



INKESTEN LOTURA

2014

Erakunde Autonomiaduna



Eustat

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA
www.eustat.eus



EUSKO JAURLARITZA
GOBIERNO VASCO

Lanketa:

EUSTAT

Euskal Estatistika Erakundea

Instituto Vasco de Estadística

Argitalpena:

EUSTAT

Euskal Estatistika Erakundea

Instituto Vasco de Estadística

Donostia-San Sebastián 1

01010 Vitoria-Gasteiz

© **Euskal AEko Administrazioa**

Lehen edizioa

I/2015

Imprimaketa eta Enkoadernazioa:

Eusko Jaurlaritzako Inprimategia eta Erreprografia Zerbitzua

ISBN: 978-84-7749-482-9

Lege-gordailua: VI-794/2014

INKESTEN ARTEKO LOTURA

Ines Garmendia Navarro

ines.garmendia@gmail.com



EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia kalea 1
01010 VITORIA-GASTEIZ
Tel.: 945 017 500
Faxa: 945 017 501
Helbide elektronikoa: eustat@eustat.es
www.eustat.es

Aurkezpena

Estatistika ofizialaren eremua lan handia egiten ari da integrazio teknikak bereganatzeko, erabiltzaileei ahalik eta kalitate oneneko informazioa eskaintzeko xedeaz. Zentzu horretan, bere XXVI Estatistikako Nazioarteko Mintegia "Statistical Matching: Methodological issues and practice with R-StatMatch" (*"Inkesta arteko lotura: Metodologia eta praktika R-StatMatch-ekin"*) gaiari eskaini zion Eustatek.

Bekadun batek eremu horretan egindako ikerketa-lana zabaltzea da aurkezpen honen helburua. Koaderno hau ataletan banatua dago: lehenengo atala metodologiari eskainia dago, eta bigarrenak R softwarean inplementatuak dauden teknikak azaltzen ditu. Hirugarrena Eustat-eko bi inkesta independenteren loturari eskainia dago (hots, Bizi-Baldintzen Inkesta eta Biztanleria Jardueraren Arabera Sailkatzeko Inkesta). Azkenik, laugarren atalak R-ko pakete propio baten garapena azaltzen du.

Vitoria-Gasteiz, 2014ko abendua

Josu Iradi Arrieta

EUSTAT-eko Zuzendari Nagusia

Aurkibidea

AURKEZPENA	2
AURKIBIDEA.....	3
SARRERA	4
METODOLOGIA.....	6
INKESTEN ARTEKO LOTURAREN OINARRIAK.....	6
HURBILKETAK ETA METODOAK	7
Mikro metodoak	8
Makro metodoak	13
Lagin-diseinu konplexuetan erabiltzeko berariazko metodoak.....	14
INKESTEN ARTEKO LOTURA BATEN FASEAK	15
EMAITZEN BALIOZKOTASUNA	19
Baliozkotasun-mailak	19
Eztabaida	20
SOFTWAREA.....	23
EZARPEN PRAKTIKOA.....	25
EUSTAT-EK EGINIKO BI LAGIN-INKESTEN ARTEKO LOTURA: BIZI-BALDINTZEN INKESTA ETA	
BIZTANLERIA JARDUERAREN ARABERA SAILKATZEKO INKESTA	25
Inkestak deskribatzea	25
BBI-BJA arteko lotura	27
Emitzak	34
R PAKETE PROPIOA GARATZEA.....	47
ONDORIOAK.....	49
BIBLIOGRAFIA.....	51
ERANSKINAK	53

Sarrera

Koaderno tekniko honetan bildutakoa Euskal Estatistika Erakundeak, 2012. urtean, *Inkesten arteko loturarako* estatistika- eta matematika-metodologietan prestatzeko eta ikertzeko emandako bekari esker egindako lanaren emaitza da.

Inkesten arteko lotura¹ deituriko metodologiak aukera ematen du estatistika integratuak eta adierazle konbinatuak egiteko. Independentek eta intereseko biztanle-kopuru berari buruzkoak diren inkestetatik ateratzen dira datuak. Hortaz, zenbait inkestetatik ateratako datuak, fitxategi ezberdinetan jasota daudenak, eraginkortasun handiagoz erabiltzen ditu; horixe da, hain zuzen, haren abantaila nagusia.

Metodologia horrek estatistika-teknika ugari, mota eta jatorri ezberdinetakoak, barne hartzen ditu: esaterako, falta diren (*missing*) datuak egozte, ziurgabetasuna zenbatzea eta laginketa konplexuari buruzko teoria. Teknika horiek etengabe garatzen ari dira, eta haietako asko software libreko R ingurunearen bitartez banatzen dira. Plataforma hori gero eta gehiago erabiltzen da akademia-esparruan, industrian, eta, pixkanaka, estatistika ofizialetan ere bai.

Koaderno honela egituratuta dago: lehenengo atalean sarrera hau jaso da; bigarrenetan metodologia eta teknika nagusiak azalduta daude. Horrez gain, zenbait gomendio bildu dira, teknikak praktikan erabiltzeko. Hirugarren kapituluak R inguruneak teknikak ezartzeko ematen dituen aukerak zehaztuta daude. Laugarren atalean benetako kasu bat, bi inkesten arteko loturari buruzkoa, jaso da. Erlazionatu diren inkestak Eustatek egin dituen bi inkesta independente dira: Biztanleria jardueraren arabera sailkatzeko inkesta eta Bizi-baldintzen inkesta. Adibide praktikoaren bidez, lotura batek eduki behar dituen fase nagusiak azaldu dira, bai eta zer emaitza-mota ematen dituen ere. Bosgarren kapituluak R pakete propio baten garapena deskribatu da. Seigarren kapituluak, azkenean, emaitzak bildu dira. Bukatzeko, eranskinean jaso dira taulak eta zenbakizko emaitzak.

Koaderno hau egin den bitartean, R pakete propio bat ere garatu da, koadernoan azaldutako metodologia ezartzen laguntze aldera. Gaur egun, lotura bateko faseei heltzeko funtzioak zenbait paketetan jasota daude. Horregatik, erabiltzaileak zenbait «filosofia»ri jarraituta egituratu behar du kodea, une bakoitzean erabiltzen ari den paketea zein den. Arazo hori konpontzea izan da, hain zuzen, ikerketa hau egiteko motibazioa. Garatutako pakete propioak (*micromatch*) zailtasun hori ezabatu nahi du. Horretarako, interfaze bakarra eskaintzen dio erabiltzaileari, eta bertan inkestak lotzeko

¹ Terminologia askotarikoa da. Terminoak baliokideak dira, eta iturria eta testuingurua zein diren, batzuk edo beste batzuk erabiltzen dira, ingelesez: *statistical matching*, *data fusion*, *file merging*, *survey linking* eta *synthetic matching*.

kontzeptuak era unibokoan azalduta daude. Horri esker, jada frogatu eta erkatu diren paketeak eraginkortasun handiagoz erabil daitezke. Paketea Konpostelako Santiagon antolatu ziren *VI Jornadas de Usuarios de R* jardunaldietan aurkeztu zen (2014ko urriaren 23an eta 24an), eta besteak beste, Eustat-eko web-aren bitartez eskuragarri dago.

GAKO-HITZAK: Inkesten arteko lotura, datuak bateratzea, falta diren balioak egozteak, R.

Metodologia

Inkesten arteko loturaren oinarriak

Inkestak lotzeko jardueran zenbait teknika erabiltzen dira, adierazleen edo aldagaien gaineko estatistika integratuak lortzera bideratuta. Datuak zenbait iturritatik ateratzen dira, eta oro har, lagin-inkestak biztanle-kopuru beraren gainekoak dira.

Praktika orokorrean, bi lagin-inkesta independente erabiltzen dira, biztanle-kopuru berari buruzkoak (adibidez, EAEko egoiliarak 2014an). Inkesta bakoitzean zenbait dimentsio edo adierazle (bizimoduak, lan-egoera, diru-sarrerak...) neurtzen dira, datuok bereizita daudela. Lotura sortzeko, inkestek aldagai edo neurketa komun batzuk eduki behar dituzte; eta oinarrizko aldagai soziodemografikoak izan ohi dira: adina, sexua edo ikasketa-maila.

Bi inkesta independentetik jasotako datuak «batzerakoan»² 1. irudiak erakusten duen antzeko egoera bat sortzen da. Z blokeak erakusten ditu aldagai komunak (soziodemografikoak eta beste batzuk), hots, bi inkestek jasotzen dituztenak. X eta Y blokeek irudikatzen dituzte galdera-sorta bakoitzeko item propioak. Z bloke komuneko informazioa erregistro guztietarako erabilgarri egongo da (loturaren oinarrizko premisa baita). Bloke espezifikoetan (X eta Y), ordea, inkesta bakoitzeko erregistroek soilik edukiko dituzte balio informatuak. Horrela, «falta» diren edo behatu ez diren balioek osatutako bi bloke handi izango ditugu (1. irudian zati argiagoak dira).

Aldagai komunak (inkesta bietan)	Berariazko aldagaiak (1. Inkesta)	Berariazko aldagaiak (2. Inkesta)
Z	X	
		Y

 Aldagai erabilgarriak
 Bildu gabeko aldagaiak

² Koaderno honetan bi inkestaren arteko lotura baino ez dugu aztertu, azalpenak errazte aldera, baina metodologia $n > 2$ inkestetan erabil daiteke.

1. irudia: Z bloke bat, aldagai komunak dituen, partekatzen duten bi inkesta «lotuz» gero, eskema hori sortzen da. X eta Y blokeek irudikatzen dituzte berariazko itemak edo inkestek partekatzen ez dituztenak. Prozesu horri «kateaketa» deitzen diogu.

Abiapuntua:

- * Aldagai komunak dituen Z bloke bat dago, hau da, aldagai horiek fitxategi biek dituzte.
- * Bestalde, X eta Y blokeek berariazko aldagaiak, ez-partekatuak, dituzte: X #1 fitxategian baino ez dago; eta Y #2 fitxategian soilik dago.
- * Biztanle-unitate bat lagin batean eta bestean agertzeko probabilitatea ia zero da, eta bazter dezakegu.

Inkesten arteko loturak zehazten duen egoera oso bestelakoa da, beste teknika batzuekin erkatzen bada, hala nola, **erregistroak batzeko** teknikarekin. Beste teknika horiek helburu dute fitxategien artean errepikatzen diren unitate berak zehaztea (adibidez, errolda baten eta administrazio-fitxategi baten artean). Inkesten arteko lotura ezberdina da, lagin-inkesta independenteak abiapuntutzat hartzen baitu. Horregatik, nolabait, kontrako egoera dela esan daiteke: hasieratik, unitateak ezberdinak direla *jakin badakigu*, baina antzeko bikoteak *bilatu nahi dira*. Helburua ez da *unitateak* erlazionatzea, baizik eta inkesten *aldagaiak* erlazionatzea.

Laburbilduz, inkesten arteko loturak berariazko aldagaiak lotu nahi ditu. Aldagai horiek berariazkoak dira biztanle-kopuru berari buruzko lagin-inkesta independenteetan. Horretarako, inkesten artean partekatutako informazioa «zubi» gisa erabiltzen da. Segidan, arazo horri aurre egiteko egin diren hurbilketa nagusiak aztertuko ditugu, bai eta denboran zehar erabili diren irtenbideak ere.

Hurbilketak eta metodoak

Inkesten arteko lotura bat sortzerakoan, oro har, honako hurbilketa hauetako bat erabiltzen da (M. D'Orazio, M. Di zio & M. Scanu, 2008):

- **Makro hurbilketak** zuzeneko zenbatespenak bilatzen ditu. Zenbatespenak berariazko aldagai batzuei buruzkoak dira: esaterako, X aldagai baten eta Y aldagai baten arteko korrelazio-koefiziente bat, edo bazter-banaketa bateratu bat.
- **Mikro hurbilketaren** helburua fitxategi sintetiko bat sortzea da. Fitxategi horretan biltzen dira fitxategietako berariazko aldagai guztiei buruzko datu guztiak. Fitxategi sintetiko hori analisi bateratuak egiteko erabiltzen da, eta hasieran fitxategi bereizietan dauden aldagaiak aztertzen dira.

Gaur egun, inkestak lotzeko estatistika-metodo anitz daude. Izan ere, metodologia horren hasieratik, hots, XX. mendeko 60ko hamarkadatik, irtenbide ugari garatu dira, bai

merkatuen ikerketa-esparruan (Europar), bai estatistika ofizialen esparruan (Estatu Batuetan eta Kanadan, 70eko hamarkadatik). (Interesa izanez gero, irakurleak S. Rässler (2002) liburuko hirugarren atala irakur dezake, inkesten arteko loturaren historia laburra, 2002. urtea artekoa, azaltzen baitu).

Segidan, mikro eta makro hurbilketek erabiltzen dituzten metodo nagusiak aztertuko ditugu. Horrez gain, estatistika ofizialetan garrantzitsua denez, lagin-diseinu konplexuak dituzten inkesten trataera ere berrikusiko dugu.

Mikro metodoak

Mikro hurbilketaren helburua fitxategi sintetikoa sortzea da. Fitxategi horretan biltzen da berariazko aldagai interesgarri guztiei buruzko informazio osoa, erregistro guztiak kontuan hartuta. A. Leulescu-k eta M. Agafitei-k (2013) lau mikro metodo-multzo aipatzen dituzte, eta jarraian azaldu ditugu.

Hot-deck metodoak

Inkesten arteko lotura garatu bitartean, *hot-deck* metodoen familia erabiliena izan da, alde handiz, gainera. Metodo ez-parametrikoen multzo bat da, hau da, aldagaiek ez dute alde aurretik ezarritako estatistika-banaketa.

Hau da prozedura: fitxategietako batean («fitxategi hartzailea» deritzona) erregistro bakoitzerako bilatzen dira erregistro bat edo gehiago beste fitxategi batean («fitxategi emailea» deiturikoa). Erregistroek ahalik eta antzekoenak izan behar dute, aldagai komunei dagokienez (adina, sexua, ikasketa-maila...). Aurkitutako fitxategi emaileari buruzko balioak *egotzi* egiten zaizkio fitxategi hartzaileari (ikus 2. irudia).

Egotzitako balioak beti dira balio errealak; hau da, benetan behatu diren balioak dira, eta fitxategi emailean jasota daude. Horixe da, hain zuzen, *hot-deck* prozeduraren ezaugarri nagusia.

A (hartzailea)

Z	X
Erregistro hartzailea	

B (emailea)

Z	Y
Erregistro emailea	

+

=

(A+B)

Z	X	Y
Erregistro hartzailea, egotzitako balioekin		

2. irudia: balioak *hot-deck* metodoaren arabera egozteko eskema-idea. Erregistro hartzaile bakoitzeko, erregistro emaile antzekoena bilatzen da, eta dagokion balioa «egozten» zaio.

Sarritan, datuen arabera, hartzaile bakoitzari ezin zaio balio bera duen emailea esleitu aldagai guztietan. Horregatik, aldagai komunen arteko *distantzia* zehaztu ohi da, erregistro antzekoen bikoteak bilatu ahal izateko. Distantziaren matematika-formulazioa, aldagaien motaren arabera (kategorikoak edo zenbakidunak) aldatzen da, bai eta beste alderdi batzuen arabera ere (ikus M. D'Orazio et al. (2008), C eranskinean).

1. koadroan adierazi den bezala, *hot-deck* egozte-algoritmoak zenbait aldaera onartzen ditu, ezarpenaren arabera: estratuak definitzea, erregistro emaileak ez errepikatzeo murrizketak, etab. Aldaera horiek behar bezala parametrizatzen baldin badira, ondo eginiko *hot-deck* algoritmo batek behar bezala birsortuko ditu fitxategi emailean behatu dituen banaketak egotzitako fitxategi hartzailean.

