p-balioa $p = P(\chi_{p-q}^2 > G)$ da.

Parametroen interpretazioa

Estimatutako $\hat{\beta}_j$ parametroak interpretatzeko, *Odds ratioa* (OR) erabiltzen da.

Erregresio logistikoaren ereduan: $OR_{\beta_j} = e^{\beta_j}$

- X_j **dikotomikoa denean (0 vs. 1)**

  Estatistiko honek azaltzen du zenbat aldiz handiagoa den arrakasta izateko probabilitatea $X_j = 1$ duten indibiduoen artean $X_j = 0$ dutenen artean baino, beste aldagai aske guztiengatik doituta.

- X_j **jarraitua denean**

  Estatistiko honek azaltzen du zenbat aldiz maizagoa den arrakastarako probabilitatea X_j unitate bat handitzen denean, beste aldagai aske guztiengatik doituta.

Horrela, $OR_{\beta_j} = 1$ denean, X_j-k ez du eraginik arrakasta izateko probabilitatean. Aldiz, $OR_{\beta_j} > 1$ denean, X_j-k arrakastarako probabilitatea handitzen du; eta $OR_{\beta_j} < 1$ denean, X_j-k arrakastarako probabilitatea txikiagotzen du.

Bi kasuetan, OR_{β_j}-rako konfiantza-tartea ere kalkula dezakegu. Horretarako, β_j-rentzako konfiantza-tartea kalkulatu behar dugu eta ondoren esponentzialak hartu.

2. Adibidea

Itsas hondoko 30 guneko datuak jasotzen dituen *bioenv* fitxategia (**A.1** eranskina) erabiliko dugu orain **a** espezia egoteko eragina duten ingurumen-ezaugarriak ezagutzeko eta baita haien arteko erlazioa topatzeko ere. Horretarako, erregresio logistiko anizkoitzeko eredua doituko dugu. Gure erantzun aldagaia (**y**) **a espeziearen presentzia** da. **y**=1 izango da, baldin eta **a** espeziea aurkitu bada, eta **y**=0 izango da espeziea ez badago (**y** aldagaia birkodifikatu dugu, dikotomikoa izateko). Bestalde, koaldagaiak gunearen sakonera (X_1), gunearen kutsadura-indizea (X_2), gunearen tenperatura (X_3) eta sedimentu mota (X_4) dira. Lehenengo hiru aldagaiak jarraituak dira, eta azkenengoa, kategorikoa da.

1.Eredu bakunak

Hasteko, **a** espeziearen eta aldagai askeen arteko erlazioa aztertuko dugu erregresio logistiko bakuna erabiliz. Emaitzak hurrengo taulan bildu ditugu:

Aldagaiak	**p-balioa**
Sakonera (X_1)	0.0568
Kutsadura-indizea (X_2)	0.0152
Tenperatura (X_3)	0.9960
Sedimentu mota (X_4)	0.0828*

(*) Fisher-en proba

p-balioa < 0.05 duten aldagaiak adierazgarritzat hartuko ditugu eredu logistiko bakunean; eta p-balioa < 0.25 baino txikiagoak diren aldagaiak

eredu anizkoitzerako aukeratuko ditugu [5]. Hortaz, eredu anizkoitzean aldagai guztiak sartuko ditugu, X_3 (tenperatura) aldagaia izan ezik.

2.Eredu anizkoitzak (1 Eredua)
Sartutako aldagaiak: sakonera (X_1), kutsadura-indizea (X_2) eta sedimentu mota (X_4).
Egiantz-arrazioiaren testa erabiliz, ereduaren p-balioa < 0.05 dela lortzen dugu. Beraz, 1 Eredua eredu nulua ($\mathbf{y} = \beta_0$) baino hobea dela ondoriozta dezakegu.
Hala ere, oraindik ez dugu adierazgarria den aldagairik ereduan. Hori dela eta, eredua hobetzen saia gaitezke p-balio handiena duen aldagaia (X_4) eredutik kenduz (2 Eredua).

Bi ereduak alderatzen ditugu (eredu habiaratuen alderaketa). p-balioa > 0.05 denez, H_0 ez dugu errefusatzen eta eredu txikiarekin (2 Eredua) geratuko gara.

Eredu horretan X_2 (kutsadura-indizea) aldagaia baino ez denez adierazgarria, 2 Eredua eta X_2 aldagaia baino ez duen eredua (3 Eredua) konpara ditzakegu. p-balioa > 0.05 da, eta hortaz; eredu txikiarekin geratuko gara, hots, 3 Ereduarekin.

Gainera, 3 Eredua eredu nulua baino hobea da (p-balioa$= 7.8190 \cdot 10^{-5}$).

Hortaz, erregresio logistiko bakuneko eredua geratzen zaigu:

$$\boxed{logit(p) = \frac{p}{1-p} = 7.4466 - 1.2030X_2}$$

Koefizienteen interpretazioa

Ondorioz, **a** espeziea itsas hondoan egoteko probabilitatea

$$\hat{p} = \frac{e^{\beta_0+\beta_1\times X_2}}{1+e^{\beta_0+\beta_1\times X_2}} = \frac{e^{7.4466-1.2030X_2}}{1+e^{7.4466-1.2030X_2}}$$

da. Adibidez, kutsadura-indizea 0 denean, **a** espeziea aurkitzeko probabilitatea $\hat{p} = 0.9994$ da, eta kutsadura-indizea 10 denean, aldiz, $\hat{p} = 0.0101$ da.

Odds ratioa eta odds ratioarentzako konfiantza-tartea kalkulatuko ditugu, %95eko konfiantza mailarekin.

- $OR = e^{-1.2030} = 0.3003$ da. Beraz, **a** espeziea egoteko odds-a 0.3003 aldiz biderkatzen da kutsadura-indizearen unitate bateko gehikuntza-

gatik. $OR < 1$ denez, kutsadurak **a** espeziea egoteko probabilitatea txikiagotzen du.

- Bestalde, $I^{0.95}_{-1.2030} = (-2.1745, -0.2316) \Rightarrow I^{0.95}_{OR} = (0.1137, 0.7932)$

  $1 \notin I^{0.95}_{OR}$ denez, kutsadura-indizeak eragin adierazgarria du **a** espeziea egotean (aurretik lortu dugun ondorio bera).

2 Poissonen erregresioa

Atal honetan **y** osoa deneko kasua aztertuko dugu.

y erantzun aldagai osoaren eta **X** aldagai askeen arteko erlazioa aurkitu nahi badugu, Poissonen erregresioa erabiliko dugu. Kasu honetan, **y** erantzun aldagaia kontaketa bat da eta Poissonen banaketari darraio ($\mathbf{y} \sim \mathcal{P}(\mu)$). **X** aldagai askeak, berriz, jarraituak, diskretuak edo kategorikoak izan daitezke.

Demagun $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ p aldagai aske ditugula.
Poissonen erregresio anizkoitzerako eredua ondokoa da (*log* esteka funtzioarekin) [10]:

$$log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$
$$\Longrightarrow \mu = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p} = e^{\beta_0} e^{\beta_1} \ldots e^{\beta_p}$$

non $E(\mathbf{y}) = Var(\mathbf{y}) = \mu$.

Izan bitez $y_1, y_2, \ldots, y_n$ n tamainako zorizko lagin bakunean behatutako $\mathbf{y} \sim \mathcal{P}(\mu)$ zorizko aldagaiaren balioak eta $X_1, X_2, \ldots, X_p$ aldagai askeak (jarraituak, diskretuak edo kategorikoak). $\beta_0, \beta_1, \ldots, \beta_p$ parametroak estimatu behar ditugu.

Parametroen estimazioa

Kasu honetan ere, *Egiantz Handieneko Metodoa* erabiltzen da parametroak estimatzeko. $\beta_0, \beta_1, \ldots, \beta_p$ parametroak

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$$

egiantz funtzioa [3] maximizatzen duten balioak dira.

Parametroei buruzko inferentzia

$X_1, \ldots, X_p$ aldagai askeen adierazgarritasuna aztertzeko, $\hat{\beta}_j$-rentzako

Wald-en testa erabiliko dugu $\forall j = 1, \ldots, p$, edo Wald-en testan oinarritzen den konfiantza-tartea. Egin beharreko hipotesi kontrastea erregresio logistikoan egindakoaren berdina da, eta baita erabilitako pibot estatistikoa eta p-balioa ere.

Ereduaren adierazgarritasuna

Ereduaren kontraste orokorra egiteko eta bi eredu desberdin konparatzeko erregresio logistikoan bezala jokatuko dugu, egiantz-arrazoiaren testa erabiliz eta eredu habiaratuentzako kontrastea eginez, hurrenez hurren.

Parametroen interpretazioa

Estimatutako $\hat{\beta}_j$ parametroak interpretatzeko, *arrisku erlatiboa* (RR– Relative Risk –) erabiltzen da.

Poissonen erregresioaren ereduan: $RR_{\beta_j} = e^{\beta_j}$

- X_j **dikotomikoa denean (0 vs. 1)**

  Estatistiko honek azaltzen du zenbat aldiz handiagoa den $\mathbf{y}$ gertatzeko arriskua X_j aldagaiaren presentzian ($X_j = 1$) X_j aldagaia ez dagoenean ($X_j = 0$) baino, beste aldagai aske guztiengatik doituta.

- X_j **jarraitua denean**

  Estatistiko honek azaltzen du zenbat aldiz handiagoa den $\mathbf{y}$ gertatzeko arriskua X_j unitate bat handitzen denean, beste aldagai aske guztiengatik doituta.

Horrela, $RR_{\beta_j} = 1$ denean, X_j-k ez du eraginik $\mathbf{y}$-n. Aldiz, $RR_{\beta_j} > 1$ denean, X_j-k $\mathbf{y}$ gertatzeko arriskua handitzen du; eta $RR_{\beta_j} < 1$ denean, X_j-k $\mathbf{y}$ gertatzeko arriskua txikiagotzen du.

RR_{β_j}-rako konfiantza-tartea ere kalkula dezakegu. Horretarako, β_j-rentzako konfiantza-tartea kalkulatu behar dugu eta esponentzialak hartu.

3. Adibidea

Lehenengo adibideko datu berdinak aztertuko ditugu, Poissonen erregresioa erabiliz. Erantzun aldagaia ($\mathbf{y}$) **d** espeziearen populazioa da, eta koaldagaiak hurrengoak dira: gunearen sakonera (X_1), kutsadura-indizea

(X_2) eta tenperatura (X_3).
Datu horiek aztertzeko, Poissonen erregresioa erregresio lineala baino aproposagoa da, **d** erantzun aldagaia osoa delako. Ikus dezagun ondorio berdinak lortzen ditugun.

1. Eredu bakunak
Lehenik, **d** espeziearen eta aldagai askeen arteko erlazioa aztertuko dugu Wald-en testa erabiliz. Emaitzak hurrengo taulan bildu ditugu:

Aldagaiak	**p-balioa**
Sakonera (X_1)	$1.78 \cdot 10^{-8}$
Kutsadura-indizea (X_2)	$3.39 \cdot 10^{-10}$
Tenperatura (X_3)	0.6789

Aurrekoetan bezala, eredu lineal anizkoitzean p-balioa $<$ 0.25 duten aldagaiak sartuko ditugu.

2. Eredu lineal anizkoitza (1 Eredua)
Sartutako aldagaiak: sakonera (X_1) eta kutsadura-indizea (X_2).
Egiantz-arrazoiaren testa erabiliz, ereduaren p-balioa $<$ 0.05 dela lortzen dugu. Beraz, H_0 errefusatzen dugu, eta 1 Eredua eredu nulua baino hobea dela ondoriozta dezakegu. Gainera, sakonera eta kutsadura-indizea aldagaiak adierazgarriak dira.

Hortaz, geratzen zaigun Poissonen erregresioa hau da:

$$\boxed{log(\mu) = 2.0551 + 0.0128X_1 - 0.1573X_2}$$

Koefizienteen interpretazioa

Estimatutako parametroak interpretatzeko, arrisku erlatiboa eta arrisku erlatiborako konfiantza-tartea kalkulatuko ditugu, %95eko konfiantza mailarekin.

- $RR_{\beta_1} = e^{0.0128} = 1.0129$ da. Beraz, **d** espeziearen populazioa (kantitatea) 1.0129 aldiz handitzen da sakonera metro bat handitzean (kutsadura-indizea konstante mantenduz). Gainera, $RR_{\beta_1} > 1$ denez, sakonerak **d** espeziearen populazioa handitzen du.

  Bestalde, $I^{0.95}_{\beta_1} = (0.0049, 0.0207) \Rightarrow I^{0.95}_{RR_{\beta_1}} = (1.0049, 1.0209)$

- $RR_{\beta_2} = e^{-0.1573} = 0.8544$ da. Beraz, **d** espeziearen populazioa (kantitatea) 0.8544 aldiz biderkatzen da kutsadura-indizearen unitate

bateko gehikuntzagatik (sakonera konstante mantenduz). Gainera, $RR_{\beta_2} < 1$ denez, kutsadura-mailak **d** espeziearen populazioa txikitzen du.

Bestalde, $I_{\beta_2}^{0.95} = (-0.2246, -0.090) \Rightarrow I_{RR_{\beta_2}}^{0.95} = (0.7988, 0.9140)$

Bi kasuetan, $1 \notin I_{RR}^{1-\alpha}$ denez, sakonerak eta kutsadura-indizeak eragin adierazgarria dute **d** espeziearen populazioan.

Ondorio bera lortu dugu eredu linealean, hau da, sakonerak eta kutsadura-indizeak **d** espeziearen populazioan eragin adierazgarria dutela. Gainera, bi ereduetan sakonerak **d**-ren populazioa hazten du, eta kutsadura-indizeak, ostera, txikitu egiten du. Hala ere, bietan lortutako koefizienteak desberdinak dira.

Bi arazo nagusi daude datu mota horietan erregresio lineala erabiltzearekin. Batetik, kontaketa datuetan zero kopurua handia da, eta horren ondorioz, ezin daiteke banaketa asimetriko bat banaketa normal batean transformatu (erregresio linealean beharrezkoa dena). Bestetik, eredu linealak balio negatiboak har ditzake, eta horrek praktikan ez du zentzurik, gure erantzun aldagaia diskretua eta ez-negatiboa delako.

Poissonen ereduak, berriz, suposatzen du erantzun aldagaia Poissonen banaketari darraiola, eta ez, ostera, banaketa normalari. Era berean, ereduak hartzen dituen balioak ez-negatiboak dira. Beraz, hori da eredurik egokiena kontaketa datuak tratatzeko.

3. Kapitulua

Ordenazioa eta dimentsio-murrizketa

Demagun $\mathbf{X}$ $(n \times p)$ datu-matrizean n indibiduoen balioak biltzen direla $X_1, \ldots, X_p$ p aldagaietarako. Aldagai horiek jarraituak edo kategorikoak izan daitezke, baina erregresioan ez bezala, aldagai guztiak maila berean daude, hau da, ez dago erantzun aldagairik. Gure helburua datuetako informazioa laburbiltzea eta sintetizatzea da.

Normalean, lagineko indibiduoak ez daude berez multzokatuta, p dimentsioko aldagaien espazioan sakabanatuta baino. Kapitulu honetan, indibiduoen posizioak dimentsio txikiago baten bidez deskribatzen saiatuko gara. Horrela, dimentsio horiek identifkatzeko ekologian gehien erabiltzen diren teknikak azalduko ditugu: faktore-analisia, korrespondentzia analisia eta korrespondentzia analisi kanonikoa.

3.1 Faktore-analisia

Datu-matrizeko aldagaiak jarrraituak direnean eta aldagaien artean korrelazio handia dagoenean aplikatzen da.

Faktore-analisiaren helburua datu-matrizean dagoen informazioa laburbiltzea da. Hasierako aldagaiak ezkutuan dauden aldagai kopuru txikiago batean laburbiltzen ditu, informazio galera minimoarekin. Ezkutuan dauden aldagai artifizial horiek hasierakoen konbinazio linealak dira, eta *faktore* izenaz ezagutzen dira (faktore kopurua $< p$).

Faktoreek aldagaiak ordezkatzen dituzte eta gordetzen duten informazioaren arabera ordena daitezke. Faktore bakoitzak duen informazioa neurtzeko, azaldutako bariantza-proportzioa erabiltzen da. Zenbat eta

handiagoa izan azaldutako bariantzaren zatia orduan eta handiagoa da faktore horrek jasotako informazioa. Hori dela eta, aukeratzen den lehenengo faktorea bariantzaren zatirik handiena azaltzen duena da, bigarren faktorea bariantzaren bigarren zatirik handiena azaltzen duena eta abar.

