ERREGISTROAK BATZEKO METODO AUTOMATIKOAK ETA EUSTATEN DUTEN ERABILERA

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

Aurkezpena

Argitalpen honetan, azken urteotan estatistika-sormenerako metodologia berrien aplikazioaren inguruan (erregistroak batzearen arloan) egindako lana aurkezten da.

Egindako lana 2005-2008 aldirako Euskal Estatistika Planeko eragiketetariko baten barruan sartzen da, eta eragiketa hori estatistika-metodoen I+G+b-ri buruzkoa da; lanaren helburua estatistika matematikoen metodologia berriak ikertzea eta estatistika--eragiketetan aplikatzea da. Ikerketa horiek bikaintasunaren kudeaketak eragin ditu eta horren helburua etengabeko hobekuntza da.

Estatistika-informazioaren eskari gero eta handiagoa, estatistika-bulegoen baliabide mugatuak eta inkestatuei gehiegizko erantzun-zama ekiditeko asmoa aintzat hartuta, beharrezkoa da zentsuetan eta inkestetan lorturiko informazioa eraginkortasunez erabiltzea, eta administrazioak emandako informazioa ere eraginkortasunez erabili behar da.

Erregistroak batzeko metodo automatikoen aplikazioa 90eko hamarkadan hasi zen EUSTATen. Lehenengo eta behin, metodo deterministak aplikatu ziren eta, 2002tik hona, probabilitate-metodoak garatu dira. Argitalpen honetan azken horiek aztertuko ditugu.

Argitalpen hau estatistikaren arloan interesa duten guztientzat erabilgarria izango delakoan nago.

Vitoria-Gasteiz, 2007ko martxoa
Josu Iradi Arrieta
Zuzendari Nagusia

ERREGISTROAK BATZEKO METODO AUTOMATIKOAK ETA EUSTATEN DUTEN ERABILERA

LABURPENA

Erregistroak batzea entitate bereko erregistroak identifikatzean datza, bi artxibo edo gehiagotan daudenean eta identifikatzaile desberdinak daudenean. EUSTATen[1] lan iraunkorra egin da erregistroak batzeko prozedurekin; hala eta guztiz ere, aspalditik hona teknika berriak ikertzen hasi gara, esate baterako, erregistroen probabilitate-batzea, eta hori koaderno tekniko honetan azalduta dago. Agiri honetan egiten den lehenengo gauza, izan ere, erregistroak batzearen kontzeptua labur-labur azaltzea da, eta gero, EUSTATek teknika berri hori ikertzen hastean zuen egoera adierazten da. Ondoren, erabilitako metodologiaren azalpena egiten da; metodologia hori, funtsean, Ivan P. Fellegi-ren eta Alan B. Sunter-en "*A theory for Record Linkage*" lanean oinarrituta dago. Aurrerago, EUSTATen probabilitate-metodo automatikoan oinarrituriko batzea egiteko garatu den programa informatikoa azaltzen da, lehen aipaturiko metodologiaren jarraibideen arabera. Programa ikusi ondoren, Erakundean arlo honetan garatu diren aplikazio batzuk aztertu eta, azkenik, ikerketaren ondorioak eskaintzen dira.

[1] EUSTATek eskerrak eman nahi ditu lehenengo Leire Legarretak eta gero Laura Oterok egindako ikerketa lanarengatik; lan hori EUSTATek sustaturiko ikerketa beken eta metodologia estatistiko-matematikoaren arloan garatu dute.

Aurkibidea

Kapitulua
1

Sarrera

Sarrera eta helburuak

Erregistroen batzea teknika algoritmikoen erabileran oinarritzen da, bi artxibo desberdineko erregistro bikoteak identifikatzeko, identifikatzaile desberdinak daudenean.

Orain arte, lan hau administrazioko langileek egin dute beti; langileok erregistro zerrendak eskuz ikuskatu eta informazio gehigarria lortu behar izaten zuten, halakorik ez zegoenean edo kontraesankorra zenean; gainera, batzeko erabakiak hartu behar izaten zituzten, aldez aurretik ezarritako arau jakinak kontuan hartuta. Oro har, zerrenda horiek izenaren edo helbidearen ezaugarriren baten arabera antolatzen ziren, prozesua errazteko. Izenen batek ezohiko aldaketa tipografikorik izanez gero, batzuetan ez zen horren loturarik aurkitzen. Fitxategiak oso luzeak zirenean, zerrendak zenbait orritan banatuta zeuden eta, zenbait kasutan, bikote batzuk galdu egiten ziren. Entrenamendu zehatza egon arren, hartzen ziren erabakiak beti ez ziren irmoak. Horren guztiaren ondorioz, teknika konputazionalak ikertu ziren, batze lan hori modu automatikoan egiteko.