HOT-DECK METODOAK

Sailkapena

- Mikro metodoa.
- Ez-parametrikoa.
- Maiztasunaren arabera.

Algoritmoa

- A-ko erregistro hartzaile bakoitzari emaile bat bilatzen zaio B-n, Z aldagai komunak ahalik eta antzekoenak izanda. Erregistro emaileko Y balioak erregistro hartzaileari egozten zaizkio.
 - B-n hartzailearen bezalako erregistro emailek ez badago, distantzia ezartzen da, antzekoena aurkitu ahal izateko. Distantzia-funtzio batzuk erabil daitezke: Manhattan distantzia, euklidearra, Gower-ena (aldagai kategorikoak eta zenbakizkoak nahastuta daudenean), etab.

Aldaerak

- Estratuak zehaztea: *hot-deck* berezi bat erabiltzen da, aldagai komun maila jakin batzuetarako. Adibidez, ohikoa da sexu-aldagai bakoitzerako *hot-deck* bat egitea, edo adinaren eta sexuaren arteko gurutze bakoitzerako. Erregistro emaileak estratu bakoitzaren barnean bilatzen dira. Horrela, koherentziarik gabeko emaitzak baztertzen dira, eta distantzien kalkulua nabarmen errazten da.
- Emaileen erabilera mugatzea: Printzipioz, B erregistroak behin baino gehiagotan erabil daitezke emaile gisa. Eta horrek jatorrizko banaketak aldatzeko arriskua dakar. Arrisku hori saihesteko, *hot-deck* metodo mugatuak garatu dira. Metodo horiek emaileak behin baino gehiagotan ez erabiltzeko muga dute.
- Hartzaile bakoitzerako emaile bakarra erabili beharrean, antzekoenak diren «K» erregistro erabiltzeko aukera dago. Kasu horietan, egotzitako balioa K balio horien arteko konbinazioa da. Beste aldaera bat «d» distantzia jakin batean dauden erregistroak erabiltzea da.

Kalkulua

- `R > StatMatch > NND.hotdeck(), RANDwNND.hotdeck(), rankNND.hotdeck()`

Erreferentziak

- [1] M. D'Orazio et al. (2008). C eranskina: Distantziak hautatzea (34-45 orr.).
[2] A. Leulescu et al. (2013).

Literaturan askotan erakutsi den bezala, inkesta independenteak *hot-deck* metodoaren bitartez lotzen baldin badira, garrantzitsua da kontuan hartzea onartu egin dela «baldintzatutako independentzia-hipotesia» deituriko egoera dagoela.

Hipotesi horri jarraituta, X eta Y berariazko aldagaien artean dauden erlazio guztiak (ez-behatuak) bildu dituzte (Z ,X) eta (Z ,Y) aldagaien arteko erlazio partzialek (behatuak). Emaila-hartzaile bikoteen arabera bilaketa zehatza baldin bada, (hau da, algoritmoak erregistro hartzaileari emaila berdina aurkitu ahal izan badio aldagai komun guztietan), hipotesi horren arabera, *hot-deck* metodoak berariazko aldagaien (X eta Y) arteko erlazio «erreala» (ez-behatua) ondo birsortuko du. Egoera ideal horretan, Z aldagaiek emandako informazioa nahikoa da X eta Z aldagaien arteko erlazioa birsortzeko.

Hala eta guztiz ere, hipotesi hori murriztailea da, eta ez da beti betetzen. Adibidez, «A» inkesta batek banakako diru-sarrerak biltzen ditu, eta «B» inkesta batek lan-egoera biltzen du («langabea», «landuna», «inaktiboa»). Adin-aldagaia erabili da inkesta horiek *hot-deck* metodoaren bitartez lotzeko. Horren arabera, A fitxategian (hartzailea) B fitxategiko (emaila) «lan-egoera» deritzon aldagaia gehitzen da. Hortaz, egotzitako A fitxategian diru-sarreraren banaketak, lan-egoerak (adibidez, inaktiboena) baldintzatuak, soilik erakutsiko du adinarekiko erlazioa edo mendekotasuna. Bestela esanda, horrela (hau da, adinaren aldagaia soilik erabilita) egotzitako fitxategi batean ez da erakutsiko diru-sarreraren eta inaktiboaren egoeraren arteko erlazioa, erlazio hori adinak zehaztu ez badu, behintzat.

Praktikan, komeni da hipotesi horretara hurbiltzea, eta horretarako, eskuragarri dauden datu guztiak erabili behar dira. Koaderno tekniko honetan ikusiko dugun bezala, erronka handiena baldintzatutako independentziaren egoera idealera ahalik eta gehien hurbiltzean datza.

Erregresioan oinarritutako metodoak

Hot-deck metodoek ez bezala, erregresio linealean oinarritutako metodoak erabat parametrikokoak dira, hau da, estatistika-eredu espezifikoak ezartzen diete aldagaiei. Kasu honetan, baldintzatutako independentzia-hipotesiak haxe du oinarri: banaketa bateratuko funtzioa bazter-banaketen funtzioen produktua da. Honela:

$$f(x,y,z) = f(x|z) \times f(y|z) \times f(z)$$

(Bestela esanda, onartu egiten da fitxategi partzialetako datuek aldagai guztiak dituen fitxategi oso bat eraiki dezaketela).

Erregresioaren arabera egozteak erregresio-eredua erabiltzen du, «falta diren» behaketei iragarri zaizkien balioak eskuratzeko. Hori horrela, ez da balio errealek (behaturik) egozten, baizik eta informazio komunaren arabera iragarri den balio bat. Baina prozedura sinpleak desabantaila batzuk ditu: sentikorra da zehaztutako ereduaren desegokitasunekin; erregresioa batez bestekorantz doa; bariantza azpiestimatzeko (erregresio-zuzenean dauden balioak hartzeagatik).

Irtenbide bat *erregresio estokastikaren arabera* egozte da: ausazko hondar-balio bat bariantza hobeto erakusteko sartzean datza. Aldagai-mota horiek, segidan adieraziko dugun bezala, metodo mistoak, sofistikatuagoak, garatzeko balio dute.

Metodo mistoak

Metodo mistoak sortzen dira aurreko hurbilketa biak konbinatzen ditugunean: batetik, *hot-deck* familiako metodo ez-parametrikokoak, sendoak, eta alde aurretik eredurik zehazten ez dutenak; eta, bestetik, metodo parametrikokoak, erregresioan oinarrituak eta motelagoak, distantziak kalkulatzeko hautatu diren aldagaien mende ezinbestean ez daudelako.

Familia horren barnean honako egozte-metodo hau aipatu behar da: *predictive mean matching* deiturikoa, Rubin-ek aurkeztua (1986). Prozedura horren arabera, «falta diren» balioak egozteko, erregresio batek iragarritako balioei jarraitzen zaie. Zehatzago esanda: lehenik, X erregresioa, Z -ren gainekoa, kalkulatu da B fitxategi emailean. Ekuazio horren bidez, iragarritako batez besteko balioa kalkulatu da, $\hat{X}$, A fitxategi hartzailean. Segidan, emaile-hartzaile bikoteak bilatze aldera, *hot-deck* metodo bat erabiltzen da, $d(X, \hat{X})$ distantzia batean oinarrituta. Azkenik, behatutako balioak egozten dira. Laburbilduz, *predictive mean matching* prozedura *hot-deck* egozte bat da, baina batez besteko balioetan oinarrituta, eta balio horiek erregresio-ereduen bitartez lortzen dira.

Nabarmendu beharreko beste metodo misto bat *propensity score* da (S. Rässler (2002), A. Leulescu et al. (2013)). Fitxategi biak, emailea eta hartzailea, A fitxategiko (emailea) erregistro guztietarako 1 balioa eta B fitxategiko (hartzailea) erregistro guztietarako 0 balioa hartzen dituen aldagai osagarri baten bidez «hedatu» egiten da. Erregistro guztiak fitxategi bakarrean bilduta, «logit» edo «probit» eredu bat kalkulatu da: mendeko aldagaia gehitutako aldagai osagarritzat hartzen da, eta aldagai komunak aldagai independentetzat hartzen dira. *Propensity score* prozedura honela definitzen da: unitate batek fitxategi batekoa edo bestekoa izateko duen probabilitate baldintzatu zenbatetsia. Azkenik, emaile antzekoenak hautatuz egiten da lotura, eta horretarako *propensity score* prozedurako balioak oinarri hartzen dira.

Egozte sinplean oinarritutako metodoak

Egozte anitza Rubin-ek sartu zuen 1970eko hamarkadan, falta diren balioen esparruan. Harrezkero, askotan erabili da inkesten arteko loturak sortzeko testuinguruan. Hauxe da haren funtsa: falta den balio bakoitzari (ez-behatua) $m > 1$ balio onargarriak esleitzea, balio bakarra esleitu beharrean. Horrela, balio horren ziurgabetasuna agerikoa da. Atera diren m balio horien bidez *pooling* (edo konbinazioa) egiten da, balio bakarra sortzeko, bai eta balio horri lotutako ziurgabetasuna (edo *intra-egoztearen bariantza*) zenbatesteko ere (A. Leulescu et al., 2013).

Inkesten arteko loturaren testuinguruan, datu osoak dituzten fitxategiak sortzeko erabili ohi da egozte anitza. Hala eta guztiz ere, egozte anitza testuinguru konplexuagoetan ere erabil daiteke. Horren froga da erregresio sekuentzialaren bidezko egozte anitza

(*sequential regression multiple imputation*). Kasu horretan, eredu independenteak erabiltzen dira aldagai bakoitza iterazio batzuetan egozteko.

Informatika aldetik, metodo aurreratu horiek esfortzu handia eska dezaketen arren, malgutasun handia eman dezakete. Adibidez, zenbati aldagai aldi berean egozteko bide ematen dute. Gaur egun, erraz samarra da metodo horiek erabiltzea, software libreko R paketeetan ezarrita daudelako: `mice`, adibidez. (Informazio gehiago nahi izanez gero, ikus hirugarren atala: Softwarea).

Makro metodoak

Makro metodoek honako helburu hau dute: X eta Y berariazko aldagaiei lotuta dagoen intereseke parametro baten gaineko zenbatespen zuzena lortzea. Prozedura horiek azaltzeko, demagun bi fitxategi bereizita ditugula, biztanleria-kopuru berari buruzkoak:

- A fitxategiak ausazko lagin sinplea dauka, Z eta X aldagaien n_A behaketarekin.
- B fitxategiak ausazko lagin sinplea dauka, Z eta Y aldagaien n_B behaketarekin.

Egoera sinpleenean, demagun hiru aldaera (X,Y,Z) dituen banaketa, ez-behatua, normala dela; eta honako parametro hauek dituela:

$$\mu = \begin{pmatrix} \mu_Z \\ \mu_X \\ \mu_Y \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_Z^2 & \sigma_{ZX} & \sigma_{ZY} \\ \sigma_{XZ} & \sigma_X^2 & \sigma_{XY} \\ \sigma_{YZ} & \sigma_{YX} & \sigma_Y^2 \end{pmatrix}$$

Banaketa bateratuaren ezaugarriak zehazte aldera, A fitxategia erabil dezakegu σ_{ZX} zenbatesteko, eta B fitxategia σ_{ZY} zenbatesteko. σ_Z^2 zenbatesteko, A edo B fitxategietako erregistroak erabil ditzakegu; eta hobe, $n_A + n_B$ bidez sorturiko erregistroak, $A \cup B$ kateatutako erregistrokoak, erabiliko bagenitu.

Helburua σ_{XY} baliorako zenbatespen bateratua zehaztea baldin bada –C fitxategi osagarririk ez edukiz gero, banaketa bateratuko behaketekin –, beharrezkoa da hipotesi osagarriren bat gehitzea, hala nola baldintzatutako independentzia-hipotesia. Hipotesi horren arabera, σ_{XY} kalkula daiteke kobariantza partzialen bidez:

$$\sigma_{XY} = \frac{\sigma_{ZX} \sigma_{ZY}}{\sigma_X^2}$$

StatMatch paketeen, makro metodoak honako funtzio hauetan ezarrita daude: `mixed.mtc()` eta `comb.samples()`. Azken horrek lagin-diseinua kontuan hartzeko aukera ematen du, lagin konplexuak direnean (ikus hurrengo atala).

Lagin-diseinu konplexuetan erabiltzeko berariazko metodoak

Askotan, lotu beharreko A eta B fitxategiek laginketa konplexuak dituzte. Hau da, biztanle-unitateen ausazko laginketa egiten ez duten inkestak abiapuntu hartuta sortu dira.

Inkestak lotzeko prozeduretan lagin-diseinua sartzeko zenbait modu daude. Guk **Renssen prozedura** nabarmenduko dugu. R-ko StatMatch paketeko `comb.samples()` funtzioan ezarrita dago. Prozedura horretan fitxategi bakoitzeko erregistroei esleitutako pisuak era jarraituan kalibratzen dira, eta lagin-diseinua irudikatzen dute. (Laginketa konplexuko terminologietan, kalibratzea pisuak berriro kalkulatzeari datza, oro har. Horrela, lortzen diren balioak eta diseinuaren balio teorikoak ahalik eta antzekoenak dira, baina, horrez gain, beste baldintza batzuk ere betetzen dira: esaterako, aldagai nagusietan biztanleei buruz ezagutzen diren erreferentziazko guztikoak birsortzea).

Renssen prozedurak, oro har, aldagai kategorikoekin erabiltzen da, bai eta makro helburu batekin ere (adibidez, $X \times Y$ kontingentzia-taula zenbatestea, era bateratuan behatu ez diren aldagaiei dagokienez). Bi faseko prozedura osoan, A eta B fitxategiak eta C fitxategi osagarria (hala dagokionean) bereizita egoten dira.

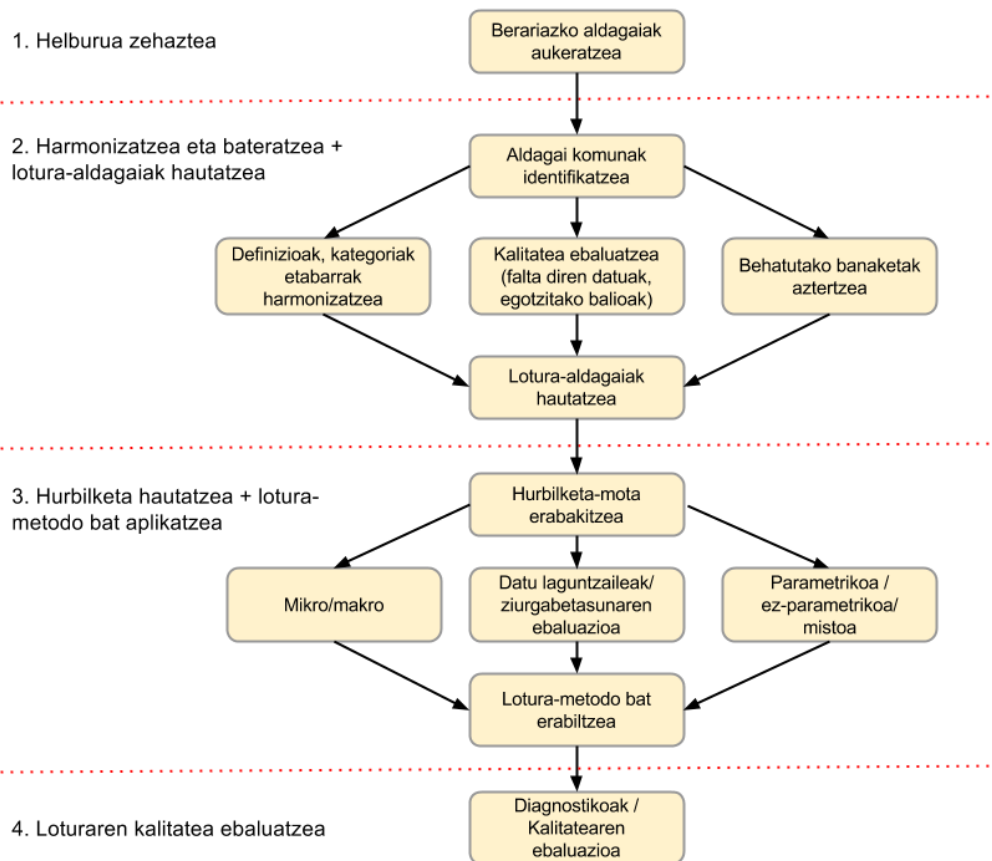
Lehenengo fasean, A eta B fitxategietako pisuak (w_A eta w_B , hurrenez hurren) berriro kalkulatu dira, X aldagai komunetarako guztizkoak lortzeko xedez (gustizkoak ezagunak edo A eta B fitxategi berberen bidez zenbatetsiak izango dira). Bigarren fasean, honako bi kasu hauek aintzat hartzen dira:

- C fitxategi osagarri bat, datu guztiak dituen, erabilgarri baldin badago, fitxategi horretako pisuak, w_C , kalibratzen dira A eta B fitxategietako guztizkoekin lerroka daitezen (bakoitzari dagozkion kalibraketak 1. urratsean egin ostean). Ondoren, haxe zenbatesten da: $X \times Y$
- C fitxategi osagarria erabilgarri ez badago, baldintzatutako independentzia-hipotesia erabiltzen da, zenbatespena lortzeko

Interesa izanez gero, irakurleak haxe irakur dezake: «*Old and new approaches in statistical matching when samples are drawn with complex survey designs*» (D'Orazio et al., 2010). Bertan Renssen kalibraketa-prozesua eta antzekoak diren beste metodo batzuk erkatuta daude.