Faktore-analisiaren aplikazio nagusiak bi dira:

(i) Aldagai kopurua murriztea

(ii) Aldagaien arteko erlazioa aurkitzea, hau da, aldagaiak sailkatzea.

Horrela, indibiduoen posizioen arteko konparaketa errazagoa da eta datuen interpretazioa errazten da.

Faktore-eredua

Izan bedi $\mathbf{x}$ $(p \times 1)$ aldagaien bektorea, non indibiduo batek p aldagaitan dituen behaketak jasotzen diren.

Bektore horretarako faktore-eredua hurrengo moduan adieraz daiteke [6]:

$$\mathbf{x} = \mu + \mathbf{A} \cdot \mathbf{f} + \mathbf{u}$$

non
$-$ $\mathbf{f}$ $(k \times 1)$ faktore izendatuko dugun aldagai ezkutuen bektorea den. $\mathbf{f} : N_k(0, I_k)$ dela suposatzen dugu.

$-$ $\mathbf{A}$ $p \times k$ tamainako matrizea da eta konstante ezezagunez osatuta dago $(k < p)$. Matrizeko koefizienteak $\mathbf{f}$ faktoreek behaturiko aldagaietan duten eragina deskribatzen dute. $\mathbf{A}$ matrizeari **zama-matrize** deituko diogu.

$-$ $\mathbf{u}$ $(p \times 1)$ behatzen ez diren perturbazioen bektorea da. Faktoreez aparte, beste aldagaiek $\mathbf{x}$-n duten eragina adierazten du. $\mathbf{u} : N_k(0, \phi)$ suposatuko dugu, ϕ matrize diagonala izanik, eta perturbazioak $\mathbf{f}$ faktoreekiko korrelazio gabekoak izanik.

Eredu faktoriala, matrize-eran, horrela adierazten da:

$$\mathbf{X} = \mathbf{1} \cdot \boldsymbol{\mu}^t + \mathbf{F} \cdot \mathbf{A}^\mathbf{t} + \mathbf{U}$$

non
$\mathbf{1}$ $(n \times 1)$ bektorea, balio guztiak 1 izanik
$\mathbf{F}$ k faktoreak batzen dituen $(n \times k)$ matrizea
$\mathbf{A}^\mathbf{t}$ zama-matrizearen iraulia eta
$\mathbf{U}$ $(n \times p)$ perturbazioen matrizea diren.

Korrelazio-matrizearen azterketa

Faktore-analisia burutzeko baldintzetako bat da aldagaien arteko korrelazioak handiak izatea, hau da, korrelazio-matrizea ezin da identitatea izan. Hori egiaztatzeko metodo desberdinak daude, korrelazio partzialetan oinarrituta. Rafael Bisquerra-k [9] honako hauek aipatzen ditu: korrelazio-matrizearen determinantea aztertzea, Bartlett-en testa burutzea, Kaiser-Meyer-Olkin (KMO) indizea aztertzea edo laginaren egokitasun-neurria (MSA indizeak) aztertzea, besteak beste. Hala ere, Bartlett-en testari ematen zaio lehentasuna, estimazioetan oinarritu baino, adierazgarritasun estatistikoa neurtzen duelako.

Barttlet-en testan kontrastatzen diren hipotesiak dira $H_0 : \mathbf{R} = \mathbf{I}$ vs. $H_1 : \mathbf{R} \neq \mathbf{I}$.

Kontrasterako χ^2 estatistikoa korrelazio-matrizearen determinantetik abiatuz kalkulatzen da:

$$\chi^2 = -\left[n - 1 - \frac{1}{6}(2p + 5)\right] \cdot ln|\mathbf{R}|$$

non n datu kopurua, p aldagai kopurua eta $ln|\mathbf{R}|$ korrelazio-matrizearen determinantearen logaritmo nepertarra diren.

Estatistiko horren askatasun-graduen kopurua $\nu = \frac{1}{2}(p^2 - p)$ da.

Analisi faktoriala burutzea egokia izango da baldin eta egindako hipotesi nulua ezin badugu onartu.

Faktoreak lortu eta biratu

Behin faktore-analisia burutzea egokia dela ziurtatuta, faktore kopurua zehaztu behar dugu. Horretarako, korrelazio-matrizearen balio propioak lortuko ditugu. Faktore kopurua zehazteko erabiliko dugun irizpidea hurrengoa da: 1 baino handiagoak diren balio propioak hartzea [6].

Faktoreak lortzeko $\mathbf{R}$ korrelazio-matrizea diagonalizatzen da. Faktoreak determinatzeko, berriz, metodo ezberdinak daude, baina ekologia-datuak aztertzeko gehien erabiltzen den metodoa **Osagai Nagusietako metodoa** da. Izan ere, metodo honek ez du aldagaien normalitatea eskatzen.

Bestalde, faktoreen interpretazioa errazteko, faktoreak biratuko ditugu. Biraketa egiteko *Varimax* izeneko irizpidea erabiliko dugu. Irizpide

hori faktore bakoitzak behaturiko aldagaietan duen eraginak adierazten dituzten koefizienteen bariantza maximizatzean datza.

4. Adibidea

Adibide honen helburua *SparrowDA.txt* fitxategiko informazioa laburtzea da, non 1126 txolarreko informazio morfologikoa biltzen den (xehetasun gehiagorako, ikusi **A.2** eranskina). Aldagaiak jarraituak direnez, faktore-analisia burutuko dugu. Datu-basean dauden aldagaiak kopuru txikiago baten bidez azaltzen saiatuko gara, faktoreen bidez.

Faktore-analisia burutu aurretik, aldagaien normalitatea aztertuko dugu. Horretarako, bi test erabiliko ditugu: Kolmogorov-Smirnov Lilliefors-en testa eta Shapiro-Wilk-en testa. Lehenengoa aplikatzen hasiko gara. Emaitzak ondorengo taulan batu ditugu:

Aldagaiak	**K-S Lilliefors p-balioa**
Hegokorda	$< 2.2 \cdot 10^{-16}$
Hegolau	$< 2.2 \cdot 10^{-16}$
Tartso	$1.647 \cdot 10^{-8}$
Buru	$1.551 \cdot 10^{-14}$
Mokolum	$< 2.2 \cdot 10^{-16}$
Mokonar	$< 2.2 \cdot 10^{-16}$
Pisu	$3.338 \cdot 10^{-15}$

Aldagai guztien normalitatea errefusatzen dugu, adierazgarritasun maila 0.01 izanik. Beraz, ez dugu zorrotzagoa den Shapiro-Wilk-en testa egin beharrik.

Korrelazio-matrizearen azterketa
Orain, faktore-analisia aplikatzea egokia den erabakiko dugu. Horretarako, korrelazio-matrizea aztertu behar dugu.

Korrelazio-matrizearen determinantea= 0.0061 txikia da, baina begiratuko dugu Bartlett testa. p-balioa < 0.0001 denez, H_0 errefusatzen dugu eta beraz, korrelazio-matrizea $\mathbf{R} \neq \mathbf{I}$ da. Hortaz, egokia da faktore-analisia burutzea.

Posible da ere KMO adierazlea kalkulatzea eta bertan agertzen dira aldagai bakoitzari dagozkion MSA adierazleak kalkulatzea. KMO (overall MSA)= 0.72 handia da eta MSA adierazleak ere nahiko handiak dira; beraz, aurretik lortutako emaitzarekin bat datoz.

Faktoreak lortu eta biratu

Faktore kopurua zehazteko, irizpide gisa 1 baino handiagoak diren balio propioak aukeratuko ditugu.

```
> eigen(korrelazio.matrize)
$values
[1] 3.26485878 1.59678585 0.76531242 0.52321853 0.44111625 0.39189131
0.01681685
```

Kasu honetan, bi faktore hartuko ditugu.

Adibideko aldagai guztiak normalak ez direnez, faktoreak determinatzeko Osagai Nagusietako metodoa erabiliko dugu, faktoreak *varimax* metodoaz biratuz. Bi faktoreekin, bariantzaren %69 azaltzen dugu. Bi faktoreek azaldutako aldagai bakoitzaren bariantzaren proportzioa h2 zutabean agertzen dira. Argi dago gutxien azaldutako aldagaia *tartso* dela. Aldagai honen komunalitatea oso baxua denez (0.34) analisitik kentzea erabaki dugu.

PC1 eta PC2 zutabeetan bi faktoreen eta aldagaien arteko korrelazioak agertzen dira.

```
> ON
Principal Components Analysis
Call: principal(r = Y[, c(1, 2, 3, 4, 5, 6, 7)], nfactors = 2, rotate
= "varimax",
    scores = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
           PC1  PC2   h2    u2 com
hegokorda 0.96 0.02 0.92 0.083 1.0
hegolau   0.95 0.02 0.91 0.090 1.0
tartso    0.53 0.25 0.34 0.658 1.4
buru      0.36 0.75 0.70 0.304 1.4
mokolum   0.07 0.82 0.67 0.326 1.0
mokonar   0.05 0.85 0.73 0.269 1.0
pisu      0.66 0.40 0.59 0.408 1.7

                       PC1  PC2
SS loadings           2.67 2.19
Proportion Var        0.38 0.31
Cumulative Var        0.38 0.69
Proportion Explained  0.55 0.45
Cumulative Proportion 0.55 1.00

Mean item complexity =  1.2
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.09
 with the empirical chi square  363.2  with prob <  1.4e-73

Fit based upon off diagonal values = 0.96>
```

Faktore-analisia – *tartso* aldagaia gabe

Korrelazio-matrizearen azterketa
Berriz ere, korrelazio-matrizea aztertzen dugu. Bartlett-en testak berriro adierazten digu $\mathbf{R} \neq \mathbf{I}$ dela (p-balioa < 0.0001). Ondorioz, egokia da faktore-analisia burutzea.

Gainera, KMO (0.7) eta MSA adierazleek nahiko handiak izaten jarraitzen dute aldaketaren ostean.

Faktoreak lortu eta biratu
Korrelazio-matrizearen balio propioak kalkulatuko ditugu.

```
> eigen(kor.mat_ZUZ)
$values
[1] 3.01258691 1.58805081 0.53430846 0.44123146 0.40696837 0.01685399
```

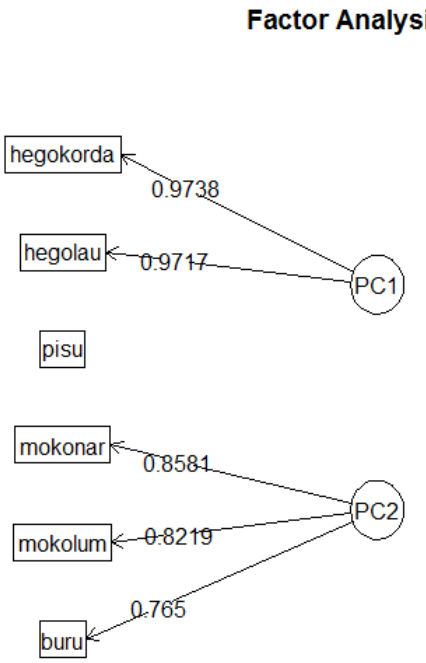
Berriz ere, nahikoa izango dugu bi faktore hartzearekin.

Osagai Nagusietako behin betiko analisia egiten dugu, faktoreak *varimax* metodoaz biratuz; eta faktore-puntuazioak gordeko ditugu grafikoak egin ahal izateko.

```
> ON_zuz
Principal Components Analysis
Call: principal(r = Y[, c(1:2, 4:7)], nfactors = 2, rotate =
"varimax",
    scores = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
           PC1  PC2   h2    u2 com
hegokorda 0.97 0.06 0.95 0.049 1.0
hegolau   0.97 0.06 0.95 0.053 1.0
buru      0.33 0.76 0.69 0.309 1.4
mokolum   0.05 0.82 0.68 0.322 1.0
mokonar   0.03 0.86 0.74 0.263 1.0
pisu      0.64 0.42 0.60 0.405 1.7

                       PC1  PC2
SS loadings           2.42 2.18
Proportion Var        0.40 0.36
Cumulative Var        0.40 0.77
Proportion Explained  0.53 0.47
Cumulative Proportion 0.53 1.00

Mean item complexity =  1.2
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.08
 with the empirical chi square  235.46  with prob <  8.8e-50

Fit based upon off diagonal values = 0.97>
```

Ikusten denez, orain, bi faktoreekin bariantzaren %77 azaltzen dugu eta komunalitate guztiak nahiko altuak dira.

Faktoreak eta jatorrizko aldagaiak lotzen dituen grafikoa hurrengoa da:



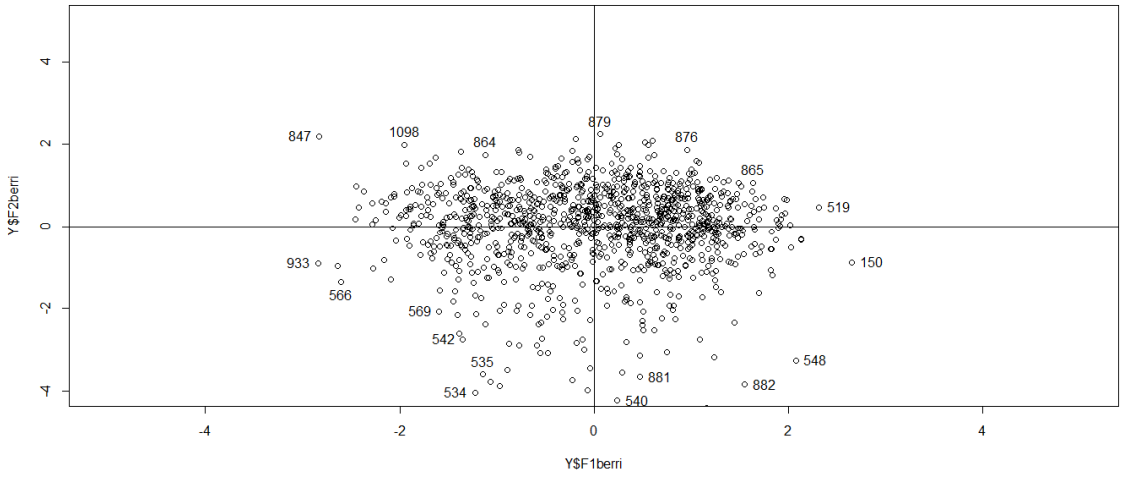
Faktoreen interpretazioa

- **Lehenengo faktorea:**
  Lehenengo faktorea, bariantzaren %40 azaltzen duena, lotuta dago gehienbat hegoen luzerarekin (*hegokorda* eta *hegolau*). Txolarren pisua (*pisu*) ere faktore honekin erlazionatuta dago, baina ez aurreko aldagaiak beste. Faktore honi "hegoen luzera" izena emango diogu.

- **Bigarren faktorea:**
  Faktore honek bariantzaren %37 azaltzen du, eta lotuta dago batez ere mokoaren luzerekin (*mokolum* eta *mokonar*) eta buruaren tamainarekin (*buru*). Txolarren pisua (*pisu*) ere faktore honekin erlazionatuta dago, baina neurri txikiagoan. Beraz, bigarren faktorea honela izendatuko dugu: "buruaren tamaina".

Faktore-puntuazioak

Indibiduoak (txolarreak) irudikatzeko, bi faktore-puntuazioak gordetzen ditugu fitxategian. Barreiadura-diagrama hurrengoa da:



Ikusten denez, behatutako txolarre gehienak grabitate zentroaren inguruan daude, hau da, txolarren hegoen eta buruen tamainak batezbestekoen inguruan murgiltzen dira. Hala ere, badaude zenbait txolarre (s566, s847 eta s933, esaterako) buruaren batez besteko tamaina baino txikiagoa dutenak. Beste txolarre batzuk, aldiz, buruaren batez besteko tamaina baino handiagoa izateagatik nabarmentzen dira, s150, s519 eta s548, kasu. Hegoen luzerari dagokionez, hainbat txolarrek batez besteko tamaina baino txikiagoa dute (adibidez, s534, s540, s881 eta s882); ostera, s864, s876, s879 eta s1098 eta txolarrek (beste batzuen artean) hegoen batez besteko luzera baino txikiagoa izateagatik bereizten dira.

Hortaz, faktore-analisia *outlierrak* detektatzeko tresna baliagarria da.

Hala ere, faktore-analisiak zenbait eragozpen ditu ekologia-datuak aztertzerakoan.

Aipatu dugunez, faktore-analisia datu-matrizeko aldagaiak jarraituak direnean eta aldagaien artean korrelazio handia dagoenean baino ezin daiteke egin. Oro har, ekologia-datuetan aztertzen diren aldagaiak ez dira denak jarraituak eta ez dute zertan korrelazionatuta egon; beraz, baldintza horiek bete ezean, teknika hau ez da oso lagungarria.