Lan hau egiteko teknika desberdinak daude. Bi talde nagusitan sailka daitezke: metodo deterministak eta probabilitate-metodoak.

Erregistroen batze deterministaren helburua bi fitxategi edo gehiago batzearen aldagai bati edo gehiagori buruzko adostasun nahiz desadostasun zehatzak aurkitzea da. Esate baterako, bi fitxategiren NAN eremu komuna erabil daiteke. Hala eta guztiz ere, aldagai hau edozein erregistrotan erregistratzean gertaturiko kodifikazio akatsen ondorioz, benetako *match*etariko batzuk galdu egiten dira (*match*: bi erregistroren konparazio bikotea, pertsona beraren fitxategi desberdinetan).

Erregistroen probabilitate-batzeak aldagai kopuru handiagoaren informazioa erabiltzen du, eta aintzat hartzen du batzearen aldagaiei buruzko edozein adostasunek edo desadostasunek emandako informazioa. Esate baterako, gizarte segurantzaren zenbakiari buruzko adostasuna errazago izan daiteke *matcha*, sexuari buruzko adostasuna baino. Halaber, batze bateko aldagai baten balio arraroei buruzko adostasunak (adibidez, Galzarsoro abizena) adierazgarriagoak dira ohiko balioei (esaterako, García) buruzko adostasuna baino.

Argitalpen honen helburu nagusia Eustatek erregistroak batzeko teknikei (probabilitate-metodoen bidezko batzeak) buruz egindako lana aurkeztea da. Erakundean erregistroak batzeko metodo berriak ikertzeko erabakia hartu genuen, metodo deterministen bitartez lorturiko emaitzetariko batzuk hobetzeko, une horretara arte azken metodo hori erabiltzen baitzen. Fellegik eta Sunterrek garaturiko lanean oinarrituz, SAS hizkuntzan programaturiko batze-programa bat diseinatu da eta horrek bi fitxategi batzeko aukera ematen du, probabilitate-teknikak erabiliz.

Koaderno honetan, lehenengo eta behin, batze-programa garatzeko jarraitu den metodologia azaltzen da. Halaber, programa hori azaldu eta berori osatzen duen makro bakoitza ikusten da; makro bakoitzaren funtzioak eta batze-programaren barruan duten posizioa ere azaltzen dira. Ondoren, Erakundeak garatu dituen programa horren aplikazio batzuk zerrendatu eta lorturiko emaitzak adierazten dira. Azkenik, EUSTATek halako probabilitate-prozedurei buruz ateratako ondorioak azaltzen dira.

Proiektuaren azalpena

Helburu hauek lortzeko, lehenengo eta behin, erregistroak batzeko metodoei buruzko dokumentazioaren ikerketa zehatza egin zen, batik bat probabilitate-metodoei buruzkoa. Horren gaineko informazio nahikoa eduki ondoren, berori gauzatzea erabaki zen eta, horretarako, SAS hizkuntzako makroak garatu ziren, horiek pertsonen erregistrodun bi fitxategi modu automatikoan batzeko.

Programa garatu ondoren, zenbait fitxategirekin exekutatu zen, berorren fidagarritasuna ikusteko eta metodo honen bidez lorturiko emaitzak teknika deterministetan oinarrituriko prozeduren bitartez lorturiko emaitzekin konparatzeko (azken horiek koefiziente jakinak finkatzen zizkien aldagaiei).

Aurrekariak

Probabilitate-metodoak ikertzen hasi baino lehen, EUSTATen teknika deterministetan oinarrituriko batze-prozedurak garatzen ziren; prozedura horiek aldagaien izaera berezi eta bakoitzari koefiziente desberdina esleitzen zioten.

Batze-prozedura horren aplikazioetariko bat, bestalde, Biztanleen Udal Erroldari honako hau lotzeko egin zen: Biztanleriaren Estatistika Erregistroan dauden pertsonen gakoak (Oinarrizko Biztanleria Unitatea) eta Lurraldearen datu-baseko lurralde-gakoak (Kokalekuko Lurralde Unitatea).