Inkesten arteko lotura baten faseak

Edozein metodori jarraituta ere, inkesten arteko lotura egiteko, beharrezkoa da haien artean oso lotuta dauden fase batzuk garatzea. Kontuan hartu behar inkesten arteko loturaren metodoa fase horietako bat baino ez dela, eta, askotan, ezta garrantzitsuen ere (A. Leulescu et al., 2013).



3. irudia: inkesten arteko lotura baten faseak. «Report WP2 ESS-net, Statistical Methodology Project on Integration of Surveys and Administrative Data» 34. orr. Koaderno honetan agertzeko egokituta dago.

Lehen fasean, lotura-helburua zehaztu behar da: (i) mikro edo makro izango den; eta (ii) berariazko aldagaiak aldeztu aurretik ezarri behar dira. Erabaki horiek funtsezkoak dira, gainerako fase guztiak zehaztuko dituztelako.

Bigarren fasean bi ataza nagusi egin behar dira: Lehenik, lagin bietako datuen arteko koherentzia aztertu behar da. Horretarako, iturrien arteko bateratze- eta elkartze-maila guztiz aztertu behar da. Edonola ere, gutxienez, honako alderdi hauek ikertu behar dira (M. D’Orazio et al., 2008):

- Unitateen zehaztapenak bat etortzen ote diren, bai eta erreferentzia-aldiak ere

- Aldagaiak eta neurri komunak, haien sailkapenekin batera (aldagai kategorikoak direnean), bat etortzen diren
- Erantzun-eza partziala edo totala: *missing* edo falta diren balioen trataera.
- Erroreen kalkulua (laginen alborapena eta zehaztasuna)
- Jatorrizko aldagaien (balizko) prozesamendua adierazle sintetikoetan

Ohikoa da desadostasunak agertzea alderdi horietako batean edo batzuetan, eta, batez ere, datuen bilketari dagokionez (kategoria ezberdineko neurtu diren aldagaiak...). Era berean, litekeena da datu-bilketaren ondorengo trataera (pisuen kalibraketak, adierazle sintetikoen kalibraketa) bestelakoa izatea lagin bakoitzean. Alderdi horiek guztiak kontuan hartuta, zerrenda bat identifikatu behar da, aldagai baliokideak edo bi iturrien arteko definizio alderagarria duten aldagaiak dituen. Orduan, ikerlan konparagarria egingo da, fitxategi batean eta bestean aurkitutako banaketa enpirikoak aztertzeko, eta aldagai bateraezinak behin betiko baztertuko dira

Segidan, xede diren berariazko aldagaiak iragarri ahal izateko, hautatu diren aldagai komunek ematen duten informazioa garrantzitsua den aztertu behar da. Esaterako, nork bere osasunari buruz duen pertzepzioa (lehen inkestan neurtuta) eta diru-sarreraren maila erlazionatu nahi izango bagenitu, zera aztertu beharko genuke: hautatutako zer aldagai komunek (adina, sexua, ikasketa-maila) bi aldagai horiekin, biekkin batera, lotura (hots, iragarleak dira) handia duten.

Egoera ezin hobean, laginen arteko *bat-etortze handiko* aldagai komunak hautatu beharko genituzke, hau da, aldagai horiek baliagarriak izan beharko lukete, xede ditugun berariazko aldagaiak iragartzeko. Horrez gain, komeni da aldagai erreduanterik hautaketan ez sartzea (adibidez, adin kategorizatua eta kategorizatu gabea).

Batzuetan, erabilgarri dagoen informazio osoa baliatu ahal izateko, komeni da jatorrizkoetatik eratorritako aldagaiak sortzea.

Hein handi batean, nolakoa den hautaketaren kalitatea, halakoa izango da loturaren arrakasta. Horregatik, hain zuzen, funtsezkoa da fase hori. Lotura-aldagaien hautaketa ahalik eta hoberena egiten laguntzen dute estatistika-tresna batzuek: ikus 2. taulako zerrenda.

ALDAGAIK HAUTATZEKO TRESNAK

- Z: aldagai komuna
- X: A fitxategiko berariazko aldagaia
- Y: B fitxategiko berariazko aldagaia

1. Bat-etortzeko mailaren ebaluazioa

a) Z aldagai kategorikoak

- o Antzekotasun-indizea
- o Gainezarpén-indizea
- o Bhattacharyya-ren koefiziente
- o Hellinger-en distantzia
- o Azterketa grafikoak: barra-diagramak, sektoreak...

Kalkulua: `R > StatMatch > comp.prop()`

b) Z aldagai zenbakidunak

- o Estatistika-zehaztapenak: minimoa, maximoa, batez bestekoa, desbiderapena, estandarra, aldakuntza-koefiziente, ehunekoak
- o Azterketa grafikoak: qqplot-ak, histogramak, zenbatetsi diren dentsitate-funtzio gainjarriak

Kalkulua: R-ko pakete anitzak: esaterako, `Hmisc > describe()`

2. Balio iragarlearen ebaluazioa (garrantzia)

a) Ordinala den X edo Y jarraitua edo kategorikoa; ordinala den Z jarraitua edo kategorikoa

- o Spearman-en korrelazio-koefiziente zuzendua

b) Ordinala den X edo Y jarraitua edo kategorikoa, Z kategoriko nominala

- o Eta-karratuari dagokion determinazio-koefiziente (Kruskal-Wallis-en testuari lotuta)

c) X edo Y kategoriko nominala, Z kategoriko nominala edo ordinala

- o Elkartze-neurriak, estatistikako khi karratuan oinarrituta, adibidez, Cramer-en V
- o Bariantzaren laburketan oinarritutako batez bestekoak (bariantzaren laburketa) edo entropiaren laburketa ()

Kalkulua: `R > Hmisc > spearman2()` edo `StatMatch > pw.assoc()`

3. Erredundantzia ebaluatzea

- a. Erredundantzia aztertzea, antzeko datuak ematen dituzten iragarleak baztertzeko

- b. Miaketa-metodoak, aldagaiak multzokatzean (*clustering*) oinarrituak

Kalkulua: `R > Hmisc > redun(), varclus()`

4. Aldagai anitzeko metodoak

- a. Estatistika-metodo generikoak: aldagai jarraituetarako erregresio-azterketa, sailkapen-zuhaitzak eta erregresio-zuhaitzak (CART), bai eta tipologia mistoetarako *random forest* ere. Kontuz erabili behar dira.

Kalkulua: `R > randomForest`

- b. Inkesten arteko loturarako berariazko metodoak: banaketa bateratuko parametroak zenbatestekoan, ziurgabetasuna gehien laburtzen duten aldagaiak hautatzea.

Aldagaiak kategorikoak baldin badira, *Frèchet* bandak erabil daitezke.

Kalkulua: `R > StatMatch > Fbwidths.by.x()`

Erreferentziak

A. Leulescu et al. (2013)

A. Agresti (2014).

2. koadroa: aldagaiak hautatzeko tresnak

Hirugarren fasean, fitxategi sintetikoa, behar bezala fusionatua, lortuko bada, lotura-metodo egokia aukeratu behar da (mikro kasu-mota), edo, bestela, intereseko estatistika-parametro baten zenbatespena (makro kasu-mota) (Metodo horiek aurreko atalean azaldu ditugu, Hurbilketak eta metodoak). Metodoa aukeratzeko, zer datu erabilgarri ditugun kontuan hartu behar da: esaterako, C fitxategi osagarria (adibidez, duela zenbait urte biztanle-kopuru antzekoaren gainean egin zen inkesta bati buruzkoa) eskuragarri baldin badugu, eta informazio osatua baldin badauka, informazio hori erabil daiteke loturaren emaitza hobetzeko.

Laugarren eta azken fasea, fitxategi fusionatuaren aplikagarritasuna bermatze aldera, emaitzak baliozkotzean datza. Fase hori garrantzitsu samarra denez, hurrengo atalean zehaztasun handiagoz deskribatuko dugu.

Emaizten baliozkotasuna

Inkesten arteko lotura baten baliozkotasuna aztertzerakoan (aplikagarritasun-mailatzat³ hartuta; 2. irudiaren azken fasea da), prozesuko fase guztiak aintzat hartu behar ditugu. Haien artean, honako hauek bereziki garrantzitsuak dira: jatorrizko iturrien kalitatea eta koherentzia; baldintzatutako banaketen gainean egindako suposizioak (hots, baldintzatutako independentzia-hipotesia) eta lotura-metodoa bera.

S. Rässler-ek (2002) lau maila ezarri zituen, lotura baten baliozkotasuna era sistematikoan ebaluatu ahal izateko. Aurrerantzean, kontzeptu horiek argiago azaltze aldera, fitxategi fusionatu bat lortu dugula joko dugu (mikro hurbilketa). Makro hurbilketan era analogoan aplikatuko lirateke alderdi horiek.

Baliozkotasun-mailak

1 maila: banakako balioak ez aldatzea

Lotura egin ostean, Y aldagaien balio errealak (ez-behatuak) fitxategi hartzailean dauden bezala birsortu behar dira: $y_i = \hat{y}_i \ i = 1, 2, \dots, n_A$ baliorako. Honako hauek dira osagaiak: y_i balio errealak (ez-behatuak) dira; eta $\hat{y}_i$ egotzitako balioak dira; eta n_A fitxategi hartzailearen tamaina da. Egotzitako balioa eta balio «erreala» zenbat aldiz datozen bat, hots, igartze-tasa kalkulatu nahi da.

2 maila: banaketa bateratuak ez aldatzea

Lotura egin ostean, X, Y eta Z aldagaien hiru multzoen banaketa bateratu erreala (ez-behatua) behar bezala erakusten da egotzitako fitxategian. Honela: $f(X, Y, Z) = \hat{f}(X, Y, Z)$. Honako hauek dira osagaiak: f balioak irudikatzen du banaketa bateratu behatua eta $\hat{f}$ balioak irudikatzen du egotzitako fitxategiaren gainekoa.

3 maila: korrelazioen egitura ez aldatzea

Lotura egin ostean, fitxategi fusionatuak bere horretan dauzka korrelazioen egitura eta goragoko mailako uneak. Honela: $cov(X, Y, Z) = \widehat{cov}(X, Y, Z)$. Honako hauek dira osagaiak: cov bariantzen eta kobariantzen matrize behatua da; eta $\widehat{cov}$ fitxategi fusionatuaren gainekoa.

³ Kasu honetan, *baliozkotasuna* aipatu da, eta ez *eraginkortasuna*: kasu honetan ez dugu batez besteko errore koadratikoa bezalako baliorik erabiliko, estatistikako beste esparru batzuetan erabili arren. Kasu honetan, helburua erreproduktzio-mailak ebaluatzea da, bai eta jatorrizko banaketak eta elkartzeak ez aldatzea ere.

4 maila: bazter-banaketak ez aldatzea

Lotura egin ostean, fitxategi emailean behatu diren bazter-banaketa bateratuak behar bezala birsortzen dira egotzitako fitxategian. Zehazki, honako hauek bete behar dira: $f(Y) = \hat{f}(Y)$ eta $f(Y|Z) = \hat{f}(Y|Z)$. Honako hauek dira osagaiak: f balioak irudikatzen du bazter-banaketa behatuak, y errealak eta $\hat{f}$ egotzitakoak.

Eztabaida

Kontuan hartzen badugu, oro har, egotzitako fitxategian Y balio «errealak» ez direla eskuragarri egongo (bestela lotura planteatzea ere zentzugabea izango litzateke), ohikoena da baliozkotasunaren **lehen maila** ez aztertzea. Izan ere, maila horrek simulazio azterlan batean baino ez luke zentzurik izango: A fitxategi bat era artifizialean bitan zatitzen da; haietako batean « Y aldagai» batzuk ezabatzen dira; eta Z aldagaiei eutsi behar zaie. Gero, ezabatutako aldagaiak loturaren bidez «berreskuratu» egiten dira; eta igartze-tasa kalkulatu da.

Egoera erreal batean (hau da, simulatu ez dena), oro har, balioduna den fitxategi sintetiko bat lortu nahi da, estatistika-azterketa konbinatuak egin ahal izateko. Egoera horretan baliagarriagoak dira bigarren, hirugarren eta laugarren mailak.

Bigarren mailak (banaketa bateratua ez aldatzea) eta **hirugarren mailak** (korrelazio-egitura edo, era orokorragoan, goragoko mailako uneak ez aldatzea) bermatzen dute fitxategi sintetikoaren baliozkotasuna, hots, biztanleriari buruzko aldagaien gaineko erlazio errealari buruzko estatistika egokiak erakusteko gaitasuna. Hala eta guztiz ere, maila horiek ezin dira zuzenean alderatu, arrazoi berarengatik: definizioz, ez da ezagutzen nolakoa den $f(X, Y, Z)$ banaketa edo korrelazioaren egitura.

Izan ere, praktikan, **laugarren maila** baino ezin da zuzenean erkatu. Maila hori betetzeak ziurtatzen du bazter-banaketak, fitxategi emailean behatuak, fitxategi hartzailean ondo birsortzen direla. Prozedura sendoak eta kalitatezko datuak erabiliz gero, maila hori lortzea erraz samarra da; horregatik, lotura-ariketa guztiek bete behar duten gutxieneko eskakizuna da (A. Leulescu et al., 2013).

Adierazi bezala, baliozkotasun-maila betetzen duen fitxategi sintetiko bat sortzeak ez du esan nahi fitxategi horrek batera behatu ez diren aldagaien arteko erlazio zuzenak automatikoki erakutsiko dituenik. Horregatik, beste lan bat ere egin behar da, eta datu laguntzaileak izatearen mende dago. Datu laguntzaileak eskuragarri baldin badaude, datu horiek loturan gehitu behar dira, fitxategi sintetikoaren baliozkotasuna areagotzeko. Bestela, datu gehigarriak ez dagoenean, ziurgabetasuna aztertu behar da.