Bestalde, aldagaien arteko erlazioa lineala ez denean, faktoreek ez

dute beti datuetako informazioa modu eraginkorrean azaltzen. Horrek arazoak dakarzkie ekologia-adituei, espezieen arteko erlazioak ez-linealak izaten baitira gehienetan.

Horrez gain, batzuetan oso zaila da faktoreak interpretatzea arlo horretan aditua ez bazara. Horrela, zientifikoki zentzugabeak diren erlazioak aurkitzera eraman gaitezke.

3.2 Korrespondentzia analisia

Korrespondentzia analisia aurreko arazoak gainditzen dituen teknika deskribatzailea da. Honen helburua datu-matrizean dagoen informazioa laburbiltzea da ere, eta horretarako, hasierako aldagaiak dimentsio txikiago baten bidez deskribatzen dira. Beraz, korrespondentzia analisia faktore-analisi mota bat da, baina aldagaiak ordinalak edo kategorikoak direnean erabil daiteke. Datu ekologikoek mota honetako aldagaiak dituztenez, korrespondentzia analisia ekologia-datuak laburtzeko eta ordenatzeko oso metodo baliagarria da.

Korrespondentzia analisia kontingentzia-taula bateko bi aldagaietako kategoria ezberdinen artean dagoen erlazioa (erakarri/aldendu) hauteman eta irudikatzeko teknika bat da, bere bertsiorik sinpleenean. Era berean, aldagai bateko kategorien arteko gertutasuna edo antzekotasuna hautemateko ere erabiltzen da.

Korrespondentzia analisia burutzeko, $\mathbf{K}$ $(n \times p)$ kontingentzia-taulatik abiatuko gara non K_{ij} elementuak X aldagaiaren i modalitatea $(i = 1, \ldots, n)$ eta Y aldagaiaren j modalitatea $(j = 1, \ldots, p)$ batera aurkezten dituzten elementu kopuruak diren.

3.1. Taula: Kontingentzia-taula

X/Y	Y_1	$\ldots$	Y_j	$\ldots$	Y_p	Y-ren bazter-banaketa
X_1	K_{11}	$\ldots$	K_{1j}	$\ldots$	K_{1p}	$K_{1.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
X_i	K_{i1}	$\ldots$	K_{ij}	$\ldots$	K_{ip}	$K_{i.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
X_n	K_{n1}	$\ldots$	K_{nj}	$\ldots$	K_{np}	$K_{n.}$
X-ren bazter-banaketa	$K_{.1}$	$\ldots$	$K_{.j}$	$\ldots$	$K_{.p}$	$K_{..}$

K_{ij} kopuruen ordez, $f_{ij} = \frac{K_{ij}}{K_{..}}$ maiztasun erlatiboak kontsideratuz gero, **korrespondentzia-matrizea** deritzogun $\mathbf{P}$ matrizea lortuko dugu:

$$\mathbf{P} = \frac{1}{K_{..}}\mathbf{K} = (f_{ij})$$

non $K_{i.} = \sum_{j=1}^{p} K_{ij}$, $K_{.j} = \sum_{i=1}^{n} K_{ij}$ eta $K_{..} = \sum_{i=1}^{n}\sum_{j=1}^{p} K_{ij}$ diren.

$\begin{pmatrix} f_{i1}/f_{i.} \\ f_{i2}/f_{i.} \\ \cdots \\ f_{ip}/f_{i.} \end{pmatrix}$ bektoreari **errenkada-soslaia** deituko diogu eta $\begin{pmatrix} f_{1j}/f_{.j} \\ f_{2ji}/f_{.j} \\ \cdots \\ f_{nj}/f_{.j.} \end{pmatrix}$ bektoreari **zutabe-soslaia**.

Errenkada-soslaia izeneko bektoreek $\mathbf{R}=\begin{pmatrix} \tilde{r}_1^t \\ \tilde{r}_2^t \\ \cdots \\ \tilde{r}_n^t \end{pmatrix}=D_r^{-1}\cdot\mathbf{P}$ matrizea osatzen dute, non $D_r^{-1} = \begin{pmatrix} \frac{1}{f_{1.}} & 0 & \cdots & 0 \\ 0 & \frac{1}{f_{2.}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{f_{n.}} \end{pmatrix}$ den.

Zutabe-soslaia izenekoek, berriz, $\mathbf{C}=\begin{pmatrix} \tilde{c}_1^t \\ \tilde{c}_2^t \\ \cdots \\ \tilde{c}_p^t \end{pmatrix}=D_c^{-1}\cdot\mathbf{P}^t$ matrizea osatzen dute, non $D_c^{-1} = \begin{pmatrix} \frac{1}{f_{.1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{f_{.2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{f_{.p}} \end{pmatrix}$ den.

Soslaiak erabiltzeak modalitate guztiei garrantzia bera ematen die.

Horrela, bi puntu-hodei ditugu **N(I)** eta **N(J)**:

$$\mathbf{N}(\mathbf{I}) = \{(\tilde{\mathbf{r}}_i^t, f_{i.})|i = 1, \ldots, n\} \subset \mathbb{R}^p$$

$$\mathbf{N}(\mathbf{J}) = \{(\tilde{\mathbf{c}}_j^t, f_{.j})|j = 1, \ldots, p\} \subset \mathbb{R}^n$$

Puntuen arteko distantzia neurtzeko, khi karratu distantzia erabiltzen da, hots, haztaturiko distantzia euklidearra.

$\mathbb{R}^p$ espazioan haztapena D_c^{-1} matrizearen bidez egiten da:

$$d^2(i,i') = \|\tilde{\mathbf{r}}_i - \tilde{\mathbf{r}}_{i'}\|^2_{D_c^{-1}} = (\tilde{\mathbf{r}}_i - \tilde{\mathbf{r}}_{i'})^t D_c^{-1} (\tilde{\mathbf{r}}_i - \tilde{\mathbf{r}}_{i'}) = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

Eta $\mathbb{R}^n$ espazioan, D_r^{-1} matrizearen bidez:

$$d^2(j,j') = \|\tilde{\mathbf{c}}_j - \tilde{\mathbf{c}}_{j'}\|^2_{D_r^{-1}} = (\tilde{\mathbf{c}}_j - \tilde{\mathbf{c}}_{j'})^t D_r^{-1} (\tilde{\mathbf{c}}_j - \tilde{\mathbf{c}}_{j'}) = \sum_{j=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$$

Laburbilduz,

Puntu-hodeiak	$\mathbb{R}^p$ espazioan	$\mathbb{R}^n$ espazioan
Koordenatuak	$D_r^{-1} \cdot \mathbf{P}$ (errenkadak)	$D_c^{-1} \cdot \mathbf{P}^t$ (errenkadak)
Pisuak	D_r	D_c
Distantzia	D_c^{-1}-k haztatua	D_r^{-1}-k haztatua

Korrespondentzia analisiaren helburua da soslai-puntuen bi hodeien hurbilketa lortzea, dimentsio txikiagoko espazio baten bidez.

Horretarako, Michael Greenacre-k [7] egiaztatzen duen moduan, matrizeen balio singularretako deskonposaketan oinarritu gaitezke.

Balio singularretako deskonposaketa orokortua

$\mathbf{A}$ $(n \times p)$ matrizea emanda, horrelako deskonposaketa lor dezakegu:

$$\underset{(n\times p)}{\mathbf{A}} = \underset{(n\times k)}{\mathbf{N}} \cdot \underset{(k\times k)}{\mathbf{D}_\alpha} \cdot \underset{(k\times p)}{\mathbf{M}^\mathbf{t}}$$

non $\mathbf{N^t} \cdot \mathbf{\Omega} \cdot \mathbf{N} = \mathbf{M^t} \cdot \mathbf{\Phi} \cdot \mathbf{M} = \mathbf{I}$ diren, $\mathbf{\Omega}$ $(n \times n)$ eta $\mathbf{\Phi}$ $(p \times p)$ matrizeak simetrikoak eta positiboki definituak izanik.

Gainera, $\mathbf{M}$-ren zutabeak $\mathbf{A^t} \cdot \mathbf{\Omega} \cdot \mathbf{A} \cdot \mathbf{\Phi}$ matrizearen bektore propioak dira eta $\mathbf{N}$-ren zutabeak, $\mathbf{A} \cdot \mathbf{\Phi} \cdot \mathbf{A^t} \cdot \mathbf{\Omega}$ matrizearen bektore propioak dira, $\mathbf{D}_\lambda = \mathbf{D}_\alpha^\mathbf{2}$ balio propioei dagozkionak.

Eckart-Young teoremak [7] egiaztatzen duenez, A matrizearen k^* balio singular handienak eta beraiei dagozkien ezker- eta eskuin-bektore propioak hartuta lortzen da A matrizearen k^* heinako *hurbilketa minimo koadratikoa*. Beraz, bektore propio horiek sortutako espazioak dira bi puntu-hodeiei hobekien doitzen zaizkienak.

Korrespondentzia analisian, faktoreak lortzeko, $\mathbf{P} - \mathbf{r} \cdot \mathbf{c}^t$ (hau da, korrespondentzia-matrize zentratua) deskonposatu behar dugu.

Emaitzen interpretazioa

N(I) eta **N(J)** puntu-hodeien inertzia (edo bariantza) $\chi^2/K_{..}$ da, non χ^2 X eta Y aldagaien ateko independentzia neurtzen duen estatistikoa den. Bestalde, **N(I)** eta **N(J)** puntu-hodeiek duten propietate bat da biek inertzia berdina dutela da, eta bere balioa $\sum_k \alpha_k^2$ balio propioen batura da.

Inertzia hori faktoreetan zehar eta puntuetan zehar deskonposa daiteke, elementu ezberdinen ekarpenak definituz:

- Ekarpen absolutua ("ctr") = modalitate batek azaldutako faktorearen inertzia zatia
- Ekarpen erlatiboa ("cor") = faktore batek azaldutako modalitatearen inertzia zatia

Ekarpen absolutuak dimentsioak definitzeko erabiltzen dira. Dimentsioaren inertzia ondoen azaltzen duten puntuak (modalitateak) bilatzen ditugu. Ekarpen erlatiboek, berriz, modalitateak ondo adierazita dauden azaltzen dute. Zenbat eta handiagoak izan, orduan eta hobeto adierazita egongo dira.

5. Adibidea

bioenv.txt fitxategian itsaso hondoko 30 guneko **a, b, c, d, e** espezieen oparotasuna bilduta dago (**A.1** eranskina). Espezie horien eta guneen artean dagoen erlazioa aztertzeko korrespondentzia analisia burutuko dugu, fitxategiko lehen bost zutabeak erabiliz.

Lehenik, balio propioak kalkulatuko ditugu. Guztira 4 balio propio izango ditugu, kontingentzia taularen dimentsioa $5 - 1 = 4$ delako.

```
Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.288241  53.0  53.0  *************************
 2      0.120474  22.2  75.2  *******
 3      0.073523  13.5  88.7  *
 4      0.061395  11.3 100.0
        -------- -----
 Total: 0.543633 100.0
```

2 dimentsio (edo faktore) hartzen baditugu, inertzia osoaren %75.2 azaltzen dugu.

Lor ditzagun orain ekarpen absolutuak eta ekarpen erlatiboak:

Rows:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	s1	20	550	30	-630	500	28	-199	50	7
2	s2	40	89	20	-31	4	0	152	85	8
3	s3	20	587	24	-519	417	19	331	170	18
4	s4	13	942	69	-1621	942	123	-39	1	0
5	s5	28	984	16	289	280	8	-459	704	50
6	s6	64	636	5	147	473	5	86	163	4
7	s7	21	487	11	369	486	10	16	1	0
8	s8	2	810	9	800	300	5	-1042	509	20
9	s9	40	864	4	15	4	0	-215	860	15
10	s10	30	987	84	-1223	982	156	87	5	2
11	s11	22	529	34	-285	97	6	-602	432	65
12	s12	40	940	7	-185	359	5	235	580	18
13	s13	19	849	101	-1306	584	111	-880	265	121
14	s14	16	279	25	468	262	13	120	17	2
15	s15	19	604	26	-664	603	30	18	0	0
16	s16	53	723	66	693	714	89	77	9	3
17	s17	3	356	13	885	340	8	191	16	1
18	s18	67	937	32	-459	805	49	186	132	19
19	s19	28	640	23	-509	586	26	-154	53	6
20	s20	25	810	33	-719	714	44	263	96	14
21	s21	13	985	38	574	217	15	-1080	768	131
22	s22	53	936	81	631	481	74	-613	454	166
23	s23	31	620	46	255	80	7	-664	540	112
24	s24	43	592	21	-319	378	15	239	213	20
25	s25	61	936	41	451	550	43	377	385	72
26	s26	49	932	40	575	746	56	287	186	33
27	s27	54	659	24	403	659	30	-4	0	0
28	s28	40	155	14	-170	151	4	-27	4	0
29	s29	19	108	13	-183	93	2	73	15	1
30	s30	64	616	51	305	215	21	416	400	91

Columns:

	name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr
1	a	303	682	188	475	669	237	66	13	11
2	b	196	445	128	233	154	37	321	291	168
3	c	189	960	384	-1030	960	695	-18	0	0
4	d	245	25	95	-73	25	4	6	0	0
5	e	67	952	206	339	68	27	-1217	883	821

“Rows:”eta “Columns:”ataletan, lehenengo bi dimentsioetako (k=1 eta k=2) koordenatu nagusiak agertzen dira. Horiekin batera, ekarpen erlatiboak (“cor”) eta ekarpen absolutuak (“ctr”) agertzen dira (milako portzentajetan). Taula horietan eskaturiko korrespondentzia analisiaren

kalitatea ("qlt") ere agertzen da. Kalitatea lehenengo bi dimentsioen ekarpen erlatiboen arteko batura da.

Kasu honetan, dimentsioak edo faktoreak zehazteko zutabeak (espezieak) erabiliko ditugu. Ekarpen absolutuak dira dimentsioak interpretatzeko erabiltzen direnak:

Lehenengo dimentsioak inertzia osoaren % 53 azaltzen du. Ekarpen absoluturik handienak **a** eta **c** espezieei dagozkie. **a** espeziearen koordenatua positiboa da, eta **c**-rena, aldiz, negatiboa. Hortaz, lehenengo faktoreak **a** eta **c** espezieak banatzen ditu (**a** alde positiboan eta **c** alde negatiboan).

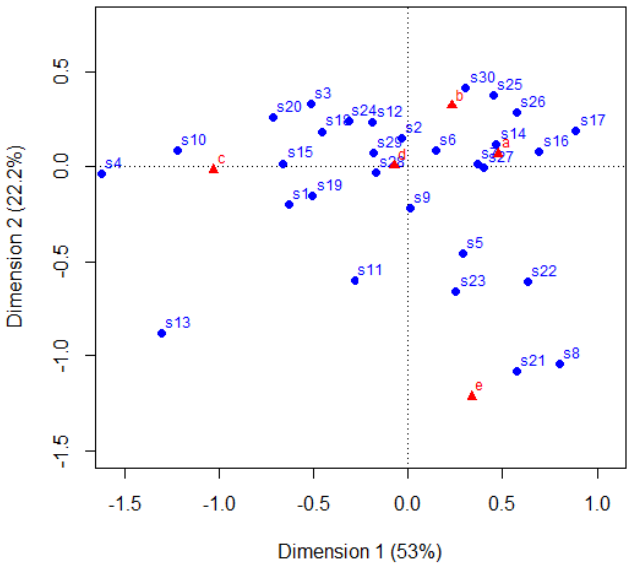
Bigarren dimentsioak inertzia osoaren % 22.2 azaltzen du. Ekarpen absoluturik handienak **b** eta **e** espezieenak dira. Faktore honek **b** eta **e** espezieak banatzen ditu: **b** alde positiboan dago eta **e** alde negatiboan.

Dimentsioak definitzeko, espezieen ekarpen absolutuak erabili ditugu. Ekarpen erlatiboek, berriz, modalitateak ondo azalduta dauden ala ez adierazten dute. Ekarpen absolutu handiak dituzten modalitateek ekarpen erlatibo handiak ere badituzte. **d** espeziea, ostera, gaizki adierazita dago lehenengo plano faktorialean, kalitate baxua baitu (qlt=%2.5). Beraz, ez dugu **d** espezieari buruzko ondoriorik aterako.

Ondoren, gune bakoitza espezieekiko nola kokatzen den aztertuko dugu. Horretarako, ekarpen erlatiboak izango ditugu kontuan eta lehenengo plano faktorialean ondo adierazita dauden guneak aztertuko ditugu soilik. Beraz, gaizki adierazita dauden s2, s14, s28 eta s29 tokiak alde batera utzi eta besteak interpretatuko ditugu soilik. Interpretazioa errazteko, itsas hondoko guneak eta espezieak adieraz ditzakegu lehenengo plano faktorialean (**3.1 Irudia**).