Pertsonen Errolda Mugimenduetarako pertsonen gakoak baino ez dira lotzen. Prozesu hau zenbait pausotan edo modulutan egiten da, kasuan kasuko fitxategiaren arabera.

Honako modulu hauek azaltzen dira:

1) Batze zorrotzagoa.

Biztanleen Udal Erroldaren kasuan, pertsonen gakoak eta lurraldearen gakoak aurkitzeko ahalegina egiten da, hurrenez hurren, Biztanleriaren eta Lurraldearen Estatistika Erregistroan. Kokalekuko Lurralde Unitatearen gakoa, lehenengo eta behin, Oinarrizko Biztanleria Unitatearen gakoa erabiliz lortzeko ahalegina egiten da.

Biztanleriaren Udal Erroldaren erregistroak lotu egingo dira jaiotza-datari, jaiotza--lekuari, sexuari, posta-helbideari, izen-abizenei eta NANari buruzko datuak dituzten Biztanleriaren Estatistika Erregistroko taulen erregistroekin, honako irizpide hauek aplikatuz:

- Berdintasuna NANean. NANa osorik eta letra barik erabiltzen da. Maiztasunik handienekoak baztertu egiten dira, bikoiztu asko ez egoteko, horiek NAN akastuntzat hartzen baitira.

- Berdintasuna izen-abizenetan. Lehenengo eta behin, izen-abizenak normalizatu eta alfagakoak sortzeko funtzioa ematen zaie. Alfagako horiek hitzez hitzezko baliokideak dira eta aldagaion tratamendua errazten dute. Lotura egiteko, hiru alfagako erabiltzen dira aldi berean.

- Berdintasuna posta-helbidean, jaiotza-datan eta sexuan berdintasuna dagoela egiaztatu ondoren.

Pertsonen Errolda Mugimenduen taulako erregistro bakoitza Biztanleriaren Estatistika Erregistroarekin lotu behar da, Oinarrizko Biztanleria Unitatearen gakoaren bitartez. Biztanleriaren Udal Erroldarako erabilitako prozesu bera erabili behar da, posta-helbidearekiko konparazioak aintzat hartu barik, hori ez baitator Pertsonen Errolda Mugimenduen fitxategian.

Gero, koefizienteak kalkulatzen dira, eta horrela, aldagaiekiko konparazioak eginez, puntuazioak ematen zaizkie doitasunei, gero baliorik handienak dutenekin gelditzeko.

2) Hain zorrotza ez den batzea.

Baliozko lotura batzuk kanpo geratzeko arriskua zegoen, lehenengo batze-irizpideak zorrotzegiak izatearen ondorioz, eta beraz, baldintza batzuk malgutzeko erabakia hartu zen.

Lehenengo egokitzapenari dagokionez, lehenengo prozesuaren bidez batu gabeko erregistroetarako berdintasunak bilatu ziren, izena eta bi abizenak konbinatuz: izena eta lehenengo abizena, izena eta bigarren abizena, lehenengo eta bigarren abizenak. Prozesu honetan erabilitako koefizienteak aurreko modulukoak dira.

Egokitzapen horren ostean batu gabe gelditzen direnei beste irizpide bat aplikatzen zaie, eta, horren arabera, jaiotza-datako hiru elementuetako biren berdintasunak bilatzen dira: eguna eta hilea, eguna eta urtea, hilea eta urtea.

Hain zorrotzak ez diren batze horien bidez, batzuetan baliozkotzat hartzen ziren mugatik gorako puntuaziodun loturak, baina elementu isolatuen doitasunaren eraginez. Beraz, bi egoera akastun sor zitezkeen: alde batetik, pertsona bera izatea, baina beste gako bat edukitzea; bestetik, gako bereko pertsona desberdinak izatea. Kasu horiek hobetzen ari da, izen-abizenen normalizazioa eginez.

Kapitulua
2

Metodologia

Eustaten fitxategien probabilitate-batzea egiteko diseinaturiko programa oinarritzen duen eredu teorikoa Fellegik eta Sunterrek 1969an “A *theory for Record Linkage*” [1] artikuluan aurkeztutakoa da. Eredu horren oinarriak honako hauek dira.

Eredu teorikoa

Erregistroak batzeko probabilitate metodoek bi artxibo desberdineko erregistroak konparatzen dituzte. Konparatu beharreko artxiboak A eta B dira, eta horien elementuak, hurrenez hurren a eta b.