Datu laguntzaileek lotura bat hobetzeko duten garrantzia aintzat hartuta, hurrengo atalean berariaz landuko ditugu.

Datu laguntzaileak gehitzea

Frogatu egin da datu laguntzaileek lotura baten emaitzak nabarmen hobetu ditzaketela. Adibidez, M. D'Orazio et al. (2008) azterlanean, hileko diru-sarreraren dezilak erabiltzen dira diru-sarreraren eta gastuen gaineko zenbatespen batzuk, zehatzagoak, hobetzeko.

Datu laguntzaileak lortzeko iturriak askotarikoak izan daitezke. Hona hemen haietako batzuk:

- a. Datu osagarriak dituen C fitxategi bat, ziur asko lehengo urteetakoak izango diren (X,Y,Z) behaketak –edo beste iturri independente batzuetatik etorritakoak– dituenak.
- b. Datu parametrik laguntzaileak, kanpoko zenbatespen baten bidezkoak.
- c. Aztergaiari buruzko alde zurrerako informazioa. (Ohikoenak dira aldagaiek har ditzaketen balioen gaineko murrizketa logikoak).

StatMatch paketeko funtzio batzuek datu laguntzaileak sartzeko aukera ematen dute. Adibidez, `comb.samples()` funtzioa diseinatu dago fitxategi bereizietan behatu diren aldagai kategorikoetarako kontingentzia-etaulak diseinatzeko. C fitxategi osagarri baten informazioa, behaketa bateratuak barne, `svy.C` parametroaren bitartez sar daiteke, zenbatespenak hobetzeko xedez.

Pakete bereko beste funtzio bat `mixed.mtc()` da. Funtzio horrek `rho.yz` parametroa⁴ onartzen du, batera behatu ez diren aldagaien arteko korrelazioari buruzko *a priori* zenbatespen bat (kanpoko) lortu ahal izateko.

Ziurgabetasuna aztertzea

Datu laguntzailerik ez izanez gero, komeni da lotura baten ziurgabetasuna zenbatesteko metodoak erabiltzea.

Esku artean dugun testuinguruan, *ziurgabetasun* hitzak esan nahi du batera behatu ez diren aldagaien arteko erlazioetarako balioen balizko hein bat dagoela, behatutako datuekin bateragarria dena. *Identifikazio-eza* ere esaten zaio.

Lehen ikusi dugun bezala, fitxategietako berariazko aldagaien eta aldagai komunaren arteko erlazioa zenbat eta estuagoa izan, hainbat eta txikiagoa izango da loturaren ziurgabetasuna. Aukera bat baino gehiago daude lotura baten ziurgabetasuna aztertzeko:

- Aldagaiak kategorikoak direnean, *Frèchet bandak* kalkula daitezke. Banda horiek gorako eta beheko kotak ematen dituzte kontingentzia-tauletako gelaxketarako. Koten arteko bitarteak behatutako datuekin bateragarriak diren

⁴ Koaderno honetako notazioa eta StatMatch-ena ez datoz bat: Z aldagai komunaren denotazioa da; eta X e Y berariazko aldagaiak.

balio guztiak ditu. `StatMatch` paketea, kalkulu hori `Frechet.bounds.cat()` funtzioan ezarrita dago.

- Egozte anitza testuinguru naturala da ziurgabetasuna ebaluatzeko. Izan ere, tresna horren bitartez eta, bereziki testuinguru bayestar batean⁵, Z-ren arabera baldintzatuta dauden erlazioei buruzko hasierako hipotesiak eta emaitzak azter daitezke. Bide horri ekiteko, `mice` paketea erabil daiteke (Buuren, S. eta Groothuis-Oudshoorn, K., 2011).

⁵ S. Rässler (2002) liburuan gaia sakon aztertuta dago.

Softwarea

Koaderno honetan aurkeztu diren metodo asko ezarrita daude R estatistika-konputazioaren inguruneke hainbat paketetan. Pakete horietako batzuk estatistika ofizialetara bideratuta daude (*StatMatch*, adibidez) edo inkesta-datuen azterketetara (*survey*, adibidez). Beste batzuk, ordea, (*Hmisc*) generikoak dira, eta askotariko funtzioak ematen dituzte datuak aztertzeko.

3. koadroan R ingurunean erabilgarri dauden paketeen zerrenda ez-exhaustiboa bildu da. Pakete horiek zenbait funtzio dituzte inkesta independenteak lotu ahal izateko.

R INGURUNeko PAKETEAk, INKESTA INDEPENDENTEAK LOTZEKO

- *StatMatch*: Marcello d'Orazio-k (ISTAT) aurkitua. Pakete hori sortu zen Europako Estatistika Sistemari (European Statistical System) egin ziren datuen integrazioari buruzko bi proiektuen ondorioz. Ikus ESSnet⁶. Paketea bereziki pentsatuta dago inkesta independenteetako datuak lotzeko eta egozteko. Zenbait funtzio ditu, inkestak lotzeko fase batzuei heltzen dietenak. Honako hauek dira funtzio nagusiak:
 - *hot-deck* egozte-metodo ez-parametrikokoak,
 - *predictive mean matching* metodoan oinarritutako metodo mistoak,
 - lagin konplexuekin lan egiteko metodoak,
 - lotura baten testuinguruan ziurgabetasuna aztertzeko metodoak.
- *survey*: Thomas Lumley-k egina. Tresna-sorta zabala dauka, lagin konplexuetako datuak aztertu ahal izateko: estatistika-datu deskriptiboak, testak, eredu lineal orokortuak, Cox ereduak, azterketa faktorialak eta osagai nagusien azterketak, etab.
- *Hmisc*: *Harrell Miscellaneous*, Frank E Harrell Jr.-ek egina, Charles Dupont-en ekarpenekin batera. Datuen azterketaren zenbait alderdi lantzen dituzten funtzioak ditu: grafiko aurreratuak, taulak sortzea, aldagaien *clustering*-a, karaktereen bektoreak manipulatzeko, aldagaiak berriro kodetzea.
- *mice*: *Multiple Imputations via Chained Equations*, Stef van Buuren-ek egina. Egozte sinplea erabiltzen du, baldintzatutako espezifikazio osoan oinarrituta (FCS, *Fully Conditional Specification*), MICE algoritmoak ezarrita. Hauxe da haren funtsa: aldagai bakoitzari bere egozte-eredua esleitzea. Paketeak ereduak ematen ditu, zenbait aldagai-motatarako. Hona hemen aldagai-motak: aldagai jarraituak (*predictive mean matching*, normala); aldagai dikotomikoak (erregresio logistikoa); ordenarik gabeko aldagai kategorikoak (erregresio logistikoa multinomiala) eta aldagai kategoriko ordinalak (*odd* proportzionalak). Paketeak diagnostiko-grafikoak ere ematen ditu. Grafiko horien bitartez, egozteen emaitzak berrikusi daitezke.

⁶ Proiektuak *Data Integration* (12/2009-12/2011) eta ISAD dira. *Integration of Survey and Administrative Data* (12/2006-06/2008), biak ISTAT-ek zuzenduak.

- *Amelia: Amelia II: A Program for Missing Data*. Egileak: James Honaker, Gary King eta Matthew Blackwell. Inkesten egozte anitza egiteko funtzioak ditu, eta *bootstrap* teknikan oinarritutako algoritmo batean ezarrita dago. Egileek beraiek sortu zuten: antzekoak diren beste soluzio batzuk baino aurreratuagoa da, eta aldagai batzuk aldi berean erabiltzeko aukera ematen du. GUI edo erabiltzaile-interfaze grafikoa dauka, eta R erabiltzen ez duten erabiltzaileek ere erabil dezakete.
- *BaBooN: Bayesian Bootstrap Predictive Mean Matching - Multiple and single imputation for discrete data*, Florian Meinfelder-ek egina. *Bayesian Bootstrap Predictive Mean Matching* algoritmoaren bi bertsio ditu, falta diren datuen egozte sinplea egiteko. Komeni da bigarren aldaera erabiltzea inkesten arteko loturak sortzeko bezalako kasuetan (edo datuak fusioatzeko); eta, orokorrean, aldagaiek falta diren datuen eredu bera dutenean, egokia da bigarren aldaera erabiltzea.

Erreferentziak

`StatMatch`

Marcello D'Orazio (2013). *StatMatch: Statistical Matching (data fusion* deiturikoa). <http://CRAN.R-project.org/package=StatMatch>

D'Orazio, M. (2013). Statistical Matching and Imputation of Survey Data with StatMatch: StatMatch Bineta.

`survey`

Thomas Lumley (2012) *survey*: analysis of complex survey samples. <http://CRAN.R-project.org/package=survey>

`mice`

Stef van Buuren, Karin Groothuis-Oudshoorn (2011). *mice*: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 1-67. URL-a: <http://www.jstatsoft.org/v45/i03/>
www.multiple-imputation.com

`Hmisc`

Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2014). *Hmisc*: Harrell Miscellaneous. R package version 3.14-3. <http://CRAN.R-project.org/package=Hmisc>

`Amelia`

James Honaker, Gary King, Matthew Blackwell (2011). *Amelia II: A Program for Missing Data*. *Journal of Statistical Software*, 45(7), 1-47. URL <http://www.jstatsoft.org/v45/i07/>.

`BaBooN`

Florian Meinfelder (2011). *BaBooN: Bayesian Bootstrap Predictive Mean Matching – Multiple and single imputation for discrete data*. <http://CRAN.R-project.org/package=BaBooN>

3. koadroa: R paketeak, inkesta independenteak lotzeko.

5. kapituluari, R pakete propioa garatzea deiturikoan, *micromatch* pakete propioa aurkeztu da. Aurreko paketeetan oinarrituta dago, eta proiektu honetan garatu da; besteak beste, Eustat-eko web-aren bidez dago eskuragarri.

Ezarpen praktikoa

Eustat-ek eginiko bi lagin-inkesten arteko lotura: Bizi-baldintzen inkesta eta Biztanleria jardueraren arabera sailkatzeko inkesta

Azterlan honen bitartez, Metodologiari eskainitako atalean azaldu diren teknikak frogatu nahi ditugu (6. orr.). Horretarako, bi lagin-inkesta independente, Eustatek eginak eta errealak, erabiliko ditugu: Bizi-baldintzen inkesta (2009) eta Biztanleria jardueraren arabera sailkatzeko inkesta (2009ko laugarren hiruhilekoa).

Bi inkesta horietako datu eskuragarriak erabili ditugu, eta loturaren fase nagusiak urratsez urrats azalduko ditugu: lehena helburuak zehaztea izango da; ondoren, iturrien eta lotu beharreko aldagaien hautaketaren arteko koherentzia aztertuko dugu; eta, azkenik, emaitzen baliozkotasun eta aurkezpen laburrak egingo ditugu. Erakutsiko diren kalkulu guztiak software libreko R ingurunearen bidez egin dira.

Inkestak deskribatzea

Biztanleria jardueraren arabera sailkatzeko inkesta (aurrerantzean, BJA) hiruhileko panel jarraitua da. Eustat-ek panel hori garatu du 1985. urteaz geroztik. Helburua Euskal Autonomia Erkidegoko lan-merkatuaren ezaugarriak eta dinamika ezagutzea da. BJA inkestak etxebizitza-panel baten gaineko probabilitatezko laginketa egiten du, eta hiru hilez behin aldatzen da (ikus Bibliografia atalean: metodologia-fitxa). Egungo laginketaren tamaina gutxi gorabehera 5.000 etxebizitzakoa da (eta 13.500 pertsona inguru barne hartzen ditu). Errotazioa hiruhileko baten zortzirenekoa da.

Inkestak bi helburu nagusi ditu:

- Estatistika-informazio jarraitua ematea. Euskadiko biztanleek ekonomia-jardueretan izaten duten parte-hartzearen arabera, kolektibo nagusi batzuk zehaztuta daude; eta inkestako datuak kolektibo horien bolumenari eta ezaugarriei buruzkoak dira. Horrez gain, pertsonen egoera-aldaketak ere kontuan hartzen dira.
- Estatistika-informazioa ematea: biztanle horien ezaugarri demografikoei eta sozialei buruzkoa, bai eta etekin ekonomikorik ez dute jardueretan biztanle horiek izaten duten parte-hartzeari buruzkoa ere.

BJA inkestaren emaitza garrantzitsuenetakoa biztanleak jardueraren arabera sailkatzea da, eta sailkapen hori landunak, inaktiboak eta langabeak bereizten dituen sailkapena baino zehatzagoa da (ikus 1. taula)

BJA INKESTAKO SAILKAPEN-SISTEMA

- **Jarduera egiten duten biztanleak:** ondasunak eta zerbitzuak sortzeko jarduerak egiten dituztenak, etxeko lanak nahiz ikasketak egiten dituztenak edo soldaduskan daudenak. Bi taldetan banatzen dira:
- = **Lan-jarduera egiten duten biztanleak:** lanpostu bat duten pertsonak. Halaber, lana esaten zaio ordainsaria edo etekina dakarren jarduerari, hau da, enplegatzaile-enplegatuaren arteko harremana duen ordaindutako lanari edo lan independenteari. Era berean, familiako lana, ez ordaindua, izan daiteke (familien arteko laguntzak).
 - = **Jarduera ez-laborala egiten duten biztanleak:** lan bila dabiltzalarik, etxeko lanak nahiz ikasketak egiten dituztenak edo soldaduskan daudenak dira. Pertsona horiek ez dute enplegu-emaila baten kontura lan egiten. Lan bila dauden ala ez sailkatuta daude, bai eta lan batean hasteko prest egoteko joera subjektiboaren arabera ere. Bi azpitalde bereizi dira:
 - ≡ **Jarduera ez laboralak egiten dituzten biztanleak:** lan bila ez dabiltzanak, eta etxeko lanak, ikasketak edo soldaduska egiteari soilik ekiten diotenak.
 - ≡ **Gainerakoak:** lan-jarduerarik ez dutenak, lan bila dabiltzanak.
- **Jarduerarik gabeko biztanleak:** bi azpitaldetan banatzen dira:
- ≡ **Soilik langabeak:** lan bila dabiltzanak, eta lanpostu bat segituan hartzeko prest daudenak.
 - **Soilik erretiratuak eta beste batzuk:** adina edo egoera fisikoa zein den, jarduerarik egiten ez duten pertsonak (pentsiodunak, lanerako ezinduak, etab.).

1. taula. *Biztanleen segmentazioa, jardueraren arabera, BJA inkestan jasotakoari jarraituta. Galdera-sortan parte hartu duten pertsonen datuak biztanle guztiei estrapolatzen zaizkie.*

Bizi-baldintzen Inkesta (aurrerantzean BBI), lagin-inkesta da, eta 1989. urteaz geroztik Eustat-ek bost urtez behin egin du. Inkestaren helburua Euskadiko AEko familien, banako pertsonen eta ingurunearen bizi-baldintzen gaineko informazio zehatza ematea da.

BBI inkesta egiteko, bi galdeketa-mota erabiltzen dira (bata banakakoa da eta bestea familiakoa). Bi etapatan estratifikatu den ausazko laginketan oinarritzen da (ikus Bibliografia atalean: metodologia-fitxa). Lehenengo etapan estratu (esparru geografiko) bateko etxebizitzak hautatzen dira. Familiako galdera-sorta horiei buruzkoa izango da. Bigarren etapan, etxebizitzako pertsona bat ausaz hautatzen da, banakako galdera-sorta erantzun beharko duena. Laginketaren hasierako tamaina 7.500 etxebizitzakoa da.