Itsas hondoko gune bat espezie batetik hurbil egoteak adierazten du espezie hori oparoagoa (batez beste baino handiagoa) dela gune horretan. Adibidez, s10 eta s4 guneetan **c** espeziea orokorrean baino gehiago agertzen da; s21 eta s8 guneetan, berriz, **e** espeziea; eta s30 eta s25 guneetan **b** espeziea. s14 **a** espezietik hurbil dago, baina kalitate gutxi duenez, ezin dugu **a** eta s14-ren arteko erlazioari buruzko ondoriorik atera. Esan dezakeguna da **a** espeziea s7 eta s27 guneetan aurkitzen dela batez beste baino gehiagotan.

3.1. Irudia: Itsas hondoko guneak eta espezieak lehenengo plano faktorialean



3.3 Korrespondentzia analisi kanonikoa

Korrespondentzia analisi kanonikoa indibiduoen maiztasun taula bat (kontaketak, erantzun bitarrak edo portzentajeak) eta aldagaien taula bat (kuantitatiboak, kualitatiboak edo biak) aldi berean aztertzea ahalbidetzen duen metodoa da.

Korrespondentzia analisi kanonikoa ekologian gehien erabiltzen den metodoen artean dago. Kasu honetan, maiztasun-taulan gune ezberdinetan jasotako datu biologikoak biltzen dira (adibidez, espezie baten oparotasuna edo espeziearen presentzia/gabezia) eta aldagaien taulan gune horien ingurumen-ezaugarriak jasotzen dira (pH maila, lurraren egitura, altitudea, tenperatura, etab.).

Korrespondentzia analisi kanonikoaren oinarriak

Analisiaren helburua datu biologikoen maiztasun taula aztertzea da, baina ingurumen-ezaugarriak kontuan izanda, bien arteko erlazioa topatzeko. Hau da, ingurumen-ezaugarriek azaltzen duten bariantza (inertzia)

zatia zein den jakin nahi dugu eta erlazioa interpretatu.

Hori dela eta, bi aldagaien multzoak ez ditugu simetrikotzat hartuko, hots, ez dira maila berean egongo. Datu biologikoak **erantzun aldagaiak** izango dira (erregresioan **y** aldagaia bezala), eta ingurumen-ezaugarrien aldagaiak, aldiz, aldagai askeak edo azaltzaileak (erregresioan **X** aldagaiak ziren bezala).

Beraz, korrespondentzia analisi kanonikoak bi teknika desberdin bateratzen ditu: korrespondentzia analisia eta erregresioa. Beste ordenazio metodoek bezala, korrespondentzia analisi kanonikoak faktoreak sortzen ditu non datuak proiektatzen diren. Halaber, erregresio anizkoitza erabiltzen da, ardatzak aldagai askeen konbinazio lineal moduan adierazteko.

Hortaz, korrespondentzia analisi kanonikoa korrespondentzia analisia bezalakoa da baina ardatzak aldagai askeen konbinazio linealak izanik.

Inertzia

Korrespondentzia analisi kanonikoan inertzia totala korrespondentzia analisian bezala identifikatzen da. Ingurumen-ezagugarri guztiek azaldutako inertzia balio propioen batura da. Balio propioei balio propio kanonikoak deritzegu.

Gaur egun, R-ko bi pakete erabil ditzakegu korrespondentzia analisi kanonikoa burutzeko: *Vegan* eta *ADE4*. Algoritmo eta eskala desberdinak erabiltzen dituzten arren, biekin emaitza bera lortzen dugu. Guk *Vegan* paketea erabiliko dugu.

6. Adibidea

bioenv.txt fitxategian itsaso-hondoko 30 gunetan neurtutako bost espezieen oparotasuna eta gune horien zenbait ingurumen-ezaugarri biltzen dira (**A.1** eranskina). Ingurumen-ezaugarrien eta espezieen artean dagoen erlazioa aztertzeko korrespondentzia analisi kanonikoa burutu dugu.

```
Call: cca(formula = espezie ~ Tenperatura + Sakonera + Kutsadura +
Sedimentua, data = aske)

              Inertia Proportion Rank
Total          0.5436     1.0000
Constrained    0.2490     0.4580    4
Unconstrained  0.2946     0.5420    4
Inertia is mean squared contingency coefficient

Eigenvalues for constrained axes:
   CCA1    CCA2    CCA3    CCA4
0.20083 0.03881 0.00603 0.00331

Eigenvalues for unconstrained axes:
    CA1     CA2     CA3     CA4
0.10409 0.08642 0.05703 0.04710
```

Ikusten denez, datuen inertzia (bariantza) totala 0.5436 da, eta ingurumen-ezaugarriek azaldutako inertzia 0.2490 da, hau da, inertzia totalaren %45.8 azaltzen dute.

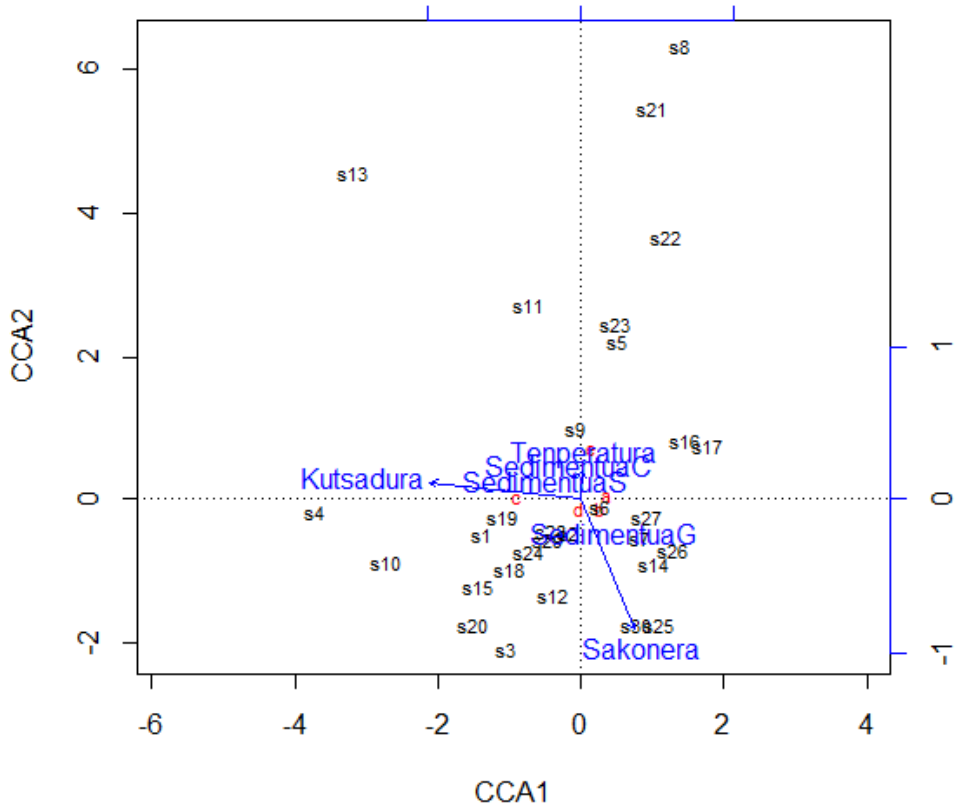
Taula horretan balio propioak ere agertzen dira (*eigenvalues for constrained axes*), dimentsioak definitzeko erabiliko ditugunak. Bi dimentsio edo faktore hartuz gero, %45.8 horren %96.24 azaltzen dugu, hau da, inertzia osoaren %44.08. Kopuru hau nahiko handia da beste ikerketa ekologikoekin konparatuz gero.

Analisiaren laburpena eskatuz gero, dimentsioen koordenatu nagusiak agertzen dira (**3.2 Irudia**). Bi dimentsio hartu ditugunez, lehen bi zutabeetan agertzen diren koordenatuei jarriko diegu arreta.

Lehenengo dimentsioa, inertzia osoaren % 36.94 azaltzen duena, lotuta dago, batez ere, kutsadura-indizearekin. Alde negatiboan batez besteko kutsadura-indizea baino handiagoa duten guneak eta espezieak egongo dira, eta alde positiboan, indize txikiagokoek.

Bigarren dimentsioa, inertzia osoaren %7.14 azaltzen duena, lotuta dago, batez ere, sakonerarekin. Alde negatiboan batez besteko sakonera baino handiagoa dutenak egongo dira, eta alde positiboan, sakonera txikiagoa dutenak.

Itsas hondoko guneak eta espezieak ingurumen-ezaugarriekiko nola kokatzen diren aztertuko dugu orain. Interpretazioa errazteko, guneak, espezieak eta ingurumen-ezaugarriak batera adieraz ditzakegu, bi dimentsioen arabera:



Guneei dagokienenez, kutsatuen dauden guneak s4 eta s10 dira. Bestalde, s8, s21 eta s22 guneak sakonera txikikoak izateagatik bereizten dira; s25 eta s30 guneak, berriz, batez besteko sakonera baino handiagoa izateagatik. s13 gunea lehenengo koadrantean dago; beraz, kutsadura-indizea batez bestekoa baino handiagoa eta sakonera batez bestekoa baino txikiagoa izateagatik nabarmentzen da.

Espezieei buruz esan dezakegu **c** espeziea kutsadura-indize altuko guneetan aurkitzen dela. **e** espeziea, ostera, sakonera gutxiko lekuetan dago. **a**, **b** eta **d** espezieak grabitate zentrotik nahiko hurbil daude; hortaz, ez dute ezaugarri berezirik.

3.2. Irudia: Espezieen, guneen eta ingurumen-ezaugarrien koordenatuak

```
Species scores

      CCA1     CCA2     CCA3     CCA4      CA1     CA2
a  0.36071  0.02857  0.09068  0.02990  0.25894  0.3250
b  0.25827 -0.13408 -0.12219  0.05214 -0.14333  0.1616
c -0.88223  0.01756  0.01725  0.03455 -0.45602 -0.0956
d -0.01031 -0.12699 -0.00265 -0.09385 -0.06306 -0.2797
e  0.13817  0.68191 -0.09104 -0.04224  0.76941 -0.6525


Site scores (weighted averages of species scores)

         CCA1     CCA2     CCA3      CCA4      CA1       CA2
s1   -1.34466 -0.50020 -1.89245 -11.00612 -0.85088 -1.680324
s2   -0.10795 -0.45913  6.33540   2.25488 -0.76514  0.178815
s3   -1.00318 -2.09813 -6.67778   0.91350 -1.98912 -0.287707
s4   -3.66922 -0.16836  2.30985   3.97615 -1.31145 -0.134338
s5    0.55008  2.20838 -0.19246  -3.82697  0.70658 -0.534659
s6    0.32786 -0.09713 -0.05433   2.66515 -0.09096  0.513638
s7    0.88186 -0.53412 -0.75913  -5.77434  0.60328  0.197976
s8    1.42669  6.34709  4.99007   1.76945  3.95298  0.807815
s9   -0.01822  0.97154  0.84540  -1.95078 -0.19034  0.004228
s10  -2.70612 -0.87397 -0.77159   2.37648  0.54120  0.681826
s11  -0.70161  2.73573 -8.53115  -1.72200  0.58567 -0.990257
s12  -0.35067 -1.33762  0.34396   0.19167 -0.35789  0.126864
s13  -3.17342  4.56041 -1.44815   4.87402  1.12301 -0.718848
s14   1.04031 -0.90359  8.70160  -6.26461  2.01798  0.513062
s15  -1.43691 -1.22288  3.20944  -7.68529  1.00686 -1.065316
s16   1.48068  0.80873  1.89230   7.50838  1.27979  1.593129
s17   1.79606  0.73598 15.02988   9.03796  4.61571  4.294089
s18  -0.99979 -0.97610  1.09468   2.28681 -1.17196  0.101381
s19  -1.09076 -0.26354 -2.34690  -7.10465 -1.40886 -1.583334
s20  -1.48791 -1.74726 -5.04367   1.47034 -1.24577 -0.261664
s21   1.01617  5.45646  1.55249  -6.54306  2.66069 -0.861313
s22   1.22473  3.66824  0.88799   0.02212  0.97758 -0.621366
s23   0.52177  2.45751 -5.05097 -10.60830  1.41724 -3.091723
s24  -0.70972 -0.73796  1.43208   6.35321 -1.53194  1.039175
s25   1.10777 -1.74716 -0.68773   0.57430  0.01349  0.366808
s26   1.31296 -0.71653  0.32440   4.78413  0.28256  1.644190
s27   0.94184 -0.24229 -1.18181  -4.18018  0.12366 -0.664342
s28  -0.37300 -0.44261  3.30327  -5.20136 -0.09533 -0.534638
s29  -0.43860 -0.57357  6.99355  -2.09283  0.22331  1.240928
s30   0.80574 -1.75550 -4.67431   3.86917 -0.49164  0.081291


Biplot scores for constraining variables

                CCA1     CCA2     CCA3     CCA4 CA1 CA2
Tenperatura  0.01419  0.18227 -0.70664 -0.68132   0   0
Sakonera     0.34931 -0.85185 -0.27964 -0.07052   0   0
Kutsadura   -0.98994  0.09845  0.08005  0.02130   0   0
SedimentuaG  0.39698 -0.40572 -0.51505  0.49458   0   0
SedimentuaS -0.33834  0.17178 -0.11693 -0.19503   0   0


Centroids for factor constraints

               CCA1    CCA2    CCA3    CCA4 CA1 CA2
SedimentuaC -0.1762  0.4842  1.2418 -0.6208   0   0
SedimentuaG  0.4706 -0.4649 -0.5935  0.5703   0   0
SedimentuaS -0.5035  0.2546 -0.1634 -0.2914   0   0
```

4. Kapitulua

Sailkapen metodoak

Lehenengo kapituluan aipatu dudan moduan, Greenacrek eta Primiceriok [1] lau multzo nagusi bereizten dituzte aldagai anitzeko tekniken artean. Kapitulu honetan, beraiek bereizten dituzten Sailkapena (analisi diskriminatzailea) eta Klusterra azalduko ditugu. Azken finean, biak erabiltzen dira indibiduoak edo aldagaiak taldekatzeko. Bien arteko desberdintasun nagusia da analisi diskriminatzailean taldeak aldez aurretik ezagutzen ditugula eta indibiduoak zein taldetan sailkatu behar diren jakin nahi dugu; kluster analisian, bestalde, ez ditugu taldeak ezagutzen, eta hori da, hain zuzen ere, zehaztu nahi duguna.

4.1 Kluster analisia

Kluster analisia edo konglomeratuen analisia indibiduoak edota aldagaiak antzeko taldetan sailkatzeko balio duen teknika estatistikoa da. Indibiduoen zenbait aldagai edo ezaugarri kontuan izanda, kluster analisiak indibiduo horiek ahalik eta homogeneoak diren taldetan sailkatzen ditu. Era berean, talde ezberdineko indibiduoak ahalik eta bereizita egotea lortu nahi da.

Ingelesez *cluster* hitzak talde edo multzo esan nahi du. Ekologian kluster analisia asko erabiltzen da animaliak eta landareak sailkatzeko, "zenbakizko taxonomia" izenez ezagutzen delarik.

Kluster analisia burutu aurretik, taldeak identifikatzeko garrantzia duten aldagaiak aukeratu behar ditugu, baita indibiduoen arteko hurbiltasun-neurria eta indibiduoak klusterretan biltzeko irizpidea ere.

Klusterrak eratzeko hainbat irizpide daude, baina irizpide guztiak distantzia- edo antzekotasun- matrize batean oinarritzen dira. Distantzia- edo antzekotasun-neurri horiek bi objektu edo banakoen arteko antzekotasun maila neurtzen dute: zenbat eta handiagoa izan banakoen arteko neurria

(distantzia), orduan eta handiagoa izango da horien arteko desberdintasuna.

Hainbat antzekotasun neurriren artean aukeratu daitekeen arren, maiz erabiltzen den neurri bat banakoen arteko **distantzia euklidear karratura** da (indibiduoak bektoretzat hartuz aldagaien espazioan). i eta j indibiduoen arteko distantzia euklidear karratura hurrengo eran definitzen da (p aldagaietarako):

$$d(x_i, x_j) = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

Metodoen sailkapena

Kluster analisirako metodoak bi talde nagusitan bana daitezke: **metodo hierarkikoak** eta **ez-hierarkikoak**. Metodo hierarkikoetan ez da talde kopurua aldez aurretik ezagutzen; eta metodo ez-hierarkikoetan, ordea, hasieratik zehaztuta dago kluster kopurua.