Bi artxibo horiek elementu komunak dituzte, eta, beraz, batzearen helburua, izan ere, sor daitezkeen AxB bikote guztien artean pertsona, gauza edo erakunde berari buruzkoak zein diren jakitea da. Hau da, helburua multzo hau banatzea da

$AxB = \{(a, b) \mid a \in A, b \in B\}$

multzo disjuntu hauen bilduran

$$M = \{(a, b) \mid a = b, a \in A, b \in B\} \quad (1)$$

eta

$$U = \{(a, b) \mid a \neq b, a \in A, b \in B\} \quad (2)$$

Horiei *match* eta *ez-match* multzo esaten zaie, hurrenez hurren.

Biztanleriaren unitate bakoitzak ezaugarri jakinak ditu lotuta, esaterako, izena, abizenak, adina edo helbidea. Pertsona, gauza edo erakunde berari buruzko erregistroak identifikatu behar dira. Hala eta guztiz ere, fitxategiak sortzeko prozesuak akatsak eta zehaztugabetasunak sar ditzake (kodifikazio, transkripzio eta tekleatze akatsak, aldaketa tipografikoak edo fonetikoak, datuen galera, etab.), sortutako erregistroetan. Akats horien ondorioz, A-ren eta B-ren bi kidek, pertsona berari buruzkoak ez direnek, erregistro berdin-berdinak sor ditzakete, eta, sarriago, A-ren eta B-ren bi kide berdin-berdinek erregistro desberdinak sor ditzakete. A-ren eta B-ren kideei dagozkien erregistroak $\alpha(a)$ eta $\beta(b)$ dira, hurrenez hurren.

Halaber, erregistratu diren pertsonak A-tik eta B-tik aukeraturiko lagin ausazkoetatik datozela jotzen da, eta horiei As eta Bs esaten zaie. Ez da baztertzen As = A eta Bs = B izateko aukera. Aurrerantzean, idazkera errazteko, ezabatu egingo da s azpiindizea.

Bi artxiboko erregistroak bikoteka jartzeko lehenengo pausoa horiek konparatzea da. Konparazioaren emaitza kode multzoa izango da, eta horiek ondorengoak

bezalako esaldietan kodifikatzen dira: “izena bat dator erregistro bietan”, “izena bat dator eta Pedro da”, “izena ez dator bat”, “izena zuriz dago erregistroetariko batean” edo “helbidean adostasuna dago hiriko zonari dagokionez, baina ez kaleari dagokionez”. Formaren aldetik, konparazio bektorea α(a) eta β(b) erregistroen funtzioko bektorea da:

$$\gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], \ldots, \gamma^k[\alpha(a), \beta(b)]\} \quad (3)$$

Ikusten denez, AxB-n oinarrituz zehazturiko funtzioa da γ. Bestalde, γ-ren balizko gauzapen guztien multzoari *konparazio espazio* esaten zaio eta Γ ikurraren bidez adierazten da.

Batze-eragiketaren prozesuan, γ(a, b) ikusten da eta erabaki behar da ea (a, b) bikotea den, (a, b) Є M (erabaki horri *link* esaten zaio eta A1-ren bidez adierazten da), edo bikotea ez den, (a, b) Є U (erabaki horri *ez-link* esan eta A3-ren bidez adierazten da). Hala eta guztiz ere, egoera batzuetarako ezinezkoa izan daiteke bi erabaki horietariko bat hartzea, akats maila zehatzetarako, eta beraz, hirugarren erabaki bat ere onartzen da, A2, eta horri *balizko link* esaten zaio.

Hortik abiatuz, L *batze-araua* Γ konparazio espazioa ausazko erabaki-funtzioen multzoan (D = {d(γ)}) aplikatzearen ondoriozkoa da, eta formula horretan:

$$d(\gamma) = \{P(A1|\gamma), P(A2|\gamma), P(A3|\gamma)\};\ \gamma \text{Є} \Gamma \quad (4)$$

eta

$$\sum_{i=1}^{3} P(Ai \mid \gamma) = 1 \quad (5)$$

Beste modu batera esateko, γ-ren balio bakoitzerako, batze-arauak baliozko hiru erabakietariko bakoitza hartzeko probabilitateak esleitzen ditu.