BBI-k hiru berariazko helburu ditu, eta hauexek dira:

1. Pertsonen osasun-, ikasketa-, lan-, aisia- eta gizarte-harremanen berri izatea
2. Pertsonen bizilekuaren ingurune fisikoaren eta sozialaren egoera deskribatzea
3. Familiaren barne-harremanak eta baliabide ekonomikoak aztertzea, baita etxeko ekipamenduak ere

BBI-BJA arteko lotura

Hurrengo ataletan azalduko dugu Eustaten BBI eta BJA inkesten arteko loturaren faseak. Horretarako, 3. irudian jasotako eskemari jarraituko diogu (Inkesten arteko lotura baten faseak atala).

Datu erabilgarriak eta erreferentziazko biztanleak

Koaderno tekniko hau idazterakoan, BBI-ren gaineko datu argitaratuak eta erabilgarriak 2009ko azken hiruhilekoak dira. Hori dela eta, lotura sortzeko, BJA inkestan denboraldi bereko datuak hautatu eta erabili egin dira.

Inkesta batean eta bestean 16 urteko edo urte gehiagoko pertsonak hautatu dira. Hori horrela, gure erreferentzia-biztanleria 16 urteko edo urte gehiagoko pertsonak dira, eta Euskadiko AEn bizi dira 2009ko azken hiruhilekoan. Erabilgarri dagoen laginketak 12.658 behaketa ditu BJA inkestan; eta BBIn 5.242 behaketa. 16 urtetik beherako pertsonak baztertu direnez, 10865 eta 4749 behaketa dira, hurrenez hurren.

1. fasea: loturaren helburua zehaztea

Hau da azterlan honen helburua: estatistika integratuak, bizimoduaren zein bizi-baldintzen gaineko aspektuak eta lan-merkatua erlazionatzen dituztenak, sortu ahal diren aztertzea. Horretarako, era independentean BJA eta BBI inkestek eman duten informazioa abiapuntutzat hartuko da.

Zehatzago esanda, bi inkesta horietako aldagaiak konbinatuta, fitxategi sintetiko bat lortu nahi dugu. Horretarako, *hot-deck* metodo baten bitartez, datuak egotziko dira: BBI inkesta hartzailatzat hartuko da, eta BJA inkesta emailatzat. BJA inkestaren aldagai nagusia –1. taulan adierazi dugun segmentazioa– BBI inkestako fitxategiari egotzi zaio.

Horren arabera, BBI inkestari (jasotzen dituen item guztiak barne hartuta, hala nola ikasketa-maila, osasun-egoera, gizarte-harremanak, ingurumena, ekonomia-egoera...) gehitu zaio BJA inkestaren aldagai nagusia, hots, biztanleriaren segmentazioa, jardueraren arabera (1. taula). Fitxategi sintetiko horrek aukera emango du askotariko bizi-baldintzak aztertzeko, lan-merkatuaren segmentazioa zein den. Azterlan hori orokorrean ezin da zuzenean egin, aldagaiak inkesta (fitxategi) independenteetakoak baitira.

Azterlan honen helburua errazago azaldu aldera, BBI inkestaren sei berariazko aldagai hautatu dira, eta inkesta horretan jasotako dimentsio nagusiei heltzen diete, ikus 2. taula.

(Eranskinean: A fitxako taulan bildu dira aldagai horien jatorria eta trataera, mikrodatuen fitxategiak aintzat hartuta).

BBI INKESTAREN BERARIAZKO ALDAGAIK (laginketa)

- Osasun-arazoak: {1-arazoren bat; 2-arazorik ez}
- Hizkuntzak jakitea: {1-gaztelania soilik; 2-gaztelania eta beste batzuk; 3-gaztelania eta euskara; 4-gaztelania, euskara eta beste batzuk}
- Aisialdia: {1-bi ordutik behera; 2 ordukoa edo 2-4 ordu artekoa; hiru ordukoa-4 ordutik gorakoa}
- Egoera ekonomiko objektiboa: {1-txarra; 2-normala; 3-ona}
- Motordun ibilgailuak edukitzea: {1-bat ere ez; 2-bat; 3-bi edo gehiago}
- Etxeko ekipamendua^a: {1-ekipamendu eskasa; 2- behar beste ekipamendu }

^a Maila agregatuak

2. taula. BBI inkestaren berariazko aldagaiak, kategoriarekin batera. Ikus eranskina: A fitxa.

2. fasea: lotura-aldagaiak hautatzea

Fase honen helburua aldagaien azpimultzo optimoa zehaztea da (aldagaiak komunak edo lotura-aldagaiak dira). Aldagai horiek BBI eta BJA inkestek partekatzen dituzten aldagai guztien (jatorrizkoak edo eratorriak) artean hautatuko dira.

2-1 fasea: galdera-sorten gaineko meta-analisisa

Informazio bera jasotzen duten (edo jaso lezakete) aldagaiak zehazte aldera, galdera-sorten gaineko meta-analisisa egingo da. Helburua BJA segmentazioa BBI inkestari egozte delarik aintzat hartuta, ohiko aldagai soziodemografikoekin (adina, sexua edo familiaren tamaina) batera, berariazko aldagaiak ere, lan-jarduerari buruzko datuak ematen dituztenak, bilatuko dira. (Aldagai horiek *proxy* edo «berariazko aldagai komunak» deitzen dira). Gure azterlanean aldagai horiek bilatzeko aukera dugu, BBI deritzon inkestako «Lan-baldintzak» atala dugulako, eta atal horretatik adierazle batzuk atera ahal direlako.

Hortaz, honako aldagai komun hauek identifikatu dira BBI eta BJA inkestetako 2009ko 4. hiruhilekoko datuen artean:

Aldagai soziodemografikoak

- Adina
- Sexua
- Familiaren tamaina

Ikasketa-mailei lotutako aldagaiak

- Ikasketa arautuak egiten ari diren pertsonak (aurrerantzean, «Ikasle S/N»)
- Urrutiko ikasketak egiten ari diren pertsonak

- Analfabetoak (ez dakite irakurtzen edo idazten)

Jarduerari lotutako aldagaiak

- Langabeak/landunak/inaktiboak identifikatzea
- Landunek egin dituzten lan-orduak
- Lan bila dabiltzan pertsonak (aurrerantzean «Lan bila S/N»)
- Etxeko lanen arduradunak

Datu horiek A eranskinetako B fitxategian bildu dira. Horrela, irakurleak jatorriraino jarraitu ahal izango die datuei, eta aldagaien kodeak mikrodatuetako fitxategietan nola tratatu diren ere jakingo du.

2-2 fasea: koherentzia aztertzea

Meta-analisisa egin ondoren, galdera-sorta bietan identifikatu diren aldagaien koherentzia aztertuko dugu, behatutako bazter-banaketetan oinarrituta. Hasieratik, bi aldagai baztertu dira: «analfabetoak» eta «urrutiko ikasketak egiten ari diren pertsonak», agertzeko probabilitatea oso txikia baita. Bazter-banaketen koherentzia era globalean (beste aldagai batzuk kontuan hartu gabe) aztertu da, baita zehaztutako talde batzuen arabera ere. Talde horiek hamabi estratuk zehaztu dituzte, eta estratuek adinaren eta sexuaren arteko gurutzeak edo balio zehatzak dituzte (ikus 3. taula).

Adina (urt.)	Sexua	
	G: Gizonak	E: Emak.
16-24	G.16-24	E.16-24
25-34	G.25-34	E.25-34
35-44	G.35-44	E.35-44
45-54	G.45-54	E.45-54
55-64	G.55-64	E.55-64
65+	G.+65	E.+65

3. taula. Ikerlan honetan aztertu ditugun adinaren eta sexuaren arabera estratuen kodetzea

Ezinbestekoa da estratuak bereiztea: aldagai gehienetarako, adin- eta sexu-talde bakoitzean inplikazio ezberdinak dituztelako. Adibidez, G. +65 eta E. +65 estratuei dagokienez, «Ikasle S/N» aldagaiak ez die informaziorik gehitzen; beraz, loturatik kanpo egon beharko luke. G. 16-24, E. 16-24 estratuei dagokienez, aldiz, ezinbesteko aldagaia da.

A. Leulescu et al. (2013) lanean jasotako aholkuei jarraituta, behatutako banaketak erkatzeko, neurketa enpiriko batzuk erabili ditugu. Azterlan honetan Hellinger-en distantzia erabili dugu: 0 (banaketa berdina) eta 1 (banaketa ahalik eta antzekoenak) arteko balioak erabiltzen ditu. Emaitzak –orokorrean eta adinaren eta sexuaren arabera– eranskinetako C fitxategian daude.

2-3 fasea: iragarpen-balioa aztertzea

Aldagaien hautaketa osatzeko, aldagaiek berariazko aldagaien gainean duten iragarpen-balioa aztertu dugu. Horretarako, lotura osatzeko baliozko informazioa ematen duten aldagaiak hautatu behar ditugu.

Koherentzia aztertzeko egin dugun bezala, aldagai komunak iragarpen-gaitasuna aztertu da, bai era globalean (behaketa guztiak barne hartuta), bai sexuaren eta adinaren arabera zehaztutako hamabi estraturen bidez. BBI-BJA kasuan bezala, aldagai guztiak kategorikoak dira, eta khi karratuan oinarritutako elkartze-neurketak erabili ditugu (Cramer-en V, adibidez). Berrir, eranskinen C fitxategian bildu ditugu emaitzak.

3. fasea: lotura-metodo bat erabiltzea

Azkenik, *hot-deck* metodo bat erabili dugu estratu bakoitzean, hautatutako aldagaiak oinarri hartuta. Segidan, prozedura azalduko dugu, estratu bat erreferentziatzat hartuta: 25 eta 34 urteko gizonak (adin biak barne).

Adibidea: *Hot-deck* metodoa G. 25-34 estratuan

Honako hauek dira segmentu horretarako hautatu ditugun aldagaiak: «Lan bila S/N» eta «Etxeko lanen arduradunak». «Familiaren tamaina» deritzon aldagaia bat datorren arren, ez du informaziorik ematen segmentu horretan, ia kasu guztietan familiak kide bakarrekoak baitira (TF=1). «Landunak» deritzon aldagaian ez dago behar adinako bat-etortzerik; eta «Langabeak» deritzon aldagaia erredundantea da, «lan-bila» aldagaia kontuan hartzen badugu.

Fitxategi hartzaileak 359 erregistro ditu eta emaileak 746. Erregistro hartzaile bakoitzari (BBI erantzun duten 25-34 urte bitarteko gizonak, esaterako), erregistro emaile bat bilatuko diogu (BJA erantzun duten 25-34 urte bitarteko gizonak, esaterako). Erregistro emaileak ahalik eta antzekoena izan beharko du, hautatuko aldagaien artean. Kodea R > `StatMatch` paketea 4. taulan erakutsita dago:

ADIBIDEA: HOT-DECK EGOZTEA R > `StatMatch` PAKETEAN

1. urratsa: emaile-hartzaile bikoteak bilatzea

```
out.nnd <- NND.hotdeck(data.rec = rec, data.don = don,
  dist.fun = "Gower", match.vars = c("BUSQ", "DOM"),
  constrained = TRUE)
```

2. urratsa: fitxategi hartzailea egozte (hots, `ecv`, hautatutako aldagaien bitartez filtratua). Oharra: "PRA", BJA inkestaren sigla da, gaztelaniazko izenaren arabera; eta "ecv" BBI inkestarena.

```
fecv <- create.fused(data.rec = rec, data.don = don,
```

```
mtc.ids = out.nnd$mtc.ids, z.vars = "PRA")
```

Honako hauek dira osagaiak:

- `rec`: fitxategi horrek BBI inkestaren emaitza filtratuak ditu (25 eta 34 urte bitarteko gizonak)
- `don`: fitxategi horrek BJA inkestarako filtratu diren erregistroak ditu (berdin)
- `match.vars`: hautatu diren aldagai komunen zerrenda (adibidean: BUSQ eta DOM).
- `z.vars`: berariazko aldagaien zerrenda; kasu honetan, BJA, bakarra.
- `NND.hotdeck()`: antzeko emaileak bilatzea
- `dist.fun = "Gower"`: Gower-en distantzia erabili dugu (ikus paketeo dokumentuak).
- `constrained = TRUE`: horrek esan nahi du algoritmoa mugatuta dagoela. Adibidez, erregistro bakoitza behin bakarrik erabiltzen da
- `create.fused()`: bbi fitxategiak (hots, `ecv` fitxategia) sortzen duen funtzioa, egotzitako balioak erabilia.
- `mtc.ids`: emaile-hartzaile korrespondentzia dauka, fusionatutako fitxategia sortu ahal izateko.
- `z.vars`: egotzitako aldagaiak (kasu honetan aldagai bakarra dugu: «PRA»).

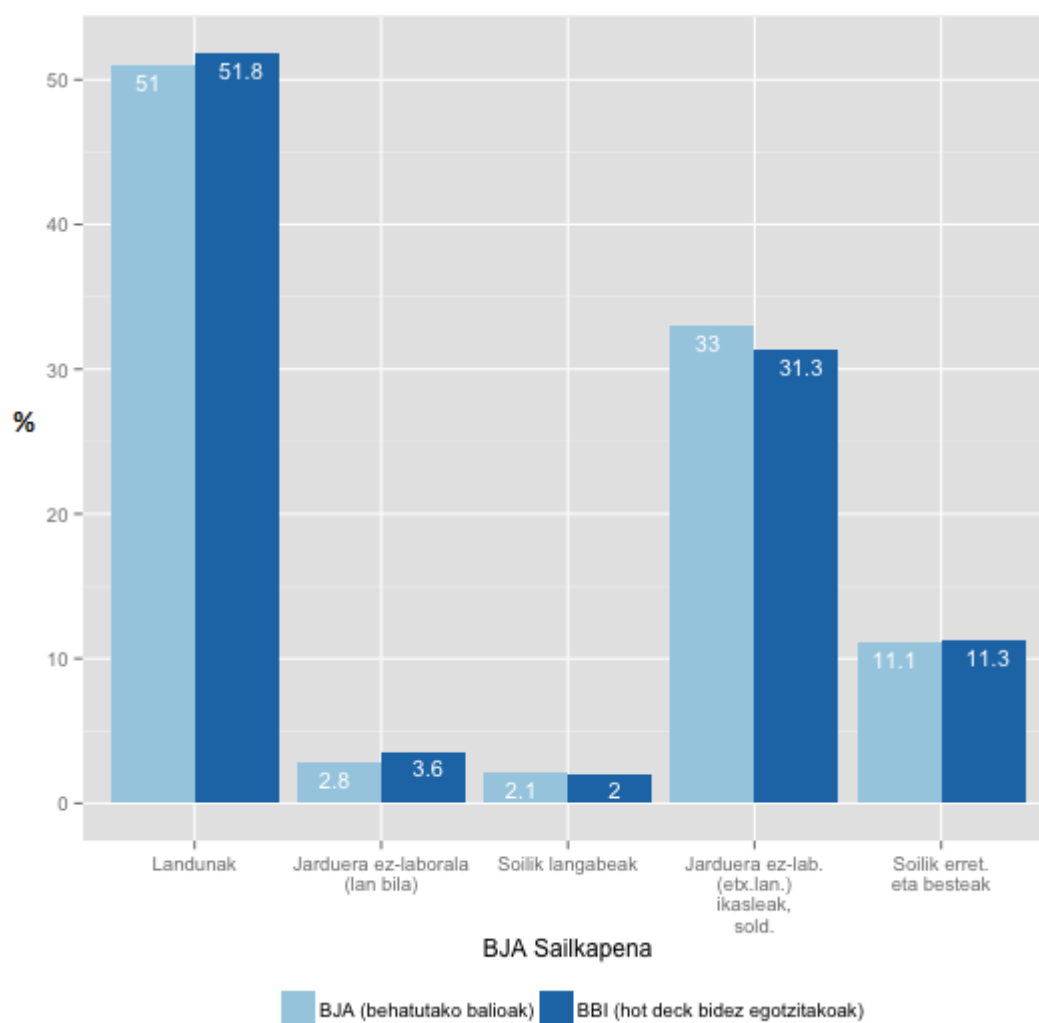
Emaita

`fecv`: `rec` hasierako fitxategia, PRA aldagai bitartez handitua.