4.1 irudian Rafael Bisquerrak [9] proposaturiko kluster analisiko metodoen sailkapena ikus daiteke.

1 Metodo hierarkikoak

Ekologian gehien erabiltzen diren metodoak hierarkikoak dira. Izan ere, metodo horiek emaitzak irudikatzeko **dendograma** izeneko sailkapen-zuhaitza eratzea ahalbidetzen digute.

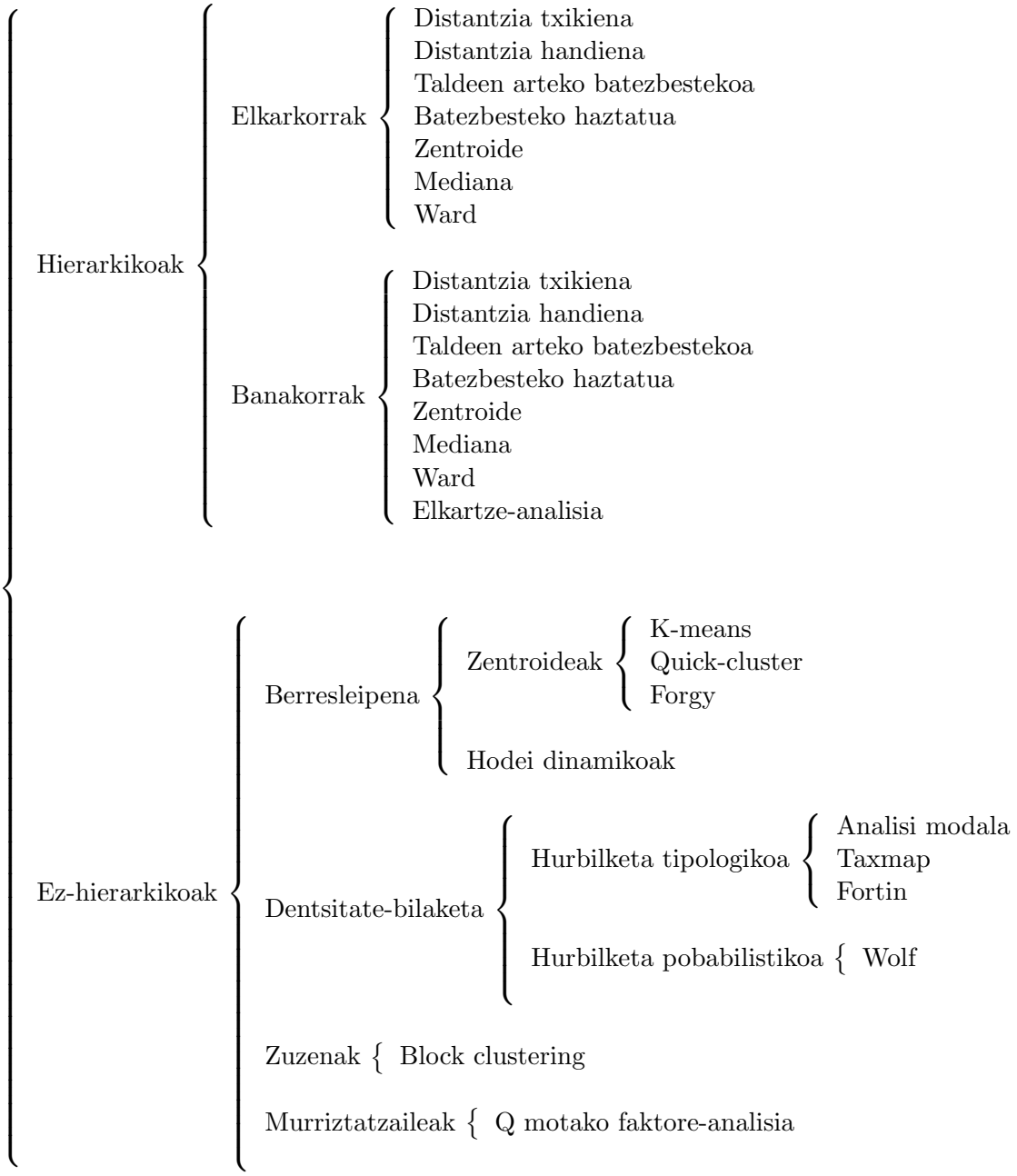
Metodo hierarkikoak bi taldetan sailkatzen dira: metodo elkarkorrak eta banakorrak.

- Metodo elkarkorretan indibiduoak beste talde daude hasieran, eta taldeak eratzen dira gorantza, indibiduoak elkartuz. Bukaeran, banako guztiak kluster bakar batean daude elkartuta.
- Metodo banakorrek alderantzizko prozesua jarraitzen dute. Hasieran indibiduoak talde bakarrean daude eta segidako zatiketen bidez, talde txikiagoak eratzen dira beherantza. Prozesuaren amaieran banakoak beste kluster daude.

Ohikoa da kluster hierarkikoa burutzea faktore- edo korrespondentzia-analisi baten ostean, lortutako faktore-koordenatuak aldagaitzat hartuta. Kasu

honetan erabiltzen den elkartze-metodoa Ward-ena edo *inertzia-galerarik txikienekoa* izaten da.

4.1. Irudia: Kluster metodoen sailkapena [9]



Ward-en metodoa

Bi konglomeratu biltzen direnean, edozein metodo erabilita ere, bariantza handiagotzen da. Ward-en metodoak taldeen barruko bariantza minimizatzea du helburu. Horretarako, lehenengo eta behin, kluster bakoitzeko aldagai guztien batezbestekoak kalkulatzen dira. Ondoren, indibiduo bakoitzaren eta bere taldearen batezbestekoaren arteko distantzia eukidearraren karratua kalkulatzen da eta distantzia guztiak batzen dira. Pauso bakoitzean lortzen diren klusterrak dira kluster barruko distantzien karratuen batura osoaren gehikuntzarik txikiena ondorioztatzen dutenak. Prozedura honek talde homogeneoak sortzen ditu.

2 Metodo ez-hierarkikoak

Metodo horien helburua indibiduoak k taldetan banatzen dituen partiketa bakarra lortzea da. Horrek eskatzen du ikertzaileak 'a priori' erabakitzea zenbat talde sortu behar diren. Bestalde, metodo ez-hierarkikoek jatorrizko datu-matrizearekin lan egiten dute, eta ez da hurbiltasun-matrizea kalkulatu behar.

Metodo ez-hierarkikoen artean, gehien erabiltzen dena *k-means* metodoa da.

K-means metodoa

K-means metodoa bereziki baliagarria da sailkatu nahi den objektu multzoa oso handia denean. Kasu honetan, dendograma handiegia izan daiteke taldeak ondo ikusi eta interpretu ahal izateko. Horrez gain, kluster analisi hierarkiko baten ostean aplika daiteke, lorturiko sailkapena egonkorra den ikusteko.

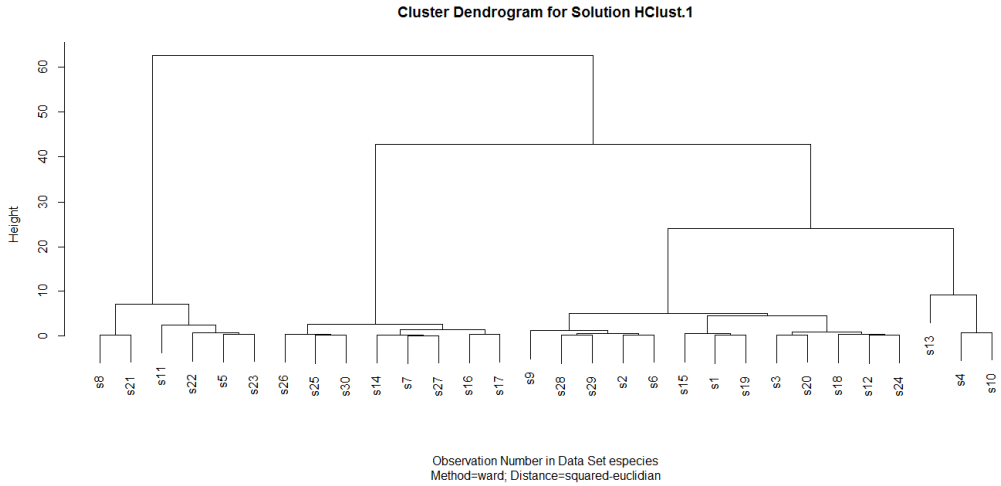
K-means metodoaren bidez, indibiduo multzoa talde homogeneoetan banatzen da eta iterazio-prozesu baten bidez sailkatzen da, taldeak egonkortu arte. Indibiduo bakoitza zentrorik hurbilena duen klusterrean egongo da. Distantzia euklidearra erabiltzen da indibiduoaren eta dagokion klusterraren zentroaren arteko hurbiltasuna lortzeko. Klusterraren zentroa da aldagai bakoitza aukeratu duen indibiduoen batezbestekoa.

5. Adibidea (Jarraipena)

Korrespondentzia analisiaren ostean, kluster analisi hierarkikoa burutu dezakegu itsas hondoko guneak sailkatzeko (espezieen oparotasunaren arabera). Horretarako, korrespondentzia analisian lortutako lehen bi koordenatuak aldagaitzat hartuko ditugu. Ward-en metodoa eta distantzia

euklidearra karratura erabiliko ditugu klusterrak eratzeko.

Hona hemen dendograma:



Dendograma ikusita, 4 kluster eska ditzakegu. 4 kluster horien laburpena honakoa da:

```
 1  2  3  4
13  3  6  8
```

Hau da, 1 klasean 13 gune daude; 2 klasean 3 gune; 3 klasean 6 gune; eta 4 klasean 8 gune.

Bestalde, kluster bakoitzaren ezaugarriak lor ditzakegu, analisian erabilitako faktore-puntuazioen zentroideen arabera:

```
INDICES: 1
        F1         F2
-0.6054478  0.2188873
------------------------------------------------------------
INDICES: 2
       F1        F2
-2.576195 -0.800184
------------------------------------------------------------
INDICES: 3
        F1         F2
 0.7029149 -2.1417443
------------------------------------------------------------
INDICES: 4
       F1        F2
0.9659978 0.5326570
```

Era berean, kluster bakoitzeko itsas hondoko guneak eska ditzakegu. Ondorengo emaitzak lortzen ditugu:

- 1 klusterra={s1, s2, s3, s6, s9, s12, s15, s18, 19, s20, s24, s28, s29}.

  1 klusterra F1 faktorearen alde negatiboan eta F2 faktorearen alde positiboan kokatzen da, baina balioak ez dira adierazgarriak; beraz, kluster honek ez du ezaugarri berezirik, hau da, batezbestekoen inguruan murgiltzen dira.

- 2 klusterra={s4, s10, s13}.

  2 klusterra bi faktoreen alde negatiboan kokatzen da, baina lehenengo faktorean solik du balio adierazgarria. Hortaz, gune horietan **c** espezie oparoa da batezbestekoarekin alderatuz.

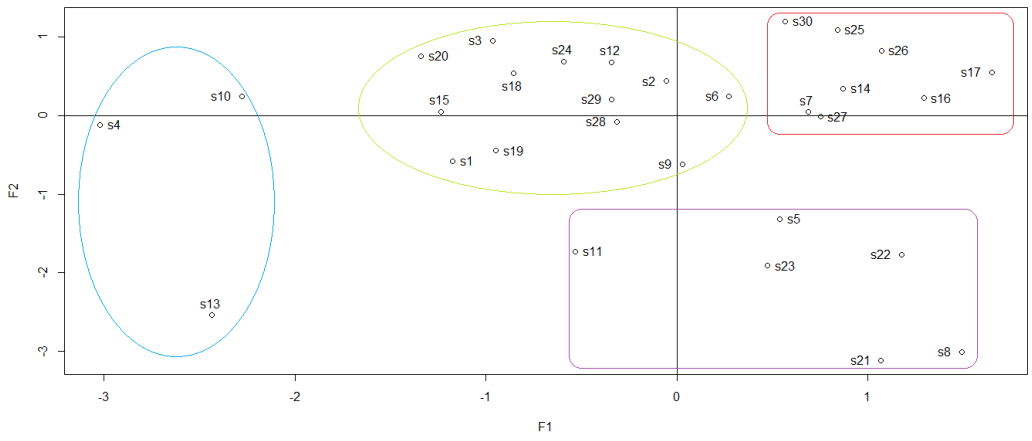
- 3 klusterra={s5, s8, s11, s21, s22, s23}.

  3 klusterra, F1 faktorearen alde positiboan eta F2 faktorearen alde negatiboan kokatzen da, baina bigarren faktorean solik du balio adierazgarria. Beraz, gune horiek **e** espeziea orokorrean baino oparoagoa izateagatik nabarmentzen dira.

- 4 klusterra={s7, s14, s16, s17, s25, s26, s27, s30}.

  4 klusterra bi faktoreen alde positiboan kokatzen da, baina balioak ez dira oso adierazgarriak. Neurri batean gune horiek **a** eta **b** espezieen oparotasunagatik bereizten dira.

Beraz, korrespondentzia analisian lortutako faktore-koordenatuen arabera, itsas hondoko guneak honela sailkatzen dira:



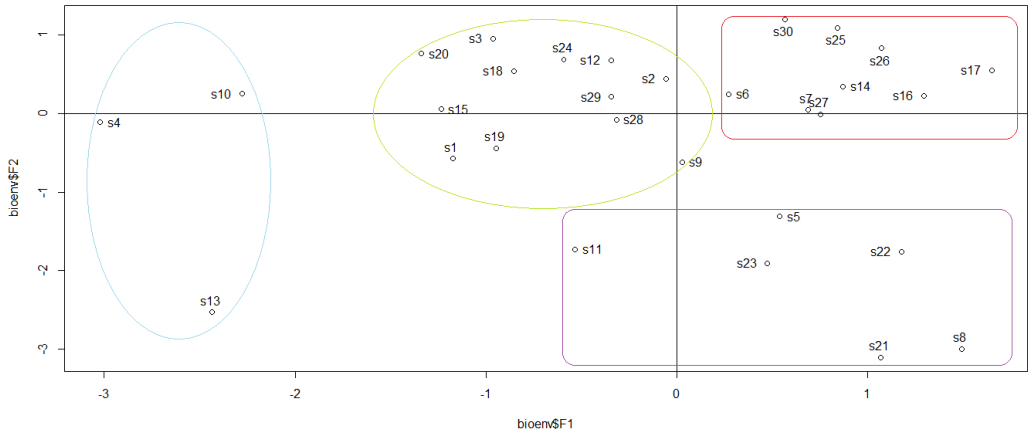
Orain, K-means metodoa aplika dezakegu aurreko emaitzen egonkortasuna aztertzeko. 4 kluster eskatuko dizkiogu.

Lortutako klusterren tamainak 9, 6, 3 eta 12 dira (aurrekoen oso antzekoak).

Eta klusterren ezaugarriak, zentroideen arabera:

```
    new.x.F1   new.x.F2
1  0.8890221  0.5010190
2  0.7029149 -2.1417443
3 -2.5761948 -0.8001840
4 -0.6786699  0.2164683
```

K-means metodoa erabilita, itsas hondoko guneak honela sailkatzen dira:



Ikusten denez, analisi hierarkikoan lortutako kluster berdinak dira eta ezaugarri berdinak dituzte, aldaketa bakarrarekin: "s6" gunea 4 klusterrean dagoela, 1 klusterrean egon beharrean.

Beraz, lortutako 4 klusterrak egonkorrak direla esan dezakegu eta sailkapen ona lortu dugula.

4.2 Analisi diskriminatzailea

Analisi diskriminatzailea talde desberdinen ezaugarriak ezagututa indibiduoak sailkatzeko erabiltzen den teknika estatistikoa da. Analisi

diskriminatzailean taldeak "a priori" zehaztuta daude eta ezagutzen dugu zein taldetan dauden indibiduoak. Informazio honekin, 'funtzio diskriminatzaile' bat edo gehiago kalkulatuko ditugu, indibiduo berrietarako iragarpenak egiteko balioko zaizkigunak.

Metodo honen bidez taldeak gehien diskriminatzen dituzten aldagaiak aztertzen dira. Ondoen diskriminatzen dituzten aldagaiak **aldagai kanoniko** deritzegun aldagai kopuru txikiago baten bidez deskribatzen dira. Aldagai kanonikoak jatorrizko aldagaien konbinazio linealak dira eta funtzio diskriminatzaile baten bidez adierazten dira.