Aintzat hartu behar dira batze-arau bakoitzari loturiko akats mailak. Onartu egiten da [α(a), β(b)] erregistro bikote bat modu ausazkoan hautatzen dela, probabilitate-prozesuren baten arabera konparatzeko. Horren ondoriozko konparazio bektorea, γ[α(a), β(b)], beraz, aldagai ausazkoa da. m(γ) γ-ren probabilitate baldintzatua (a, b) Є M emanik da, eta hauxe izango da:

$$m(\gamma) = P\left\{\gamma\left[\alpha(a), \beta(b)\right] \mid (a,b) \in M\right\} = \sum_{(a,b)\in M} P\left\{\gamma\left[\alpha(a), \beta(b)\right]\right\}.P\left[(a,b) \mid M\right] \quad (6)$$

Era berean, u(γ) γ-ren probabilitate baldintzatua (a, b) Є U emanik da. Horrenbestez,

$$u(\gamma) = P\left\{\gamma\left[\alpha(a), \beta(b)\right] \mid (a,b) \in U\right\} = \sum_{(a,b)\in U} P\left\{\gamma\left[\alpha(a), \beta(b)\right]\right\}.P\left[(a,b) \mid U\right] \quad (7)$$

Batze-arau horri loturiko bi akats mota daude. Lehenengoa, *match*ekin bat ez datozen erregistro bikoteak konparatzean horik *link* gisa hartzen direnean gertatzen da, eta horren probabilitatea hauxe da:

$$P(A1 \mid U) = \sum_{\gamma \in \Gamma} u(\gamma).P(A1 \mid \gamma) \quad (8)$$

Bigarren akats mota *mach* bat konparatzean *ez-link* gisa hartzen denean gertatzen da, eta horren probabilitatea honako hau da:

$$P(A3 \mid M) = \sum_{\gamma \in \Gamma} m(\gamma).P(A3 \mid \gamma) \qquad (9)$$

Γ espazioko batze-araua μ, λ (0<μ<1, 0<λ<1) akats maila duen batze-araua da, eta L(μ, λ, Γ) gisa adierazten da, baldin eta

$$P(A1 \mid U) = \mu \qquad (10)$$

eta

$$P(A3 \mid M) = \lambda \qquad (11)$$

Esaten da L(μ, λ, Γ) batze-araua hobezina dela, baldin eta

$$P(A2 \mid L) \leq P(A2 \mid L') \qquad (12)$$

erlazioa edozein L'(μ, λ, Γ') deritzonerako mantentzen bada, lehengo erlazioak egiaztatzen dituzten batze-arau guztien artean.

Ikusten denez, definizio honen arabera, erabakitzeko arau hobezinak maximizatu egiten ditu konparazio positiboak hartzeko aukerak (hau da, A1 edo A3 erabakia), akats maila finkoekin. Zentzuzko erabakia dela dirudi, kontuan hartuta A2 erabakia hartzeak eskuzko batze-eragiketak ondorioz ekartzen dituela eta horiek kostu handia dutela. Gainera, bestalde, A2 aukera txikia ez bada, batze-prozesuaren erabilgarritasuna zalantzazkoa izango da.

Aintzat hartu behar da, μ eta λ konbinazio batzuetarako, (10) eta (11) betetzen dituzten batze-arauen multzoa hutsa dela. Hortaz, (10) eta (11) ekuazioak aldi berean D multzoren baterako betetzeko aukera ematen duten μ eta λ konbinazioak baino ez dira kontuan hartzen, eta multzo horretan (4) eta (5) gisa zehazturiko erabaki-funtzioena da. Hori betetzen denean, esaten da (μ, λ) *akats mailen bikote onargarria* dela.

Autoreek batze-arau hobezina proposatzen dute. Horretarako, lehenengo eta behin, ordena bakarra zehazten dute γ gauzatzeko aukera guztien arteko multzo finituan, ondoren adierazitakoaren arabera: γ-ren balioa m(γ) eta u(γ) zero izateko modukoa bada, γ hori gertatzeko aukera zero da, eta beraz, ez dago Γ espazioan sartu beharrik. Orduan, m(γ) > 0, baina u(γ) = 0 betetzen duen edozein γ-rentzat ordena arbitrarioa esleitzen da.

Gainerako γ-ak m(γ) / u(γ) sekuentzia beheranzkoa izateko moduan antolatzen dira. Bestalde, m(γ) / u(γ) deritzenen balioa γ batentzat baino gehiagorentzat berbera denean, ordena arbitrarioa esleituko da.