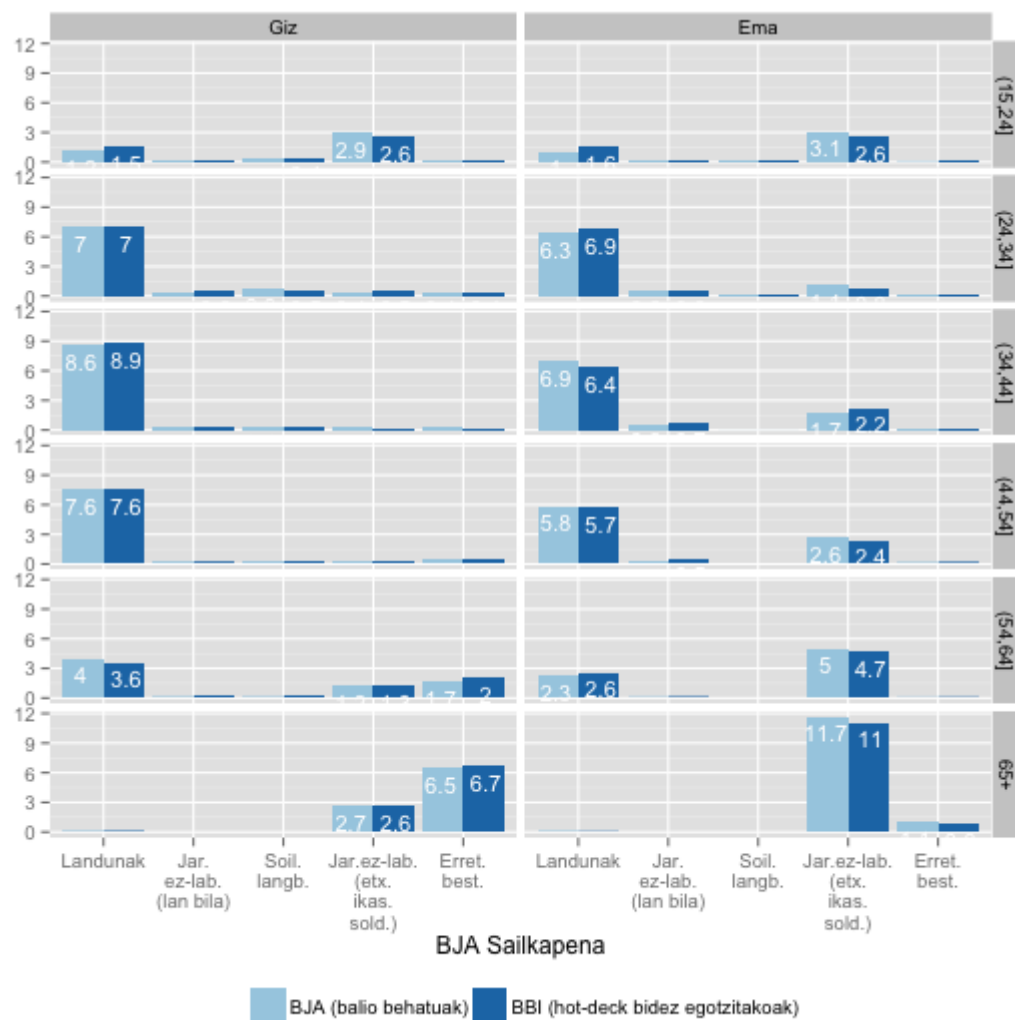
4. koadroa: Adibidea: *hot-deck bitartezko egozte*a, BBI-BJA lotura sortzeko, G. 25-34 segmentuan: (25-34 urteko bitarteko gizonak).

4. fasea: emaitzen kalitatea ebaluatzea

Azkenik, emaitzak ebaluatuko dira, behatutako bazter-banaketak eta egotzitakoak aztertuta (baliozkotasunaren laugarren maila). Emaita globalak 4-1 irudian erakutsita daude (emaitza globalak) eta 4-2 irudian (adinaren eta sexuaren arabera emaitzak).



4-1. irudia: BBI-BJA loturaren emaitzak: Behatu den bazter-banaketa erreala (BJA inkesta emalea), egotzitakoaren aldean (BBI inkesta, hartzailea).



4-2. irudia: BBI-BJA loturaren emaitzak, adinaren eta sexuaren arabera: Behatu den bazter-banaketa erreala (BJA inkesta emaitza), egotzitakoaren aldean (BBI inkesta, hartzailea).

Emaitzak

Zehaztu ditugun helburuak eta hautatu ditugun berariazko aldagaiak aintzat hartuta, lortutako emaitzak zenbait kontingentzia-aula dira. Taulak aurrerago jasota daude. Bi emaitza-mota ditugu: biztanleria osoari buruzkoak eta adinaren eta sexuaren arabera zenbait estratuturi buruzkoak. Estratu horiek adibide gisa hautatu ditugu.

Taula horiek baliagarriak dira, lan-merkatuaren segmentazioa (BJA inkestak adierazita) zein den, bizi-baldintzak (BBI inkestak emandakoak) aztertzeo bide ematen dutelako. Jatorriz, aldagai horiek fitxategi bananduetan daudenez, eta datu gehigarriak ez dagoenez, datuak «gurutatu» nahi baditugu, inkestak lotzeo teknikaren bat erabili beharoko dugu. Kasu honetan erabili dugun estrategia hauze izan da: BJA inkestaren (emaila) aldagai nagusia egotzi diogu BBI inkestari (hartzailea), *hot-deck* egoztearen bidez, adin- eta sexu-estratuen arabera.

Kontingentzia-taulak, bizi-baldintzen inkestako (jatorria: BBI) itemak aztertzeko, jarduerarekin (jatorrizkoa: BJA) dituzten lotura zein den.

Aldagaia: osasun-baldintzak

Iturria: BJA	Iturria: BBI		
	1-arazoren bat	2-arazorik ez	Guztira
Landunak	154.204	806.109	960.312
Jarduera ez-laborala (lan bila)	14.633	51.196	65.830
Soilik langabeak	7.553	30.079	37.633
Jarduera ez-laborala (etxeko l., ikas., sold.)	181.454	399.219	580.673
Soilik erretiratuak eta beste batzuk	86.556	121.987	208.543
Guztira	444.400	1.408.590	1.852.991
errenkada %			
	1-arazoren bat	2-arazorik ez	Guztira
Landunak	16,1%	83,9%	100%
Jarduera ez-laborala (lan bila)	22,2%	77,8%	100%
Soilik langabeak	20,1%	79,9%	100%
Jarduera ez-laborala (etxeko l., ikas., sold.)	31,2%	68,8%	100%
Soilik erretiratuak eta beste batzuk	41,5%	58,5%	100%
Guztira	24,0%	76,0%	100%
zutabe %			
	1-arazoren bat	2-arazorik ez	Guztira
Landunak	34,7%	57,2%	51,8%
Jarduera ez-laborala (lan bila)	3,3%	3,6%	3,6%
Soilik langabeak	1,7%	2,1%	2,0%
Jarduera ez-laborala (etxeko l., ikas., sold.)	40,8%	28,3%	31,3%
Soilik erretiratuak eta beste batzuk	19,5%	8,7%	11,3%
Guztira	100%	100%	100%

1-1 emaitza. Bizi-baldintzak (iturria: BBI), BJA segmentazioarekin erkatuta (iturria: BJA). Guztizkoak biztanleria osorako; eta errenkaden eta zutabeen ehunekoak.

Aldagaia: osasun-baldintzak			
45-54 urte bitarteko adina	Iturria: BBI		
Iturria: BJA	1-arazoren bat	2-arazorik ez	Guztira
Landunak	45.002	201.485	246.487
Jarduera ez-laborala (lan bila)	2.651	1.046	3.698
Soilik langabeak	1.178	3.370	4.548
Jarduera ez-laborala (etxeko l., ikas., sold.)	181.454	42.995	224.449
Soilik erretiratuak eta beste batzuk	1.467	8.534	10.002
Guztira	231.753	257.430	489.184
errenkada %	1-arazoren bat	2-arazorik ez	Guztira
Landunak	18,3%	81,7%	100%
Jarduera ez-laborala (lan bila)	71,7%	28,3%	100%
Soilik langabeak	25,9%	74,1%	100%
Jarduera ez-laborala (etxeko l., ikas., sold.)	80,8%	19,2%	100%
Soilik erretiratuak eta beste batzuk	14,7%	85,3%	100%
Guztira	47,4%	52,6%	100%
zutabe %	1-arazoren bat	2-arazorik ez	Guztira
Landunak	19,4%	78,3%	19,4%
Jarduera ez-laborala (lan bila)	1,1%	0,4%	1,1%
Soilik langabeak	0,5%	1,3%	0,5%
Jarduera ez-laborala (etxeko l., ikas., sold.)	78,3%	16,7%	78,3%
Soilik erretiratuak eta beste batzuk	0,6%	3,3%	0,6%
Guztira	100%	100%	100%

1-2 emaitza. Bizi-baldintzak (iturria: BBI), BJA segmentazioarekin erkatuta (iturria: BJA). Hautatu den estratua: 45 eta 54 urte bitartekoak (biak barne). Errenkaden eta zutabeen ehunekoak eta guztizkoak.

Aldagaia: hizkuntzak jakitea

Iturria: BJA	Iturria: BBI				Guztira
	1-gazt. soilik	2. gazt. + beste batzuk	3. gazt. + eu	4. gazt. + eu + beste batzuk	
Landunak	310.088	181.823	165.623	302.778	960.312
Jarduera ez-laborala (lan bila)	19.586	17.525	10.651	18.067	65.830
Soilik langabeak	10.867	4.874	4.762	17.129	37.633
Jarduera ez-laborala (etxeko l., ikas., sold.)	265.572	55.413	112.938	146.750	580.673
Soilik erretiratuak eta beste batzuk	113.300	27.900	46.692	20.651	208.543
Guztira	719.413	287.535	340.666	505.375	1.852.992
errenkada %	1-gazt. soilik	2. gazt. + beste batzuk	3. gazt. + eu	4. gazt. + eu + beste batzuk	Guztira
Landunak	32,3%	18,9%	17,2%	31,5%	100%
Jarduera ez-laborala (lan bila)	29,8%	26,6%	16,2%	27,4%	100%
Soilik langabeak	28,9%	13,0%	12,7%	45,5%	100%
Jarduera ez-laborala (etxeko l., ikas., sold.)	45,7%	9,5%	19,4%	25,3%	100%
Soilik erretiratuak eta beste batzuk	54,3%	13,4%	22,4%	9,9%	100%
Guztira	38,8%	15,5%	18,4%	27,3%	100%
zutabe %	1-gazt. soilik	2. gazt. + beste batzuk	3. gazt. + eu	4. gazt. + eu + beste batzuk	Guztira
Landunak	43,1%	63,2%	48,6%	59,9%	51,8%
Jarduera ez-laborala (lan bila)	2,7%	6,1%	3,1%	3,6%	3,6%
Soilik langabeak	1,5%	1,7%	1,4%	3,4%	2,0%
Jarduera ez-laborala (etxeko l., ikas., sold.)	36,9%	19,3%	33,2%	29,0%	31,3%
Soilik erretiratuak eta beste batzuk	15,7%	9,7%	13,7%	4,1%	11,3%
Guztira	100%	100%	100%	100%	100%

2-1 emaitza.Hizkuntzak jakitea (iturria: BBI), BJA segmentazioarekin erkatuta (iturria: BJA). Guztizkoak biztanleria osorako; eta errenkaden eta zutabeen ehunekoak.

Aldagaia: hizkuntzak jakitea

35-44 urte bitarteko adina		Iturria: BBI			
Iturria: BJA		1-gazt. soilik	2. gazt. + beste	3. gazt. + eu	4. gazt. + eu + beste
			batzuk	eu	batzuk
					Guztira
Landunak		76.100	59.203	50.488	96.088
Jarduera ez-laborala (lan bila)		7.081	4.679	4.096	3.464
Soilik langabeak		2.536	415	321	2.311
Jarduera ez-laborala (etxeko l., ikas., sold.)		16.426	7.238	7.056	13.743
Soilik erretiratuak eta beste batzuk		1.899	1.051	155	993
Guztira		104.042	72.587	62.116	116.599
errenkada %		1-gazt. soilik	2. gazt. + beste batzuk	3. gazt. + eu	4. gazt. + eu + beste batzuk
					Guztira
Landunak		27,0%	21,0%	17,9%	34,1%
Jarduera ez-laborala (lan bila)		36,7%	24,2%	21,2%	17,9%
Soilik langabeak		45,4%	7,4%	5,8%	41,4%
Jarduera ez-laborala (etxeko l., ikas., sold.)		36,9%	16,3%	15,9%	30,9%
Soilik erretiratuak eta beste batzuk		46,3%	25,6%	3,8%	24,2%
Guztira		29,3%	20,4%	17,5%	32,8%
zutabe %		1-gazt. soilik	2. gazt. + beste batzuk	3. gazt. + eu	4. gazt. + eu + beste batzuk
					Guztira
Landunak		73,1%	81,6%	81,3%	82,4%
Jarduera ez-laborala (lan bila)		6,8%	6,4%	6,6%	3,0%
Soilik langabeak		2,4%	0,6%	0,5%	2,0%
Jarduera ez-laborala (etxeko l., ikas., sold.)		15,8%	10,0%	11,4%	11,8%
Soilik erretiratuak eta beste batzuk		1,8%	1,4%	0,3%	0,9%
Guztira		100%	100%	100%	100%

2-2 emaitza. Hizkuntzak jakitea (iturria: BBI), BJA segmentazioarekin erkatuta (iturria: BJA). Hautatu den estratua: 35 eta 44 urte bitartekoak (biak barne). Errenkaden eta zutabeen ehunekoak eta guztizkoak.

Aldagaia: aisialdia (orduak egunean)

Iturria: BJA	Iturria: BBI			
	<2 ordu	2-4 ordu	+4 ordu	Guztira
Landunak	171.098	551.052	238.162	960.312
Jarduera ez-laborala (lan bila)	8.430	28.436	28.963	65.830
Soilik langabeak	1.161	17.341	19.131	37.633
Jarduera ez-laborala (etxeko l., ikas., sold.)	48.494	247.203	284.977	580.673
Soilik erretiratuak eta beste batzuk	11.415	48.415	148.713	208.543
Guztira	240.598	892.447	719.946	1.852.991
errenkada %				
	<2 ordu	2-4 ordu	+4 ordu	Guztira
Landunak	17,8%	57,4%	24,8%	100%
Jarduera ez-laborala (lan bila)	12,8%	43,2%	44,0%	100%
Soilik langabeak	3,1%	46,1%	50,8%	100%
Jarduera ez-laborala (etxeko l., ikas., sold.)	8,4%	42,6%	49,1%	100%
Soilik erretiratuak eta beste batzuk	5,5%	23,2%	71,3%	100%
Guztira	13,0%	48,2%	38,9%	100%
zutabe %				
	<2 ordu	2-4 ordu	+4 ordu	Guztira
Landunak	71,1%	61,7%	33,1%	51,8%
Jarduera ez-laborala (lan bila)	3,5%	3,2%	4,0%	3,6%
Soilik langabeak	0,5%	1,9%	2,7%	2,0%
Jarduera ez-laborala (etxeko l., ikas., sold.)	20,2%	27,7%	39,6%	31,3%
Soilik erretiratuak eta beste batzuk	4,7%	5,4%	20,7%	11,3%
Guztira	100%	100%	100%	100%

3-1 emaitza. Ordu libreak egunean (iturria: BBI), BJA segmentazioarekin erkatuta (iturria: BJA).