Funtzio diskriminatzailea

Funtzio diskriminatzailearen ekuazioa erregresioaren antzekoa da. Fisher-ek (1936) asmatu zuen [9]:

$$D = b_1X_1 + b_2X_2 + \cdots + b_kX_k + b$$

D menpeko aldagaia da (talde baten partekoa izatea adierazten du), X_i aldagai askeen balioak dira (aldagai iragarle ere esaten zaie, indibiduoak taldeetan sailkatzeko balio dutelako) eta b_i datuetatik abiatuz estimatutako koefizienteak dira.

Bi talde baino ez ditugunean, funtzio diskriminatzailea erregresio anizkoitzeko ekuazioa besterik ez da, non menpeko aldagaiak 0 eta 1 balioak hartzen dituen (lehenengo taldearen parte edo bigarren taldearen parte izatea). Oro har, k talde ditugunean, $k-1$ funtzio diskriminatzaile kalkula daitezke. Funtzio guztiak ez-korrelatuak dira.

Funtzio diskriminatzileak taldeka bereiztu ahal izateko, D balioak ezberdinak izan behar dira taldearen arabera. Beraz, b_i koefizienteen balioak aukeratzen dira D balioek taldeak ahalik eta gehien bereiz ditzaten. Edo beste era batean esanda, ondoko arrazoia maximizatu behar dute [9]:

$$\frac{KB_{\text{taldeen artekoa}}}{KB_{\text{taldeen barnekoa}}}$$

non KB karratuen batura den.

Funtzio diskriminatzailean oinarrituz, ale bakoitzeko puntuazio diskriminatzailea kalkula dezakegu. Eta puntuazio horretatik abiatuz, aleak taldeetan sailkatzeko baliagarria den araua lor dezakegu. Sailkatzeko erabiltzen den araua **Bayes-en teoreman** oinarritzen da. Horren arabera, D puntuazio diskriminatzailea duen alea G_k taldearen barnekoa izateko

probabilitatea ondoko eran estima daiteke:

$$P(G_k|D) = \frac{P(D|G_k)P(G_k)}{\sum_{i=1}^{g} P(D|G_i)P(G_i)}$$

Puntuazio diskriminatzailearen arabera, ale bakoitza $P(G_k|D)$ 'a posteriori' probabilitaterik handiena duen taldean sailkatzen da.

Funtzio diskriminatzailearen adierazgarritasuna

Analisi diskriminatzailean egiten den hipotesi nulua hurrengoa da: taldeen puntuazio diskriminatzaileen batezbestekoen artean ez dago diferentzia adierazgarririk. Hipotesi estatistiko hau egiaztatzeko proba bat Wilks-en *lambdan* oinarritzen da.

Bi talde besterik ez daudenean, Wilks-en *lambda* ondokoa da:

$$\lambda = \frac{KB_{\text{taldeen barnekoa}}}{KB_{\text{osoa}}}$$

Balio horrek taldeen artean dagoen diferentziak azaltzen ez duen aldakortasun osoaren proportzioa adierazten du.

Oharra.- Wilks-en *lambda* adierazgarria izateak ez du esan nahi funtzio diskriminatzailea eraginkorra denik, baizik eta batezbestekoen arteko diferentziak daudela. Taldeen arteko diferentzia txikiak estatistikoki adierazgarriak izan daitezke, nahiz eta taldeen arteko diskriminazio onik ez baimendu. Aldiz, *lambda* adierazgarria ez bada, diskriminazioa ez da posible.

Koefiziente diskriminatzaileen interpretazioa

Koefizienteen interpretazioa erregresio-koefizienteenaren antzekoa da. Funtzio diskriminatzailearen koefiziente estandarizatuak erabiltzen dira aldagaiak estandarizatu ondoren. Koefiziente horiek aldagai bakoitzak funtzio diskriminatzaileetan duen garrantzi erlatiboaren hurbilketa ematen dute.

Funtzio diskriminatzailearen eraginkortasuna

Metodo desberdinak daude funtzio diskriminatzailearen eraginkortasuna aztertzeko. Horien artean, erabilienak hauek dira:

(i) **Nahaste-matrizea**
Matrize honetan, hasierako taldea eta aurreikusitako taldea gurutza-

tzen dira, talde bakoitzean ondo sailkaturikoen portzentajeak adieraziz. Ondo sailkatutakoen portzentaje osoak funtzio diskriminatzailearen eraginkortasuna adierazten du.

(ii) **Korrelazio kanonikoa eta *eta* koefizientea**
Korrelazio kanonikoa puntuazio diskriminatzaile eta taldeen arteko lortura-neurri bat da. *eta* koefizientearen baliokidea da, non puntuazio diskriminatzailea menpeko aldagaia eta taldea aldagai askea diren.

eta koefizientea bariantza analisitik kalkula daiteke:

$$eta^2 = \frac{KB_{\text{taldeen artekoa}}}{KB_{\text{osoa}}}$$

eta koefizienteak korrelazio mota bat adierazten du. *eta* koefizientearen karratuak taldeen arteko diferentziari dagokion aldakortasun osoaren proportzioa adierazten du.

Bi talde besterik ez daudenean, korrelazio kanonikoa puntuazio diskriminatzaile eta dagokion taldearen arteko Pearson-en korrelazio-koefizientea da.

λ eta eta^2-ren arteko erlazioa honakoa da:

$$\lambda + eta^2 = 1$$

Suposizio parametrikoak

Funtzio diskriminatzaileak talde bakoitzean aleak sailkatzean egindako erroreen probabilitatea minimizatzen du, zenbait baldintza betetzen diren heinean.

Batetik, aldagaiak jarraituak eta korrelatu gabeak izan behar dute. Horretarako, korrelazio-matrizea aztertuko dugu eta bi aldagaien arteko korrelazioa 0.9 baino altuagoa bada, bietako bat analisitik kenduko dugu [2].

Era berean, aldagaien banaketak aldagai anitzeko normala izan behar du. Lehenik, aldagai bakoitzeko azterketa deskribatzailea egin beharko genuke, datuak banaketa normalari doitzen zaizkion egiaztatzeko. Aldagairen baten banaketa normala ez bada, aldagai anitzeko normalitatearen baldintza ez da betetzen.

Azkenik, talde guztietako bariantza-kobariantzen matrizeek baliokideak izan behar dute. Hipotesi hori egiaztatzeko proba ezberdinak daude. Horietariko bat da Box-en M proba. Proba hau taldeeen bariantza-kobariantzen

matrizeen determinantean oinarritzen da.

Kontuan hartu behar da laginak handiak direnean oso erraza dela adierazgarritasun estatistikoa agertzea, nahiz eta matrizeak ezberdinak ez izan. Test hori ere oso sentikorra da aldagai anitzeko normalitatearekiko desbiderapenarekin. Hau da, matrize baliokideak adierazgarriak izango dira, aldagai anitzeko normalitaterik ez badago.

Dena dela, analisi diskriminatzailea teknika sendoa da, eta aurreko baldintzak betetzen ez badira ere, teknika baliagarria da taldeak diskriminatzeko (nahiz eta diskriminazio optimorik ez lortu).

7. Adibidea

SparrowDA.txt fitxategian 1126 txolarreko informazio morfologikoa biltzen da, 10 behatzailerengandik neurtuta (**A.2** eranskina). Behatzaile-efektua dagoen aztertu nahi dugu, hau da, behatzaileek antzeko neurketak egiten dituzten ala ez. Galdera horri erantzuteko analisi diskriminatzailea burutuko dugu. Erantzun aldagaia *behatzailea* (kategorikoa) da eta aldagai askeak *hegokorda*, *hegolau*, *tartso*, *buru*, *mokolum*, *mokonar* eta *pisu* (jarraituak) dira.

Guztira 10 behatzaile daude eta bakoitzak behatutako txolarre kopurua 11 eta 332 artean dago. Argi dago ezin ditugula konparatu 11 behaketako emaitzak 332 behaketeko emaitzekin; beraz 30 behaketa baino gutxiagoko behatzaileak analisitik kenduko ditugu, talde guztiak antzeko tamainak izan dezaten.

Lehenik eta behin, analisi diskriminatzailea aplikatzeko baldintzak betetzen diren egiaztatu behar dugu.

Aurretik dakigu aldagaien normalitatea ez dela onartzen (4. adibidea).

Horrez gain, ez da bariantza-kobariantzen matrizeen arteko berdintasun-hipotesia betetzen, Box-en M proba-ren p-balioa $< 2.2 \cdot 10^{-16}$ baita (hori gerta daiteke testa oso sentikorra delako aldagai anitzeko normalitatearekiko).

Beste baldintzetako bat da aldagaiak korrelatu gabeak izatea. Korrelazio-matrizea aztertuz gero, ikusten dugu *hegokorda* eta *hegolau* aldagaien arteko korrelazioa oso altua dela (0.9835); eta, hortaz, bi

aldagaietako bat analisitik baztertuko dugu, *hegokorda* aldagaia (hautamenez).

Suposizio parametrikoak betetzen ez badira ere, analisi diskriminatzailea burutzea erabaki dugu. Izan ere, lehen aipatu dugunez, praktikan, analisi diskriminatzailea teknika sendoa da eta ondo dabil nahiz eta hasierako baldintzak ez bete. Hala ere, horrelako kasuetan, beste aukera bat erregresio logistikoa erabiltzea da; kasu honetan, logistika multinomiala.

Lehenengo, taldeen arteko diferentziak aztertzen ditugu, Wilks-en *lambda* estatistikoa erabiliz. p-balioa < 0.05 denez, taldeen arteko diferentziak daudela ondorioztatzen dugu, hau da, behatzaile efektua dagoela. Kalkula ditzagun funtzio diskriminatzaileak.

Guztira 6 aldagai ditugu eta 8 talde (behatzaile); hortaz, gehienez, $min(6, 8-1) = 6$ funtzio diskriminatzaile kalkula ditzakegu. Aldagai estandarizatuak erabiliko ditugu koefizienteak konparatu ahal izateko. **4.2 Irudian** ikus daitezke eskuratutako emaitzak.

Lorturiko lehenengo bi funtzio diskriminatzaileak honako hauek dira:

$$D_1 = -0.0594 \cdot Z.hegolau + 0.1814 \cdot Z.tartso - 0.4593 \cdot Z.buru - 0.9945 \cdot Z.mokolum + 1.5508 \cdot Z.mokonar + 0.0075 \cdot Z.pisu$$

$$D_2 = -0.5892 \cdot Z.hegolau + 1.1076 \cdot Z.tartso - 0.4508 \cdot Z.buru + 0.2413 \cdot Z.mokolum - 0.0897 \cdot Z.mokonar - 0.0811 \cdot Z.pisu$$

Lehenengo funtzio diskriminatzaileak datuen aldakortasunaren %66.72 azaltzen du, eta bigarren funtzio diskriminatzaileak %19.60. Beraz, bi funtzioek aldakortasunaren %86.32 azaltzen dute.

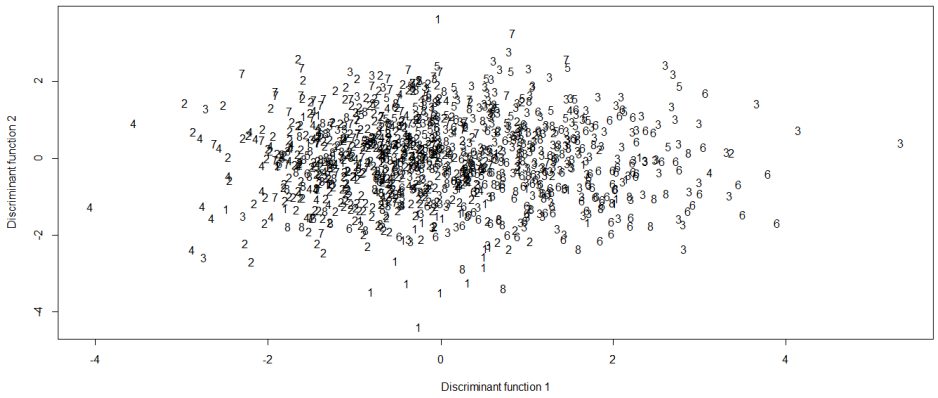
Lehenengo funtzioan *mokonar* aldagaiak koefiziente diskrimintzaile handia du; eta beraz, aldagai honek garrantzi handia du taldeak diskriminatzeko. *mokolum* eta *buru* faktoreek ere koefiziente negatibo nahiko handiak dituzte. Ematen du lehenengo funtzio diskriminatzaileak buruaren tamainaren arabera diskriminatzen dituela txolarrak.

Bigarren funtzioan, aldiz, *tartso* aldagaiak du koefiziente positibo handia, beraz, aldagai horrek taldeak diskriminatzeko eragin handia du. *hegolau* eta *buru* faktoreek ere koefiziente negatibo nahiko handiak dituzte.

4.2. Irudia

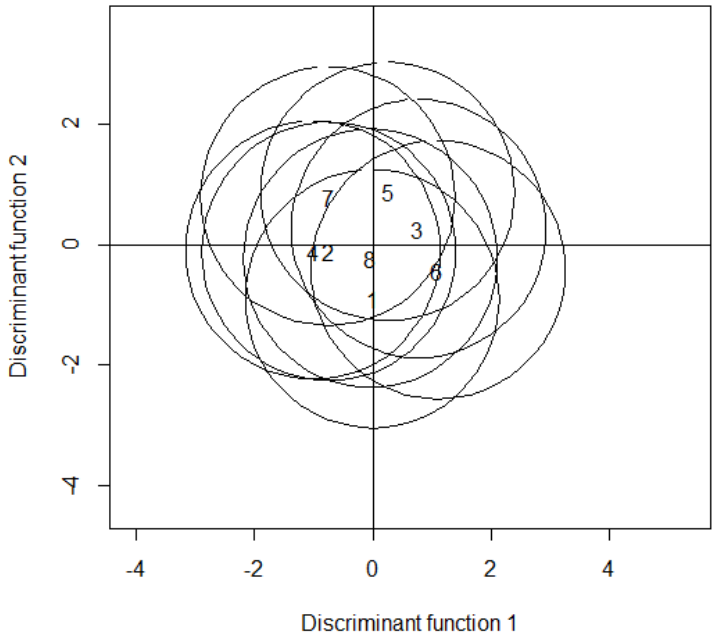
```
Call:
lda(behatzaile ~ Z.hegolau + Z.tartso + Z.buru + Z.mokolum +
    Z.mokonar + Z.pisu, data = Y_std)

Prior probabilities of groups:
         1          2          3          4          5          6
0.05081670 0.30127042 0.24591652 0.06624319 0.04900181 0.12250454
         7          8
0.06079855 0.10344828

Group means:
     Z.hegolau    Z.tartso      Z.buru    Z.mokolum   Z.mokonar
1  0.422259663 -0.62127242 -0.06356423  0.007659186  0.07358653
2  0.119645575 -0.06643853  0.05252948  0.104614692 -0.38045110
3 -0.102956735  0.27653451  0.03463864 -0.214582949  0.33677612
4 -0.063904942 -0.19571137  0.14086965  0.604935104 -0.19744123
5 -0.472485384  0.29573402 -0.52306847  0.116857956  0.03045256
6  0.009939008 -0.31583995 -0.06009229 -0.065396933  0.68947971
7 -0.047966788  0.49957006 -0.21379775  0.265306926 -0.43853953
8 -0.029967690 -0.09304835  0.15027947 -0.319531434 -0.17549342
        Z.pisu
1 -0.234093017
2  0.005569696
3  0.006000374
4  0.103040283
5 -0.201506399
6  0.136394330
7 -0.248769870
8  0.098664211

Coefficients of linear discriminants:
                    LD1         LD2         LD3         LD4         LD5
Z.hegolau -0.059443623 -0.58915156  0.2976403 -0.93934921 -0.1185005
Z.tartso   0.181410810  1.10762115 -0.1729139 -0.28691554 -0.2025169
Z.buru    -0.459278072 -0.45075908 -1.0178028  0.06470805 -0.6562522
Z.mokolum -0.994507324  0.24129661  0.9078239  0.20131190 -0.3198923
Z.mokonar  1.550784417 -0.08965449  0.4020494 -0.11109771 -0.2192435
Z.pisu     0.007525027 -0.08112482 -0.1680504  1.02503380  0.3467066
                   LD6
Z.hegolau -0.46474765
Z.tartso  -0.22002767
Z.buru     0.48015523
Z.mokolum -0.01120544
Z.mokonar  0.02226689
Z.pisu    -0.71639339

Proportion of trace:
   LD1    LD2    LD3    LD4    LD5    LD6
0.6672 0.1960 0.0920 0.0354 0.0078 0.0016
```

Funtzio diskriminatzaileak estimatu ondoren, horien puntuazio diskriminatzaileak kalula ditzakegu *predict()* metodoaren bidez. Lehenengo bi funtzio diskriminatzaileen puntuazioak irudikatuko ditugu, non indibiduoak (behaketak) behatzailearen zenbakiarekin adierazita dauden:



Aurretik aipatu dugu behatzaile efektua dagoela; beraz, taldeak sakabanatuta egon beharko lirateke. Hala ere, irudian taldeak ez dira ondo bereizten.