Beste alde batetik, ordenaturiko {γ} multzoa i azpiindizearen bidez adierazten da; (i = 1, 2,…, N_Γ); bertan, N_Γ Γ-ren puntu kopurua da eta ui = u(γi); mi = m(γi) idazten da.

Bestalde, (μ, λ) akats mailen bikote onargarria da, eta biak ez dira handiegiak izango. Halaber, n, n' zenbaki osoak izango dira; horrela,

$$\mu = \sum_{i=1}^{n} u_i \qquad \lambda = \sum_{i=n'}^{N_\Gamma} m_i \qquad 0 < n \leq n' < N_\Gamma$$

Zehazten da

$$T_\mu = \frac{m(\gamma_n)}{u(\gamma_n)}$$
$$T_\lambda = \frac{m(\gamma_n')}{u(\gamma_n')} \qquad (13)$$

Orduan Fellegik eta Sunterrek frogatu zuten batze-araurik onena, (μ, λ) akats mailetan, ondoren adierazitakoa zela:

$$d(\gamma) = \begin{cases} (1,0,0) & T_\mu \le m(\gamma)/u(\gamma) \ bada \\ (0,1,0) & T_\lambda < m(\gamma)/u(\gamma) < T_\mu \ bada \\ (0,0,1) & m(\gamma)/u(\gamma) \le T_\lambda \ bada \end{cases}$$

Aplikazio askotan, akats maila nahiko altuak onar daitezke, A2 ekintzaren aukera ezabatzeko. Kasu honetan, n eta n' edo T_μ y T_λ aurreko forman γ-ren multzo ertaina hutsa izateko moduan hartzen dira. Beste modu batera esateko, (a, b) bikote bakoitza M-n edo U-n kokatzen da. Izan ere, geuk ere erabaki hauxe hartu dugu gure batze-programan, eta horrela, muga bakarra ezartzen dugu, $T_\mu = T_\lambda$.

Sinplifikazio kasu batzuk

Praktikan, har daitezkeen balioen multzoa oso handia izan daiteke, eta m(γ) eta u(γ) probabilitateen zenbatespena egitea ezinezkoa izan daiteke. Horrenbestez, suposizio batzuk egin daitezke γ-ren banaketa sinplifikatzeko.

Bestalde, onartu egiten da γ-ren osagaiak berrantolatu eta bildu egin daitezkeela, $\gamma = \left(\gamma^1, \gamma^2, \cdots, \gamma^k\right)$ egiteko eran, eta osagai bektorialak estatistikoki independenteak izan daitezke, banaketa baldintzatu bakoitzari dagokionez. Orduan:

$$m(\gamma) = m_1(\gamma^1) \cdot m_2(\gamma^2) \cdot \cdots \cdot m_k(\gamma^k) \qquad (14)$$

$$u(\gamma) = u_1(\gamma^1) \cdot u_2(\gamma^2) \cdot \cdots \cdot u_k(\gamma^k) \qquad (15)$$

Bertan, m(γ) eta u(γ), hurrenez hurren, (6) eta (7)-rekin zehaztuta daude, eta

$$m_i(\gamma^i) = P(\gamma^i \,|\, (a,b) \in M)$$
$$u_i(\gamma^i) = P(\gamma^i \,|\, (a,b) \in U)$$

Notazioa laburragoa izateko, $m(\gamma^i)$ eta $u(\gamma^i)$ idazten da, teknikoki zehatzagoak diren $m_i(\gamma^i)$ eta $u_i(\gamma^i)$ adierazpenak erabili beharrean. Adibidetzat, pertsonen erregistroen konparazio batean, γ^1-ren barruan abizenei loturiko konparazio osagai guztiak joan daitezke, eta γ^2-ren barruan, berriz, helbideei loturiko konparazio osagai guztiak. Izan ere, γ^1 eta γ^2 osagaiak bektoreak dira berez; γ^2 azpiosagaiek, esate baterako, helbideen osagaiak konparatuta kodifikaturiko emaitzak adieraz ditzakete (hiria, kalea, zenbakia,

etab.). Bi erregistrok *match* bat osatzen badute, hau da, pertsona bera adierazten badute, desadostasuna egon daiteke konfigurazioan, akatsen eraginez. Izenean egon daitezkeen akatsak, adibidez, independenteak dira helbidean egon daitezkeenetatik, onarpenaren arabera. Bi erregistrok *ez-match* bat osatzen badute, hau da, bi pertsona desberdin adierazten badituzte, esan daiteke izeneko ezusteko adostasuna independentea dela helbideko ezusteko adostasunetik.