Guztizkoak biztanleria osorako; eta errenkaden eta zutabeen ehunekoak.

Aldagaia: aisialdia (orduak egunean)

(adin-taldea, 35-44 urte)		Iturria: BBI			
Iturria: BJA		<2 ordu	2-4 ordu	+4 ordu	Guztira
Landunak		35.916	67.250	14.515	117.680
Jarduera ez-laborala (lan bila)		2.246	4.291	7.230	13.767
Soilik langabeak		0	0	159	159
Jarduera ez-laborala (etxeko l., ikas., sold.)		4.169	23.519	7.230	34.917
Soilik erretiratuak eta beste batzuk		0	0	1.027	1.027
Guztira		42.330	95.060	30.160	167.550
errenkada %		<2 ordu	2-4 ordu	+4 ordu	Guztira
Landunak		30,5%	57,1%	12,3%	100%
Jarduera ez-laborala (lan bila)		16,3%	31,2%	52,5%	100%
Soilik langabeak		0,0%	0,0%	100,0%	100%
Jarduera ez-laborala (etxeko l., ikas., sold.)		11,9%	67,4%	20,7%	100%
Soilik erretiratuak eta beste batzuk		0,0%	0,0%	100,0%	100%
Guztira		25,3%	56,7%	18,0%	100%
zutabe %		<2 ordu	2-4 ordu	+4 ordu	Guztira
Landunak		84,8%	70,7%	48,1%	70,2%
Jarduera ez-laborala (lan bila)		5,3%	4,5%	24,0%	8,2%
Soilik langabeak		0,0%	0,0%	0,5%	0,1%
Jarduera ez-laborala (etxeko l., ikas., sold.)		9,8%	24,7%	24,0%	20,8%
Soilik erretiratuak eta beste batzuk		0,0%	0,0%	3,4%	0,6%
Guztira		100%	100%	100%	100%

3-2 emaitza. Aisialdia (iturria: BBI), BJA segmentazioarekin erkatuta (iturria: BJA). Hautatu den estratua: 35 eta 44 urte bitarteko emakumeak (biak barne). Errenkaden eta zutabeen ehunekoak eta guztizkoak.

Aldagaia: egoera ekonomiko objektiboa

Iturria: BJA	Iturria: BBI			
	Txarra	Normala	Ona	Guztira
Landunak	64.393	420.064	475.855	960.312
Jarduera ez-laborala (lan bila)	19.567	31.043	15.219	65.830
Soilik langabeak	6.935	18.128	12.570	37.633
Jarduera ez-laborala (etxeko l., ikas., sold.)	81.399	319.399	179.875	580.673
Soilik erretiratuak eta beste batzuk	40.017	112.727	55.799	208.543
Guztira	212.311	901.361	739.318	1.852.991
errenkada %				
	Txarra	Normala	Ona	Guztira
Landunak	6,7%	43,7%	49,6%	100%
Jarduera ez-laborala (lan bila)	29,7%	47,2%	23,1%	100%
Soilik langabeak	18,4%	48,2%	33,4%	100%
Jarduera ez-laborala (etxeko l., ikas., sold.)	14,0%	55,0%	31,0%	100%
Soilik erretiratuak eta beste batzuk	19,2%	54,1%	26,8%	100%
Guztira	11,5%	48,6%	39,9%	100%
zutabe %				
	Txarra	Normala	Ona	Guztira
Landunak	30,3%	46,6%	64,4%	51,8%
Jarduera ez-laborala (lan bila)	9,2%	3,4%	2,1%	3,6%
Soilik langabeak	3,3%	2,0%	1,7%	2,0%
Jarduera ez-laborala (etxeko l., ikas., sold.)	38,3%	35,4%	24,3%	31,3%
Soilik erretiratuak eta beste batzuk	18,8%	12,5%	7,5%	11,3%
Guztira	100%	100%	100%	100%

4-1 emaitza. Egoera ekonomiko objektiboa (iturria: BBI), BJA segmentazioarekin erkatuta (iturria: BJA).
Guztizkoak biztanleria osorako; eta errenkaden eta zutabeen ehunekoak.

Aldagaia: egoera ekonomiko objektiboa

35-34 urte bitarteko adina		Iturria: BBI			
Iturria: BJA		Txarra	Normala	Ona	Guztira
Landunak		21.066	108.143	128.270	257.479
Jarduera ez-laborala (lan bila)		6.179	10.297	4.341	20.818
Soilik langabeak		1.640	8.015	4.400	14.054
Jarduera ez-laborala (etxeko l., ikas., sold.)		2.194	9.478	12.087	23.760
Soilik erretiratuak eta beste batzuk		1.251	6.852	3.367	11.470
Guztira		32.330	142.785	152.465	327.581
errenkada %		Txarra	Normala	Ona	Guztira
Landunak		8,2%	42,0%	49,8%	100%
Jarduera ez-laborala (lan bila)		29,7%	49,5%	20,9%	100%
Soilik langabeak		11,7%	57,0%	31,3%	100%
Jarduera ez-laborala (etxeko l., ikas., sold.)		9,2%	39,9%	50,9%	100%
Soilik erretiratuak eta beste batzuk		10,9%	59,7%	29,4%	100%
Guztira		9,9%	43,6%	46,5%	100%
zutabe %		Txarra	Normala	Ona	Guztira
Landunak		65,2%	75,7%	84,1%	78,6%
Jarduera ez-laborala (lan bila)		19,1%	7,2%	2,8%	6,4%
Soilik langabeak		5,1%	5,6%	2,9%	4,3%
Jarduera ez-laborala (etxeko l., ikas., sold.)		6,8%	6,6%	7,9%	7,3%
Soilik erretiratuak eta beste batzuk		3,9%	4,8%	2,2%	3,5%
Guztira		100%	100%	100%	100%

4-2 emaitza. Egoera ekonomiko objektiboa (iturria: BBI), BJA segmentazioarekin erkatuta (iturria: BJA). Hautatu den estratua: 35 eta 44 urte bitartekoak (biak barne). Errenkaden eta zutabeen ehunekoak eta guztizkoak.

Aldagaia: ibilgailu propioak

Iturria: BJA	Iturria: BBI			
	Bat ere ez	Bat	Bi edo gehiago	Guztira
Landunak	114.756	558.465	287.091	960.312
Jarduera ez-laborala (lan bila)	19.094	34.008	12.728	65.830
Soilik langabeak	5.770	19.520	12.343	37.633
Jarduera ez-laborala (etxeko l., ikas., sold.)	211.396	274.670	94.608	580.673
Soilik erretiratuak eta beste batzuk	71.811	110.357	26.375	208.543
Guztira	422.827	997.020	433.145	1.852.991
errenkada %				
	Bat ere ez	Bat	Bi edo gehiago	Guztira
Landunak	11,9%	58,2%	29,9%	100%
Jarduera ez-laborala (lan bila)	29,0%	51,7%	19,3%	100%
Soilik langabeak	15,3%	51,9%	32,8%	100%
Jarduera ez-laborala (etxeko l., ikas., sold.)	36,4%	47,3%	16,3%	100%
Soilik erretiratuak eta beste batzuk	34,4%	52,9%	12,6%	100%
Guztira	22,8%	53,8%	23,4%	100%
zutabe %				
	Bat ere ez	Bat	Bi edo gehiago	Guztira
Landunak	27,1%	56,0%	66,3%	51,8%
Jarduera ez-laborala (lan bila)	4,5%	3,4%	2,9%	3,6%
Soilik langabeak	1,4%	2,0%	2,8%	2,0%
Jarduera ez-laborala (etxeko l., ikas., sold.)	50,0%	27,5%	21,8%	31,3%
Soilik erretiratuak eta beste batzuk	17,0%	11,1%	6,1%	11,3%
Guztira	100%	100%	100%	100%

5-1 emaitza. Ibilgailu propioen kopurua (iturria: BBI), BJA segmentazioarekin erkatuta (iturria: BJA).
Guztizkoak biztanleria osorako; eta errenkaden eta zutabeen ehunekoak.

Aldagaia: ibilgailu propioak

Estratua: 55-64 urte		Iturria: BBI		
Iturria: BJA		Bat ere ez	Bat	Bi edo gehiago
				Guztira
Landunak		13.009	69.510	31.889
Jarduera ez-laborala (lan bila)		970	2.345	2.036
Soilik langabeak		276	2.582	449
Jarduera ez-laborala (etxeko l., ikas., sold.)		23.395	68.120	20.161
Soilik erretiratuak eta beste batzuk		9.495	20.390	8.382
Guztira		47.144	162.947	62.917
errenkada %		Bat ere ez	Bat	Bi edo gehiago
				Guztira
Landunak		11,4%	60,8%	27,9%
Jarduera ez-laborala (lan bila)		18,1%	43,8%	38,0%
Soilik langabeak		8,3%	78,1%	13,6%
Jarduera ez-laborala (etxeko l., ikas., sold.)		20,9%	61,0%	18,1%
Soilik erretiratuak eta beste batzuk		24,8%	53,3%	21,9%
Guztira		17,3%	59,7%	23,0%
zutabe %		Bat ere ez	Bat	Bi edo gehiago
				Guztira
Landunak		27,6%	42,7%	50,7%
Jarduera ez-laborala (lan bila)		2,1%	1,4%	3,2%
Soilik langabeak		0,6%	1,6%	0,7%
Jarduera ez-laborala (etxeko l., ikas., sold.)		49,6%	41,8%	32,0%
Soilik erretiratuak eta beste batzuk		20,1%	12,5%	13,3%
Guztira		100%	100%	100%

5-2 emaitza. Ibilgailu propioak (iturria: BBI), BJA segmentazioarekin erkatuta (iturria: BJA). Hautatu den estratua: 55 eta 64 urte bitartekoak (biak barne). Errenkaden eta zutabeen ehunekoak eta guztizkoak.

Aldagaia: etxeko ekipamendua

Iturria: BJA	Iturria: BBI		
	Eskasa	Nahikoa	Guztira
Landunak	33.187	927.125	960.312
Jarduera ez-laborala (lan bila)	0	65.830	65.830
Soilik langabeak	446	37.186	37.633
Jarduera ez-laborala (etxeko l., ikas., sold.)	92.649	488.024	580.673
Soilik erretiratuak eta beste batzuk	43.476	165.067	208.543
Guztira	169.758	1.683.232	1.852.991
errenkada %			
	Eskasa	Nahikoa	Guztira
Landunak	3,5%	96,5%	100%
Jarduera ez-laborala (lan bila)	0,0%	100,0%	100%
Soilik langabeak	1,2%	98,8%	100%
Jarduera ez-laborala (etxeko l., ikas., sold.)	16,0%	84,0%	100%
Soilik erretiratuak eta beste batzuk	20,8%	79,2%	100%
Guztira	9,2%	90,8%	100%
zutabe %			
	Eskasa	Nahikoa	Guztira
Landunak	19,5%	55,1%	51,8%
Jarduera ez-laborala (lan bila)	0,0%	3,9%	3,6%
Soilik langabeak	0,3%	2,2%	2,0%
Jarduera ez-laborala (etxeko l., ikas., sold.)	54,6%	29,0%	31,3%
Soilik erretiratuak eta beste batzuk	25,6%	9,8%	11,3%
Guztira	100%	100%	100%

6-1 emaitza. Etxeko ekipamendua (iturria: BBI), BJA segmentazioarekin erkatuta (iturria: BJA). Guztizkoak biztanleria osorako; eta errenkaden eta zutabeen ehunekoak.

Aldagaia: etxeko ekipamendua

65 urtetik gorakoak	Iturria: BBI		
Iturria: BJA	Eskasa	Nahikoa	Guztira
Landunak	729	1.877	2.606
Jarduera ez-laborala (lan bila)	0	0	0
Soilik langabeak	0	0	0
Jarduera ez-laborala (etxeko l., ikas., sold.)	78.202	174.850	253.051
Soilik erretiratuak eta beste batzuk	39.083	98.975	138.058
Guztira	118.014	275.701	393.715
errenkada %	Eskasa	Nahikoa	Guztira
Landunak	28,0%	72,0%	100%
Jarduera ez-laborala (lan bila)	.	.	.
Soilik langabeak	.	.	.
Jarduera ez-laborala (etxeko l., ikas., sold.)	30,9%	69,1%	100%
Soilik erretiratuak eta beste batzuk	28,3%	71,7%	100%
Guztira	30,0%	70,0%	100%
zutabe %	Eskasa	Nahikoa	Guztira
Landunak	0,6%	0,7%	0,7%
Jarduera ez-laborala (lan bila)	0,0%	0,0%	0,0%
Soilik langabeak	0,0%	0,0%	0,0%
Jarduera ez-laborala (etxeko l., ikas., sold.)	66,3%	63,4%	64,3%
Soilik erretiratuak eta beste batzuk	33,1%	35,9%	35,1%
Guztira	100%	100%	100%

6-2 emaitza. Etxeko ekipamendua (iturria: BBI), BJA segmentazioarekin erkatuta (iturria: BJA). Hautatu den estratua: 65 urtetik gorakoak (Errenkaden eta zutabeen ehunekoak eta guztizkoak.

R pakete propioa garatzea

BBI eta BJA inkestak lotu bitartean, funtzio batzuk sortu genituen, prozesuaren faseetako kalkuluak bizkortze aldera: aldagaien hautaketa, estratuen arabera egozte eta emaitzen baliozkotzea. Funtzio horiek, frogatutako eta erkatutako paketeetan oinarrituta (ikus hirugarren kapitulua: Softwarea), R pakete propio batean sartu ziren, eta pakete horri `micromatch` deitu diogu. Asmoa funtzio horiek, koaderno tekniko honekin batera, banatzea da.

Berezko R paketeek (hots, proiektu bati lotuak) zenbait abantaila dituzte (Chambers, J., 2008):

- R funtzioak proiekturako egituratzen dituzte, osotasun integratu batean egongo balira bezala; horri esker, eraginkoragoak eta fidagarriagoak dira kalkuluak.
- Bestalde, programatzailearentzat ere oso lagungarriak diren dokumentu batzuk erraz sortzen dituzte.
- Kalkuluak orokortzeko bideak ematen dituzte, beste arazo batzuetan edo antzeko egoeretan erabilgarriak izan daitezen.
- Kodea erraz bana dakieke beste erabiltzaile batzuei.

BBI-BJA loturari buruzko proiektuan jada abantaila horiek baliatu genituen.

Proiektu honen ondorengo fase batean ikusi genuen `micromatch` paketea orokortu eta inkesta independenteen gaineko edozein loturatan era generikoan ezartzeko aukera zegoela. Urrats hori bete ahal izateko, klase- eta metodo-egitura bat diseinatu genuen helburuetara bideratutako programazioaren gaineko S4 sisteman (Chambers, J., 2008). **Klaseek** kontzeptuak *objektu* gisa kapsulatzeko bide ematen dute: esaterako, «lotu beharreko inkesta» (`micromatch` paketeko `filetomatch` klasean). Ondoren, objektu horiei eragiten dieten metodoak diseinatu genituen: esaterako, «alderatu aldagai guztiak, lotu beharreko beste inkesta bat aintzat hartuta».

Beste pakete batzuek ez bezala, `micromatch` paketeak ez du inkestak lotzeko metodo propiorik ematen. Horren orde, frogatuta eta erkatuta dauden paketeak oinarri hartzen ditu, eta soluzio generikoa aplikatzen du. Soluzio horrek lotura bateko faseko guztiei heltzen die (eta bizkortu ere egiten ditu).