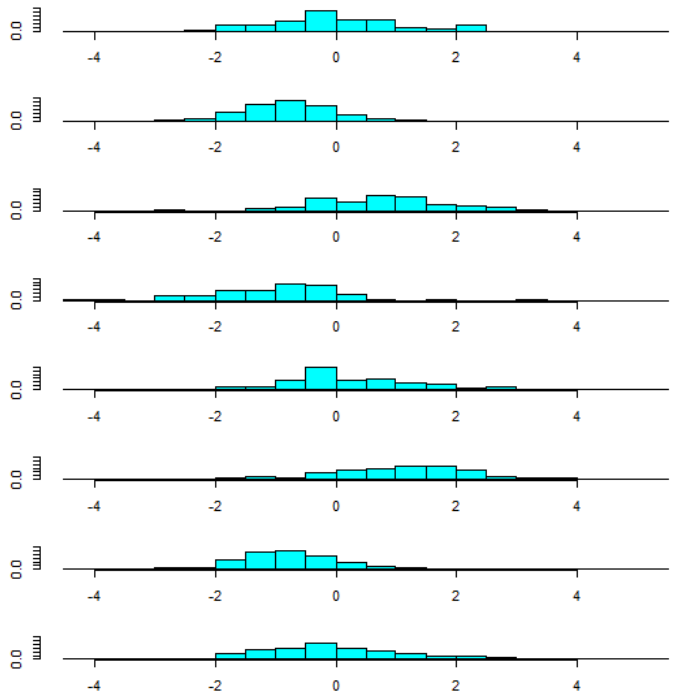
Barreiadura-diagrama egin beharrean, konfiantza-tarteak erabil ditzakegu:



Bertan taldeen batezbestekoak ageri dira, behatzailearen zenbakiarekin

adierazita. Taldeen inguruko zirkunferentziak %90eko konfiantza tarteko eremuak adierazten dituzte, hau da, talde bakoitzeko populazioaren %90 multzo horren barruan egotea espero da. Grafiko honen bidez ere nabari dezakegu taldeak ez daudela bereizita.

Horrez gain, puntuazio diskriminatzaileen histogramak irudika ditzakegu. Adibidez, lehenengo funtzio diskriminatzailearenak:



Hemen ere ikusten da taldeak ez daudela oso diskriminatuta, talde guztiek [-4, 4] tarteko balioak hartzen baitituzte.

Ondorioz, behatzaile-efektua badago ere, hau ez da oso esanguratsua.

Analisi diskriminatzailearen eraginkortasuna ondoko nahaste-matrizearen bidez azter dezakegu:

observer	1	2	3	4	5	6	7	8
1	8	29	10	0	0	7	1	1
2	1	283	35	0	2	3	5	3
3	2	67	168	0	1	31	2	0
4	0	63	3	3	1	3	0	0
5	0	20	23	0	7	4	0	0
6	1	29	57	0	0	45	0	3
7	0	46	13	0	1	0	7	0
8	0	61	39	0	0	6	0	8

Ikusten denez, lehenengo behatzaileak behatutako 56 txolarretatik 8 baino ez dira zuzen sailkatu, besteak gainerako taldeetan sailkatu dira (1 taldekoak izanik): 29 bigarren taldean, 10 hirugarren taldean, 7 seigarren taldean, eta txolarre bana 7 eta 8 taldean. Beraz, lehenengo taldeko populazioaren %14.29 (8/56) baino ez da zuzen sailkatu.

Era berean, bigarren behatzaileak behatutako 332 txolarretatik 283 ondo sailkatu dira, hau da, bigarren taldeko indibiduoen %85.24 ongi sailkatu da.

Hirugarren behatzaileak behatutako 271 txolarretatik 161 ondo sailkatu dira, hau da, talde horretako indibiduoen %61.99.

Laugarren behatzaileak behatutako 73 txolarretatik 3 baino ez dira zuzen sailkatu, hau da, laugarren taldeko indibiduoen %4.11.

Bostgarren behatzaileak behatutako 54 txolarretatik 7 ondo sailkatu dira, hots, talde horretako indibiduoen %12.96 ongi sailkatu da.

Seigarren behatzaileak behatutako 135 txolarretatik 45 ondo sailkatu dira, hau da, talde horretako indibiduoen %33.33.

Zazpigarren behatzaileak behatutako 67 txolarretatik bakarra sailkatu da ondo, hau da, talde horretako indibiduoen %10.45.

Zortzigarren behatzaileak behatutako 114 txolarretatik 8 ondo sailkatu dira, hau da, talde horretako indibiduoen %7.02.

Guztira, indibiduoen %48 baino ez da ongi sailkatu.

2, 3 eta 6 taldeetako portzentajeak nahiko altuak dira; beraz, badirudi talde horietan diskriminazioa posible dela; eta ondorioz, behatzaile-efektua dagoela. Gainerako taldeen portzentajeak, aldiz, baxuak dira; hortaz, ez da behatzaile-efekturik antzematen.

5. Kapitulua

Ondorioak

Ekologiak izaki bizidunen eta ingurumenaren arteko erlazioak kuantifikatzen ditu, zentzu horretan, zientzia kuantitatiboa da. Kantitate horiek aztertuz, bizi garen mundu konplexu honi buruzko galderei erantzuna eman nahi die. Ekologia-aditu, biologo eta beste ingurumen-ikerlari gehienek jakitun dira datu ekologikoak aztertzearen zailtasunez, baina gutxik menderatzen dituzte teknika estatistikoak arazo horiei aurre egin ahal izateko. Memoria honetan aldagai anitzeko analisiak ekologia-datuak aztertzeko guztiz erabilgarria izan daitekela ikusi dugu, eta datuek dakartzaten zaitasunak gainditzen dituztela. Beraz, nabaria da ekologian teknika estatistikoen premia dagoela.

Era berean, ekologoek sarritan beren buruari egiten dioten galdera hauxe da: "Zein metodo aplikatu behar dut?". Galdera horri erantzuna ematea izan da lan honen helburu nagusietako bat. Argi dago erantzuna aztertu nahi denaren araberakoa dela: Zer erakutsi nahi dugu? Zeintzuk dira ikerketaren azpian dauden galderak? Horrek emango digu metodorik egokiena aukeratzeko oinarria.

Horrez gain, memoria honetan zehar garaturiko adbideen bidez, metodo estatistikoak aplikatzeko jarraitu beharreko urratsak azpimarratu ditugu: (i) ikerketaren helburuak eta hipotesiak ongi finkatu; (ii) metodo estatistiko egokiena aplikatu, (iii) eredua ondo doituta dagoela ikusi eta (iv) interpretazio ekologikoa egin (nahiz eta azken erabakia ekologia-adituen esku utzi).

Aldagai anitzeko teknikei esker ekologoek jasotako datu-base neurrigabeak labur daitezke eta parametro kopuru txikiagoen bidez azaldu, interpretazioa errazteko. Lan honetan datu ekologikoak aztertzeko aldagai anitzeko analisiko hainbat teknika aurkeztu ditugu, bakoitzaren onurak eta desabantailak azpimarratuz. Adibidez, ikusi dugu erregresio lineal anizkoitzak, faktore-analisiak eta analisi diskriminatzaileak guztiz mugatzen

dutela analiza daitezken datu-ekologikoak, aldagaiak jarraituak izan behar baitute; eta oro har, ekologia datuetan mota guztietako aldagaiak biltzen dira (jarraituak, diskretuak edo kategorikoak). Korrespondentzia analisia eta kluster analisia, aldiz, edozein motako aldagaiekin aplika daiteke; beraz, ekologia-adituentzako oso tresna erabilgarriak dira.

Lan honetan garaturiko adibideetan, datu berak era ezberdinetan aztertzeko aukera landu dugu. Izan ere, datu multzo batek galdera bat baino gehiago plantea dezake eta ahal den informazio gehien lortzeko metodo bat baino gehiago aplikatzea beharrezkoa izan daiteke. Memoria honetan bi datu-base besterik ez ditugu erabili eta zortzi teknika estatistiko aplikatu ditugu horietan. Beraz, argi dago batzuetan ez dagoela metodo bakarra datuak analizatzeko, edo bi metodo ezberdin erabil ditzakegula era osagarrian (adibidez, korrespondentzia analisia eta kluster analisia, faktore-analisia eta kluster analisia, analisi diskriminatzailea eta erregresio logistikoa, etab.).

Bestalde, lan honen bidez, matematikarien eta ekologia-adituen arteko elkarlanaren beharra eta aukera azpimarratu nahi izan dut, existitzen den elkarlanerako bidean sakontzeko beharra. Izan ere, elkarlan hori ezinbestekoa da bi aldeentzako onuragarriak diren ikerketa zientifikoak burutzeko. Batetik, ekologia-adituek planteatutako galderak objektibotasunez erantzuteko oinarria lortzen dute; baina, bestalde, estatistikariek ikerketa-arlo berriak esploratzeko aukera dute.

Memoria honetan ekologia-datuak aztertzeko erabiltzen diren aldagi anitzeko analisiko teknikarik ‘ezagunenak’ aurkeztu ditugu. Horietako batzuk Matematika Graduko ‘Aldagai Anitzeko Analisia’ irakasgaian ikusitakoak dira. Beraz, proiektu honek graduan ikusitako kontzeptu teorikoak jorratzeko aukera eman dit, eta baita aldagai anitzeko analisiak errealitatean duen aplikazioa ikusteko ere. Beste teknika batzuk (Poissonen erregresioa, korrespondentzia analisi kanonikoa edota analisi diskriminatzailea), ostera, ez ditugu karreran landu, baina gai izan naiz jasotako prestakuntzan oinarrituta lan honetan garatzeko.

Hortaz, memoria hau burutzea oso esperientzia aberasgarria iruditu zait eta nire etorkizun profesionalerako baliagarria izango delakoan nago, barneratutako ezagutza tekniko guztiagatik zein proiektuaren kudeaketari dagokionez hartutako eskarmentuagatik. Zer esanik ez, hau gaiaren sarrera orokor bat besterik ez da izan, eta teknika interesgarri asko geratu dira kanpoan.

Bibliografia

[1] Greenacre, Michael & Primiceiro, Raul. *Multivariate Analysis of Ecological Data*. Fundación BBVA, 2013.

[2] Zuur, Alain F.; Ieno, Elena N. & Smith, Graham M. *Analysing Ecological Data*. Springer Science & Business Media, 2007.

[3] Faraway, Julian J. *Extending the Linear Model with R*. Chapman & Hall/CRC, 2005.

[4] Faraway, Julian J. *Linear Models with R*. Chapman & Hall/CRC, 2009.

[5] Hosmer, David W. & Lemeshow, Stanley. *Applied Logistic Regression*. John Wiley & Sons, Inc., 2000.

[6] Peña, Daniel. *Análisis de Datos Multivariantes*. Mc Graw-Hill, 2002.

[7] Greenacre, Michael. *Theory and aplication of Correspondence Analysis*. London Academic Press, 1984.

[8] Kostov, Belchin Adriyanov. *Aportación del análisis canónico de correspondencias al análisis textual*. Trabajo fin de carrera, Universitat Politècnica de Catalunya, 2006. (https://upcommons.upc.edu/bitstream/handle/2099.1/6539/Memoria.pdf?sequence=1)

[9] Bisquerra Alzina, Rafael. *Introducción Conceptual al Análisis Multivariable*. PPU, Barcelona, 1989.

[10] STAT 504. Analysis of Discrete Data. Lesson 9: Poisson Regression. The Pennsylvania State University, 2015. (https://onlinecourses.science.psu.edu/stat504/node/168)

[11] Borcard, Daniel; Gillet, François & Legendre, Pierre. *Numerical Ecology with R*. Springer Science & Business Media, 2011.

[12] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing 2013. (http://www.R-project.org).

A. Eranskina

Datu-baseak eta R-ko kodea

Memoria honetako adibideak garatzeko bi datu-base erabili ditugu: *bioenv* eta *SparrowDA* fitxategiak.

A.1 *bioenv*

Fitxategi honetan itsas hondoko 30 guneko datuak jaso dira [1]. Datu horiek gune bakoitzean dagoen **a**, **b**, **c**, **d** eta **e** espezieen indibiduo kopurua, gunearen sakonera (metrotan), tenperatura (℃-an), kutsadura-indizea eta sedimentu mota erakusten dute.

Hiru sedimentu mota daude $C = buztina/lokatza$, $S = hondarra$ eta $G = legarra/harria$. Horrez gain, kutsadura indizea metal astunen (barioa, kadmioa eta beruna, besteak beste) kontzentrazioan oinarritzen da. Zenbat eta altuagoa izan indizea, orduan eta handiagoa da kutsadura-maila.

Bost espezieak aldagai diskretuak dira; sakonera, tenperatura eta kutsadura-indizea aldagai jarraituak dira; eta sedimentu mota kategorikoa da.

Datu-base txikia da, bost espezietako datuak baino ez baitira jaso (normalean, ekologian ehunka espezie aztertzen dira). Gune kopurua, aldiz, errealagoa da.

A.1. Taula: *bioenv* datu-basea [1]

	a	***b***	***c***	***d***	***e***	***Kutsadura***	***Sakonera***	***Tenperatura***	***Sedimentua***
s1	0	2	9	14	2	4,8	72	3,5	S
s2	26	4	13	11	0	2,8	75	2,5	C
s3	0	10	9	8	0	5,4	59	2,7	C
s4	0	0	15	3	0	8,2	64	2,9	S
s5	13	5	3	10	7	3,9	61	3,1	C
s6	31	21	13	16	5	2,6	94	3,5	G
s7	9	6	0	11	2	4,6	53	2,9	S
s8	2	0	0	0	1	5,1	61	3,3	C
s9	17	7	10	14	6	3,9	68	3,4	C
s10	0	5	26	9	0	10,0	69	3,0	S
s11	0	8	8	6	7	6,5	57	3,3	C
s12	14	11	13	15	0	3,8	84	3,1	S
s13	0	0	19	0	6	9,4	53	3,0	S
s14	13	0	0	9	0	4,7	83	2,5	C
s15	4	0	10	12	0	6,7	100	2,8	C
s16	42	20	0	3	6	2,8	84	3,0	G
s17	4	0	0	0	0	6,4	96	3,1	C
s18	21	15	33	20	0	4,4	74	2,8	G
s19	2	5	12	16	3	3,1	79	3,6	S
s20	0	10	14	9	0	5,6	73	3,0	S
s21	8	0	0	4	6	4,3	59	3,4	C
s22	35	10	0	9	17	1,9	54	2,8	S
s23	6	7	1	17	10	2,4	95	2,9	G
s24	18	12	20	7	0	4,3	64	3,0	C
s25	32	26	0	23	0	2,0	97	3,0	G
s26	32	21	0	10	2	2,5	78	3,4	S
s27	24	17	0	25	6	2,1	85	3,0	G
s28	16	3	12	20	2	3,4	92	3,3	G
s29	11	0	7	8	0	6,0	51	3,0	S
s30	24	37	5	18	1	1,9	99	2,9	G

A.2 *SparrowDA*

SparrowsDA fitxategian 1126 txolarreko informazio morfologikoa biltzen da, 10 behatzailerengandik neurtuta (Chris Elphick, University of Connecticut, USA) [2]. Neurtutako aldagaiak honako hauek dira: hegoaren luzera, bi modutan neurtuta (*hegokorda* eta *hegolau*); tartsoaren luzera (*tartso*); buruaren tamaina (*buru*); mokoaren luzera, lumak hasten direnera arte (*mokolum*); mokoaren luzera, narinetara arte (*mokonar*) eta pisua (*pisu*).

Datu-basea oso luzea denez, ez dugu hemen erakutsiko. Hala ere, ondoko estekan eskura daiteke:

http://highstat.com/book1.htm (14. kapitulua)

A.3 R-ko kodea

Jarraian lanerako adibideak garatzeko R-ko kodea dago. Erabili ditugun paketeak *nortest*, *psych*, *ca*, *vegan*, *Rcmdr*, *MASS*, *biotools* eta *ade4* dira.

```
############################################################
#####                 GRADU AMAIERAKO LANA             #####
############################################################


#-----------------------------------------------------------#
#####           2. KAPITULUA - ERREGRESIOA             #####
#-----------------------------------------------------------#

###         1. Adibidea - Erregresio Lineala              ###

#Datuak irakurri
bioenv <- read.table(file="bioenv.txt", header=TRUE, dec=",", sep="\t")

#Laburpena egin
summary(bioenv)

#Aldagaiak zuzenean irakurtzeko
attach(bioenv)

#Normalitatearen azterketa
shapiro.test(d)
```

```
#Eredu bakunak doitu
ej1<-lm(d~Sakonera)
summary(ej1)

ej2<-lm(d~Kutsadura)
summary(ej2)

ej3<-lm(d~Tenperatura)
summary(ej3)

#Eredu anizkoitza doitu
eredua1<-lm(d~Sakonera+Kutsadura)
summary(eredua1)

#Ereduen konparaketa
anova(eredua1,ej2)

###        2. Adibidea - Erregresio Logistikoa           ###

#a espeziea dikotomiko bihurtu
a<-ifelse (a > 0, 1, 0)

#Eredu bakunak doitu
ej1<-glm(a~Sakonera, family=binomial)
summary(ej1)

ej2<-glm(a~Kutsadura, family=binomial)
summary(ej2)

ej3<-glm(a~Tenperatura, family=binomial)
summary(ej3)

ej<-glm(a~Sedimentua, family=binomial)
summary(ej)

#Kontingentzia-taula eta khi karratu testa
Table <- xtabs(~a + Sedimentua)
Table
fisher.test(Table)

#Eredu anizkoitzak doitu
eredua1<-glm(a ~ Sakonera + Kutsadura + Sedimentua, family=binomial)
summary(eredua1)
```

```
## Egiantz-arrazoiaren testa
    # deviance funtzioak doitutako ereduaren deviance ematen digu.
    # df -> Askatasun graduak dira, bi ereduen arteko parametroen diferentzia
pchisq(eredua1$null.deviance-deviance(eredua1),df=4, lower=FALSE)

#Aldagai guztiak ez dira adierazgarriak
#p-baliorik handiena duen aldagaia eredutik kentzen dugu:
eredua2<-glm(a ~ Sakonera + Kutsadura, family=binomial)
summary(eredua2)

#Ereduen arteko konparaketa deviance-aren bitartez (egiantz-arrazoiaren testa)
pchisq(deviance(eredua2)-deviance(eredua1),
df.residual(eredua2)-df.residual(eredua1), lower=FALSE)

#Edo anova funtzioa erabiliz
anova(eredua1,eredua2, test="Chi")

#p-baliorik handiena duen aldagaia eredutik kentzen dugu:
eredua3<-glm(a ~ Kutsadura, family=binomial)
summary(eredua3)

#Ereduen arteko konparaketa deviance-aren bitartez (egiantz-arrazoiaren testa)
pchisq(deviance(eredua3)-deviance(eredua2),
df.residual(eredua3)-df.residual(eredua2), lower=FALSE)

#Egiantz-arrazoiaren testa
pchisq(eredua3$null.deviance-deviance(eredua3),df=1, lower=FALSE)

#Edo anova funtzioa erabiliz
anova(eredua3, test="Chisq")

#Odds ratioaren eta konfiantza-tartearen kalkulua
exp(coefficients(eredua3))#OR
confint.default(eredua3)
exp(confint.default(eredua3))#Konfiantza-tartea

###        3. Adibidea - Poissonen erregresioa          ####

#Eredu bakunak doitu
ej1<-glm(d~Sakonera, family=poisson)
summary(ej1)

ej2<-glm(d~Kutsadura, family=poisson)
summary(ej2)
```

```
ej3<-glm(d~Tenperatura, family=poisson)
summary(ej3)

#Eredu anizkoitza doitu
eredua1<-glm(d~Sakonera+Kutsadura, family=poisson)
summary(eredua1)

## Egiantz-arrazoiaren testa
    # deviance funtzioak doitutako ereduaren deviance ematen digu.
    # df -> Askatasun graduak dira, bi ereduen arteko parametroen diferentzia
pchisq(eredua1$null.deviance-deviance(eredua1),df=2, lower=FALSE)

#Risk ratioaren eta konfiantza-tartearen kalkulua
exp(coefficients(eredua1))#RR
confint.default(eredua1)
exp(confint.default(eredua1))#Konfiantza-tartea


#-------------------------------------------------------------#
#### 3. KAPITULUA - Ordenazioa eta dimentsio-murrizketa ####
#-------------------------------------------------------------#

###            4. Adibidea - Faktore Analisia            ###

#Datuak irakurri eta laburpena egin
Y <- read.table(file="SparrowDA.txt", header=TRUE, dec=",", sep="\t")

summary(Y)
dim(Y)

#Aldagaien izenak aldatu
names(Y) <- c("hegokorda", "hegolau", "tartso", "buru", "mokolum", "mokonar",
"pisu", "behatzaile", "urte", "hilabete")

#Aldagaiak jarraitu bezala irakurtzeko
Y$hegokorda<-as.numeric(levels(Y$hegokorda))[Y$hegokorda]
Y$hegolau<- as.numeric(levels(Y$hegolau))[Y$hegolau]
Y$tartso<-as.numeric(levels(Y$tartso))[Y$tartso]
Y$buru<- as.numeric(levels(Y$buru))[Y$buru]
Y$mokolum<- as.numeric(levels(Y$mokolum))[Y$mokolum]
Y$mokonar<-as.numeric(levels(Y$mokonar))[Y$mokonar]
Y$pisu<- as.numeric(levels(Y$pisu))[Y$pisu]
```

```
#Aldagaiak zuzenean irakurtzeko
attach(Y)

#Normalitatearen azterketa
library(nortest)

lillie.test(hegokorda)
lillie.test(hegolau)
lillie.test(tartso)
lillie.test(buru)
lillie.test(mokolum)
lillie.test(mokonar)
lillie.test(pisu)

#Korrelazio-matrizea eta determinantea
korrelazio.matrize<-cor(Y[,c(1:7)])
korrelazio.matrize
det<-det(korrelazio.matrize)
det


#Bartlett-en testa
library(psych)
cortest.bartlett(korrelazio.matrize,n=1126)

#KMO eta MSA adierazleak
KMO(korrelazio.matrize)

#Balio propioak
eigen(korrelazio.matrize)

#Osagai nagusietako metodoa, faktoreak biratuz
ON <- principal(Y[,c(1:7)], nfactors = 2, rotate='varimax', scores=TRUE)
ON

##Faktore Analisia - Tartso aldagaia gabe
#Korrelazio-matrizea, tartso aldagaia kenduta
kor.mat_ZUZ<-cor(Y[,c(1:2,4:7)])
kor.mat_ZUZ

#Determinante berria
det<-det(kor.mat_ZUZ)
det
```

```
#Bartlett-en testa
cortest.bartlett(kor.mat_ZUZ,n=1126)

#KMO eta MSA adierazleak
KMO(kor.mat_ZUZ)

#Balio propioak
eigen(kor.mat_ZUZ)

#Osagai nagusietako metodoa, faktoreak biratuz
ON_zuz <- principal(Y[,c(1:2,4:7)], nfactors = 2, rotate='varimax', scores=TRUE)
ON_zuz

#Faktoreak eta jatorrizko aldagaiak lotzen dituen grafikoa
fa.diagram(ON_zuz$loadings,cut=.7,simple=FALSE, digits=4)

#Faktore-puntuazioak gorde
Y$F1berri <- ON_zuz$scores[,1]
Y$F2berri <- ON_zuz$scores[,2]

#Barreiadura-diagrama
plot(Y$F2berri~Y$F1berri, xlim=c(-5,5.0), ylim=c(-4,5.0))
abline(h=0, v=0)
identify(Y$F1berri, Y$F2berri,rownames(Y))


###      5. Adibidea - Korrespondentzia Analisia         ###
library(ca)

#Errenkaden (guneen) izenak zehaztu
rownames(bioenv)<- bioenv[,1]

#Korrespondentzia analisia burutu eta labupena eskatu
mat <-ca(bioenv[,2:6])
summary(mat)

#Guneen eta espezieen biplota
plot(mat, xlim=c(-1.5,1), ylim=c(-1.5,0.75))


###   6. Adibidea - Korrespondentzia Analisi Kanonikoa   ###

#Erantzun aldagaiak zehaztu
espezie <- bioenv[, 2:6]
```

```
#Aldagai askeak zehaztu
aske <- bioenv[, 7:10]

library(vegan)

#Korrespondentzia analisi kanonikoa burutu eta laburpena eskatu
espezie_cca <- cca(espezie ~ Tenperatura+Sakonera+Kutsadura+Sedimentua,
 data=aske)
espezie_cca

summary(espezie_cca)

#Ingurumen-ezaugarriek azaldutako inertzia, %:
cat(round((sum(espezie_cca$CCA$eig)/espezie_cca$tot.chi)*100,4),"%", "\n")

#Bi dimentsioetan ingurumen-ezaugarriek azaldutako inertzia zatia, %
cat(round(sum(espezie_cca$CCA$eig[1:2])/sum(espezie_cca$CCA$eig)*100,4),
"%", "\n")

#Bi dimentsioek azaldutako inertzia osoa, %:
cat(round(sum(espezie_cca$CCA$eig[1:2])/espezie_cca$tot.chi*100,4),"%", "\n")

#Lehenengo dimentsioak azaldutako inertzia:
cat(sum(espezie_cca$CCA$eig[1])/espezie_cca$tot.chi*100, "%", "\n")

#Bigarren dimentsioak azaldutako inertzia:
cat(sum(espezie_cca$CCA$eig[2])/espezie_cca$tot.chi*100, "%", "\n")

#Triplota
plot(espezie_cca)


#-----------------------------------------------------------#
#####         4. KAPITULUA - SAILKAPEN METODOAK        #####
#-----------------------------------------------------------#

###     5. Adibidea (Jarraipena) - Kluster Analisia      ###

#Guneen koordenatuak
as.data.frame(mat$rowcoord,row.names=row.names(bioenv[,2:6]))

#Lehenengo bi koordenatuak gorde
bioenv$F1<-mat$rowcoord[,1]
```

```
bioenv$F2<-mat$rowcoord[,2]

#Kluster analisi hierarkikoa - Warden metodoa eta distantzia karratura erabiliz
HClust.1 <- hclust(dist(model.matrix(~-1 + F1+F2, bioenv))^2,method= "ward")

#Dendograma
plot(HClust.1, main= "Cluster Dendrogram for Solution HClust.1",
xlab="Observation Number in Data Set especies", sub="Method=ward;
Distance=squared-euclidian")

#Klusterren tamainak eta bakoitzaren ezaugarriak
summary(as.factor(cutree(HClust.1, k = 4)))
by(model.matrix(~-1 + F1 + F2, bioenv), as.factor(cutree(HClust.1, k =+ 4)),
colMeans)

library(Rcmdr)

#Kluster bakoitzeko guneak
bioenv$hclus.label <- assignCluster(model.matrix(~-1 + F1 + F2,bioenv),
bioenv,cutree(HClust.1, k = 4))
bioenv1 <- subset(bioenv, bioenv$hclus.label==1)
bioenv2 <- subset(bioenv, bioenv$hclus.label==2)
bioenv3 <- subset(bioenv, bioenv$hclus.label==3)
bioenv4 <- subset(bioenv, bioenv$hclus.label==4)
row.names(bioenv1)
row.names(bioenv2)
row.names(bioenv3)
row.names(bioenv4)

#Guneak faktore-puntuazioen arabera:
plot(bioenv$F2~bioenv$F1)
abline(h=0, v=0)
identify(bioenv$F1, bioenv$F2,rownames(bioenv))

#Kluster analisi ez-hierarkikoa - K-means metodoa
.cluster <- KMeans(model.matrix(~-1 + F1 + F2,bioenv), centers = 4,
iter.max =10, num.seeds = 10)
.cluster$size #Klusterren tamainak
.cluster$centers #Zentroideak

#Kluster bakoitzean dauden guneak
bioenv$KMeans <- assignCluster(model.matrix(~-1 + F1 + F2, bioenv),
bioenv, .cluster$cluster)
as.data.frame(bioenv$KMeans,row.names=row.names(bioenv))
```

```
###       7. Adibidea - Analisi diskriminatzailea        ###
library(MASS)

#Behatzaileek behatutako txolarre kopurua
table(behatzaile)
#0 eta 9 behatzaileak analisitik kendu
Y<-Y[(behatzaile!= 0)&(behatzaile!= 9), ]
detach(Y)
attach(Y)
dim(Y)
table(behatzaile)

#Bariantza-kobariantza matrizeen arteko berdintasuna
library(biotools)
boxM(Y[,1:7], Y[,8])

#Korrelazio-matrizea
cor(Y[,c(1:7)])

#Aldagai estandarizatuekin
.Z <- scale(Y[,c("hegolau","tartso","buru","mokolum","mokonar","pisu")])
Y$Z.hegolau <- .Z[,1]
Y$Z.tartso <- .Z[,2]
Y$Z.buru <- .Z[,3]
Y$Z.mokolum <- .Z[,4]
Y$Z.mokonar <- .Z[,5]
Y$Z.pisu <- .Z[,6]
remove(.Z)
Y_std <- Y[,c(11:16,8)]
names(Y_std)
names(Y)
attach(Y)

#Taldeen arteko diferentziak
summary(manova(as.matrix(Y_std[, c(1:6)]) ~ behatzaile), "Wilks")

#Funtzio diskriminatzaileak
discrim1<-lda(behatzaile ~ Z.hegolau + Z.tartso + Z.buru + Z.mokolum
+ Z.mokonar + Z.pisu, data = Y_std)
discrim1

#Korrelazioak funtzio diskriminatzaileen eta aldagaien artean
correlation<-cor(x = Y[, c(2:7)], y = predict(discrim1)$x)
```

```
#Korrelazioen grafikoa lehenengo bi funtzio diskriminatzaileen
#eta aldagaien artean
plot(c(-0.8, 0.8), c(-0.8, 0.8), type = "n",
     xlab = "Discriminant function 1", ylab = "Discriminant function 2")
lines(c(-1, 1), c(0, 0), lty = 2)
lines(c(0, 0), c(-1, 1), lty = 2)
library(ade4)
s.arrow(correlation[, 1:2], grid = F, add.plot = T, addaxes = T, clabel = 1.5)

#Puntuazio-diskriminatzaileak
predict(discrim1)

#Taldeak iragarri
predict(discrim1)$class

#Lehenengo bi funtzio diskriminatzaileen puntuazioak grafikoki
plot(predict(discrim1)$x[, 1], predict(discrim1)$x[, 2], type = "n",
     xlab = "Discriminant function 1", ylab = "Discriminant function 2")
text(predict(discrim1)$x[, 1], predict(discrim1)$x[, 2], labels = behatzaile)


#Konfiantza-tarteak erabiliz
group_average.1<-tapply(predict(discrim1)$x[, 1], behatzaile, mean)
group_average.2<-tapply(predict(discrim1)$x[, 2], behatzaile, mean)
plot(predict(discrim1)$x[, 1], predict(discrim1)$x[, 2], type = "n",
     xlab = "Discriminant function 1", ylab = "Discriminant function 2")
text(group_average.1, group_average.2, 1:8,cex = 1)
Radius <- 2.15
#NG = talde kopurua
NG <- 8
for (i in 1:NG) {
     x3 <- vector(length=72)
     y3<- vector(length=72)
     t1 <- 0
     for (j in 1:72) {
         x3[j] <- Radius * sin(t1)+ group_average.1[i]
         y3[j] <- Radius * cos(t1)+ group_average.2[i]
         t1 <- t1 + 3.1416/36
     }
     lines(x3,y3)
     }
abline(0,0)
abline(h=0,v=0)
```

```
#Histogramak
par(mar = rep(2, 4))
ldahist(data = predict(discrim1)$x[, 1], g=behatzaile)
ldahist(data = predict(discrim1)$x[, 2], g=behatzaile)
ldahist(data = predict(discrim1)$x[, 3], g=behatzaile)
ldahist(data = predict(discrim1)$x[, 4], g=behatzaile)
ldahist(data = predict(discrim1)$x[, 5], g=behatzaile)
ldahist(data = predict(discrim1)$x[, 6], g=behatzaile)

#Nahaste-matrizea
table(behatzaile)
table(behatzaile, predict(discrim1)$class)

#Talde bakoitzean zuzenki sailkatutako txolarreen portzentajeak
for (i in 1:length(unique(behatzaile))) {
cat(i, round(sum(behatzaile== i&predict(discrim1)$class == i)/
                sum(behatzaile== i)*100, 2), "\n")
}

#Zuzenki sailkatutako portzentajea guztira
sum(diag(prop.table(table(behatzaile, predict(discrim1)$class))))*100
```