Beraz, `micromatch` paketeak *ingurune* bat sortzen du, eta bertan inkestak lotzeko metodo-mota guztiak erraz erabil eta froga ditzake erabiltzaileak. Laburbilduz, honako hauek dira pakete honen helburuak:

- Inkesten arteko lotura batek behar dituen kalkuluak errazteko ingurunea eskaintzea
- Konputazio-sistema eraginkorra eta sendoa eskaintzea: bertan kalkulu guztiak eraginkortasunez elkarri lotuta daude, klase-sistemari esker.
- Inkesta-metodologia definitzea, gaur egun pakete bereizietan dauden tekniken multzoa pakete bakarrean sartuta.
- Beka jaso bitartean garatutako lana zabaltzea: ingurune bat eskaintzen dugu, non erabiltzaileak BBI-BJA kalkuluak berriro sor ditzakeen.

Pakete hori Konpostelako Santiagon antolatu ziren *VI Jornadas de Usuarios de R* deituriko jardunaldietan aurkeztu zen, zehazki, 2014ko urrian, eta , besteak beste, Eustat-eko web-aren bitartez dago eskuragarri.

Ondorioak

Inkesten arteko loturak biztanleria-kopuru berari buruzkoak diren inkesta independenteen bidez jasotako informazioa era eraginkorragoa ustiatzeko bide ematen digu, estatistika eta adierazle integratuak eskuratuta.

Koaderno tekniko honetan adibide praktiko bat garatu dugu, hau da, Eustatek eginiko bi inkesta independente –Bizi-baldintzen inkesta eta Biztanleria jardueraren arabera sailkatzeko inkesta– erlazionatzeko lotura sortu dugu. Horri esker, lotura bat sortzeko prozesu osoko elementu batzuen garrantziaz ohartu gara. Hona hemen elementuok:

- **Aldagai komun** egokiak hautatuko baditugu, galdera-sorten analisi sakona egin beharko dugu, alderagarria den informazioa zehaztu ahal izateko (datu-fitxategietan behatutako banaketa enpirikoei eta definizioari dagokienez).
- **Estratu-aldagaiak** (adinaren eta sexuaren arabera BBI-BJA adibidean) baliagarriak dira, berariazko aldagaiei oso lotuta baitaude. Berariazko aldagaiak hauexek dira gure adibidean: batetik, BBI inkestatik hartutako bizi-baldintzak, hots, aisialdia, gizarte-harremanak...; eta, bestetik, lan-merkatua, BJA inkestatik hartuta.
- **Emaitzak balioztatu** behar dira; horretarako, algoritmoak bazter-banaketak sortu beharko ditu, eta dagoeneko behatu direnekin erkatu ahal izango dira. Edonola ere, kontuan hartu beharko da ziurgabetasun-iturri gehiago inplikatura egon daitezkeela.

Era orokorragoan, inkestak lotzeko metodologia bat hartzea onuragarria da: inkestak ez dira tresna independentetzat hartzen, baizik eta osotasun integratutzat. Horren harira, A. Leulescu-k eta M. Agafitei-k (2013) zenbait gomendio eman dituzte.

- Ahal den neurrian, galdera-sortak **estandarizatu** behar dira: galderek alderagarriak izan behar dute, haien arteko koherentzia bermatu ahal izateko.
- Aukera izanez gero, **inkesta guztiek partekatzen duten modulu txiki bat** sar daiteke, «berariazko» alderdi batzuk, oinarrizkoak ere badirenak, bertan biltzeko: esaterako, diru-sarrerak, nork bere osasunari buruz duen pertzepzioa. Inkestak lotzerakoan, aldagai horiek funtzio gakoa dute loturaren emaitzen kalitatea hobetzeko (hau da, ziurgabetasuna murrizteko).

BBI-BJA loturari soilik dagokionez, komeni litzateke etorkizunean diru-sarrerei edo ikasketa-mailei buruzko aldagai harmonizatuak izatea. Aldagai horiek ziurgabetasuna murrizten eta, ondorioz, loturaren emaitzak hobetzen lagunduko dute.

Aipatutako tekniken **erabilera** zehazteko, kontuan hartu behar da software libreko R ingurunea gero eta tresna gehiago ematen ari dela; hortaz, litekeena da etorkizunean pakete guztien arteko koordinazioa handiagoa izatea. Proiektu honetan, `micromatch` paketearen garapenari esker, jada urrats batzuk eman ditugu norabide horretan.

Bibliografia

Erreferentzia orokorrak

Agresti (2014). Categorical data analysis. John Wiley & Sons.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. Journal of statistical software, 45(3).

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). Statistical matching: Theory and practice. John Wiley & Sons.

D'Orazio, M., Di Zio, M., & Scanu, M. (2010). Old and new approaches in statistical matching when samples are drawn with complex survey designs. Proceedings of the 45th "Riunione Scientifica della Societa'Italiana di Statistica", Padova, 16-18.

Leulescu, A. & Agafitei, M. (2013). Statistical matching: a model based approach for data integration. Eurostat Methodologies and working papers.

Rässler, S. (2002). Statistical matching. Springer.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business & Economic Statistics, 4(1), 87-94.

Europako proiektuak

Data Integration: <http://www.cros-portal.eu/content/data-integration-finished>

ISAD Integration of Survey and Administrative Data. <http://www.cros-portal.eu/content/isad-finished>

Report WP2 ESS-net, Statistical Methodology Project on Integration of Surveys and Administrative Data. Recommendations on the use of methodologies for the integration of surveys and administrative data.

Eustat-en barneko dokumentuak

Eustat, Ficha metodológica de la encuesta Población en Relación con la Actividad (PRA) http://es.eustat.es/document/poblact_c.html

Eustat, Ficha metodológica de la Encuesta de Condiciones de Vida (ECV), http://es.eustat.es/document/ecvida_c.html#axzz37QhmZ17l

Softwarea

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

Chambers, J. (2008). Software for data analysis: programming with R. Springer.

D'Orazio, M. (2013). StatMatch: Statistical Matching (*data fusion* deiturikoa). <http://CRAN.R-project.org/package=StatMatch>

Lumley, T. (2012) survey: analysis of complex survey samples. <http://CRAN.R-project.org/package=survey>

Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL-a: <http://www.jstatsoft.org/v45/i03/>

Harrell Jr, F. E. (2014). Hmisc: Harrell Miscellaneous. R package version 3.14-3. <http://CRAN.R-project.org/package=Hmisc>

Honaker, J., King, G. & Blackwell, M. (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software, 45(7), 1-47. URL <http://www.jstatsoft.org/v45/i07/>.

Meinfielder, F. (2011). BaBooN: Bayesian Bootstrap Predictive Mean Matching – Multiple and single imputation for discrete data. <http://CRAN.R-project.org/package=BaBooN>

Eranskinak

LOTURARAKO BERARIAZKO ALDAGAIK				
BJA aldagai bakarra (PRA, gaztelaniazko izenaren arabera)				
Aldagaia	Mikrodatuen aldagaia	--	Izen laburra	Agregatu diren kategoriak ^a
Jarduerarekiko erlazio hedatua	PV1_PRA2		PRA22	1-lan-jarduera (landunak) ^a ; 2- jarduera ez-laborala (etxeko lanen arduradunak, ikasleak, soldaduskan daudenak); 3-jarduera ez-laboral (lan-bila daudenak); 4-soilik langabeak; 5-soilik erretiratuak eta beste batzuk.

BBI inkestaren aldagaiak (laginketa); ECV, gaztelaniazko izenaren arabera				
Aldagaia	Mikrodatuen aldagaia	Galder a-sorta ^c	Izen laburra	Kategoriak
Osasun-arazoak	CVI_TRASPI	ind	SAL	1-arazoren bat; 2-arazorik ez
Hizkuntzak jakitea	CVI_COIDI	ind	IDM	gaztelania eta euskara; 4-gaztelania, euskara eta beste batzuk
Aisialdia	CVI_TLIBRR	ind	LIB	1-bi ordutik behera; 2 ordukoa edo 2-4 ordu artekoa; hiru ordukoa-4 ordutik gorakoa
Egoera ekonomiko objektiboa	CVI_SITEC2	fam	ECO	1-txarra; 2-normala; 3-ona
Motordun ibilgailuak edukitzea ^b	CV1_NMOTOR; CV1_NCOCHR;	fam	VEHICL	1-bat ere ez; 2-bat; 3-bi edo gehiago
Etxeko ekipamendua	CV1_EQUIP6	fam	EQP	1-ekiporik ez; 2-ekipo batzuk; 3-ekipo asko

^a Landunen barruan, BJA inkestak azpikategoriak bereizten ditu, landun egoera-maila zein den. Azterlan honetan kategoria bakarrean gehitu dira: 'Landunak'

^b Aldagaia sortu da, honako hauek abiapuntu hartuta: CV1_NMOTOR '50 cc-tik gorako motozikleta-kopurua'; CV1_NCOCHR 'Automobil-kopurua'; CV1_NFURGR 'Furgoneta-kopurua'; CV1_OTRVEHICR 'Beste ibilgailu batzuk'

^c Jatorrizko galdera-sorta: banakakoa (ind) edo familiakoa (fam).

A fitxa: Berariazko aldagaiak, Eustat-ek eginiko BBI eta BJA inkesta independenteen arteko loturarako

	BJA-2009 4H		BBI-2009			
Aldagaia	Izena ^a	Galera-zenbakia galdera-sortan ^b	Izena ^a	Jatorrizko fitxategia ^c	Galera-zenbakia galdera-sortan ^b	Maila agregatuak ^c
Aldagai soziodemografikoak						
Sexua ('sexo')	PV1_SEXO	p12	ind: CV1_SEXOI	ind	--	G: gizonak; E: emakumeak
Adina ('edad')	PV1_EDAD	p10	ind: CV1_EDADIR	ind		01-"<=15 urte"; 02-"16-24 urte"; 03-"25-34 urte"; 04-"35-44 urte"; 05-"45-54 urte"; 06-"55-64 urte"; 07-">=65 urte"
Familiaren tamaina ('tamaño familiar')	TAMAÑO_FAM	--	fam: CV1_TFAMR	fam	--	1,2, 3+ kide
Ikasketa-maila						
Ikasketa arautuak egiten ari da S/N ^e	PV1_ENRE (!D) ^d	p43	CV1_SITES (B)	ind	pl2	1= ikasten du; 0=ez du ikasten
Urrutiko ikasketak egiten ari da S/N	PV1_ENRE (C)	p43	CVI_SISTE (F)	ind	pl4	1= bai; 0= ez
Analfabeto S/N	PV1_LEES	p34	CVI_ANALF	ind	pl15	1=bai; 0=ez
D. Lan-egoerari lotutako aldagaiak						
Jarduerarekiko erlazioa – OIT	PV1_PRA1	--	CV1_REL1	ind	--	Landunak; langabeak; inaktiboak
Lan bila S/N	PV1_BUSQ	p140	CVI_BUSQ	ind	pT23	1= bai; 0= ez
Astean egindako lan-orduak (land.)	PV1_HTRA ^g	p107	CVI_HOTAT	ind	pT22	Zenbakizkoa
Etxeko lanen arduraduna	PV1_SILH	p55	CV1_TDOME1	ind	Adierazle sintetikoa ^f	Egiten ditu; ez ditu egiten

^a Aldagaiaren izena mikrodaturen fitxategian.

^b Galdera-zenbakia fitxategian

^c Jatorrizko fitxategia, BBI inkestarako: jatorrizko datuak dituen fitxategia adierazten du (banakakoa: ind; familiakoa: fam)

^d '!' ikurra: adierazitakoa izan ezik, gainerako kategoria guztiak hartu dira.

^e Aurrerantzean, 'Ikasle S/N'.

^f pT27 galderaren 4 itemetatik eratorritako adierazlea: COMPR2-elikagaiak erostea ; COMID2-janaria prestatzea; FREG2-harrikoa egitea; ROPA2-arropa prestatzea; LIMPC2-etxea txukundu eta garbitzea.

^g Segmentua: OIT=landunak

B fitxa: BBI eta BJA inkestek partekatzen dituzten aldagaien meta-analisia

Aldagai komunak		Koherentzia	Iragarpen-balioa						
Aldagaia	Izen laburra	Hell. dist.	Cramer-en V (globala)						
			BJA: "PRA"	BBI: "SAL"	BBI: "IDM"	BBI: "LIB"	BBI: "ECO"	BBI: "VEH"	BBI: "EQP"
Adina	ED	0,002	0,372	0,304	0,267	0,230	0,152	0,289	0,396
Sexua	S	7,68E-06	0,409	0,014	0,027	0,075	0,051	0,123	0,054
Familiaren tamaina ('tamaño familiar')	TF	0,001	0,192	0,169	0,099	0,167	0,194	0,285	0,275
Ikasle S/N	EST	0,009	0,329	0,097	0,333	0,138	0,028	0,084	0,089
Landun S/N	OCP	0,002	1,000	0,275	0,291	0,402	0,345	0,373	0,285
Langabe S/N	PAR	0,012	0,998	0,021	0,046	0,105	0,142	0,029	0,073
Inaktibo S/N	INA	0,007	0,999	0,288	0,309	0,380	0,285	0,366	0,322
Lan bila S/N	BUSQ	0,025	0,850	0,067	0,118	0,037	0,073	0,035	0,087
Etxeko lanak	DOM	0,054	0,548	0,054	0,061	0,105	0,032	0,096	0,002

C1 taula. Neurketa globalak

Estratua	Antzeman diren aldagai komunak	Laginaren tamaina		Koherentzia	Iragarpen-balioa
(Izen laburra)	Estratuaren arabera hautatutako aldagaiak	BBI (hartzailea)	BJA (emaila)	Hell. distantziak (aldagaien arabera)	Cramer-en V Mendeko aldagaia: "PRA" (BJA, euskaraz)
G.15-24	EST, BUSQ	157	512	EST: 0,048 - BUSQ: 0,052	EST: 0,887 - BUSQ: 0,907
E.15-24	EST, BUSQ	166	491	EST: 0,09 - BUSQ: 0,003	EST: 0,835 - BUSQ: 0,848
G.25-34	BUSQ eta DOM	359	746	BUSQ: 0,002 - DOM: 0,064	BUSQ: 0,895 - DOM: 0,389
E.25-34	PAR, DOM	372	740	PAR: 0,004 - DOM: 0,049	PAR: 1,000 - DOM: 0,378
G.35-44	PAR, TF2	426	920	PAR: 0,011 - TF2: 0,047	PAR: 1,000 - TF2: 0,117
E.35-44	PAR, OCP	386	957	PAR: 0,019 - OCP: 0,051	PAR: 1,000 - OCP: 1,000
G.45-54	PAR, BUSQ	377	966	PAR: 0,004 - BUSQ: 0,034	PAR: 1,000 - BUSQ: 0,870
E.45-54	PAR, BUSQ	403	1034	PAR: 0,032 - BUSQ: 0,004	PAR: 1,000 - BUSQ: 0,730
G.55-64	INA, BUSQ	349	843	INA: 0,033 - BUSQ: 0,023	INA: 1,000 - BUSQ: 0,848
E.55-64	INA	391	887	INA: 0,027	INA: 1,000
G.+65	TF2	540	1158	TF2: 0,018	TF2: 0,295
E.+65	TF2	823	1611	TF2: 0,030	TF2: 0,111

C2 taula. Adinaren eta sexuaren araberako estratuen neurketak, hautatutako aldagaietarako

C fitxa: Aldagaiak estratuen arabera hautatzea. Koherentziaren neurketa enpirikoak (Hellinger-en distantziak) eta iragarpen-balioa (Cramer-en V). Neurketa globalak (C1 taula) eta estratuen araberakoak (C2 taula)

Organismo Autónomo del



www.eustat.eus