Beste modu batera esateko, onartu egiten da γ^1, γ^2,..., γ^k modu independente baldintzatuan banatuta egotea. Azpimarratu egin behar da ez dela ezer onartzen γ-ren banaketa ez-baldintzatuari buruz.

Argi dago m(γ)/u(γ)-ren edozein funtzio monotono gorakor baliozkoa izan daitekeela, batze-arauen helbururako estatistika-test gisa.

Hain zuzen ere, ratio honen logaritmoa erabili eta hauxe zehatuko dugu:

$$w^k(\gamma^k) = \log m(\gamma^k) - \log u(\gamma^k) \qquad (16)$$

Orduan, hauxe idatzi ahal da:

$$w(\gamma) = w^1 + w^2 + \cdots + w^k \qquad (17)$$

Eta w(γ) geure estatistika-test gisa erabiliko dugu; izan ere, u(γ)=0 edo m(γ)=0 izanez gero, w(γ) = +∞ (edo w(γ) = -∞) izango da, w(y) edozein zenbaki finitu baino handiagoa (edo txikiagoa) baita.

Bestalde, γ^k-k n_k konfigurazio desberdinak har ditzake, $\gamma_1^k, \gamma_2^k, \cdots, \gamma_{n_k}^k$. Beraz,

$$w_j^k = \log m(\gamma_j^k) - \log u(\gamma_j^k) \qquad (18)$$

Batze-prozesuaren intuiziozko interpretazioa egiteko, baliozkoa da koefizienteak positibotzat zehaztuta egotea, konfigurazioek $m(\gamma_j^k) > u(\gamma_j^k)$ betetzen dutenean, eta negatibotzat hartuta egotea konfigurazioek $m(\gamma_j^k) < u(\gamma_j^k)$ betetzen dutenean; propietate hori γ konfigurazio osoari loturiko koefizienteek babesten dute.

Konfigurazio kopuru osoa (hau da, γЄΓ puntu kopurua), jakina, $n_1 \cdot n_2 \cdot \cdots \cdot n_k$ da. Hala eta guztiz ere, osagaietarako zehazturiko koefizienteen propietate batukorra aintzat hartuta, nahikoa izango da $n_1 + n_2 + \cdots + n_k$ koefiziente zehaztea. Horrenbestez, bakoitzari lotutako koefizientea zehaztu eta batuketa horretan erabil daiteke; horrela, Γ-ren kardinala nabarmen murriztuko dugu.

Koefizienteen kalkulua

Batze-prozedura garatzeko funtsezko puntuetariko bat, izan ere, m(γ) eta u(γ) koefizienteak kalkulatzea da, egon eta sor daitekeen bakoitzerako. Zenbait metodo proposatu da hori egiteko; izatez ere, Fellegiren eta Sunteren lanean bertan ere bi modu desberdin agertzen dira, gai horri aurre egiteko. Lan honetan, erregistraturiko

datuen konfigurazioak batu beharreko artxiboen multzoan agertzeko maiztasunean oinarrituriko metodoa aukeratu dugu.

Pentsatu bi artxiboetako erregistroen osagaietariko bat *abizena* eremua dela. Bi erregistroetako abizenak konparatuta, konparazio bektoreko osagai bat lortuko dugu. Osagai hori "*abizena ados*" edo "*abizena ez ados*" edo "*abizena ez da artxibo batean edo bietan agertzen*" erako konparazio hutsa izan daiteke.

Bi artxiboetariko edozeinetan, abizena akatsen batekin erregistra daiteke. Akatsik gabeko bi artxiboetako abizen guztien zerrenda egin daitekeela onartzen da, baita artxiboetan abizen horiek dituzten pertsona-kopuruen zerrenda ere.

A-n eta B-n maiztasunak honako hauek dira:

$$f_{A1}, f_{A2}, \cdots, f_{Am}; \quad \sum_{j=1}^{m} f_{Aj} = N_A$$

eta

$$f_{B1}, f_{B2}, \cdots, f_{Bm}; \quad \sum_{j=1}^{m} f_{Bj} = N_B$$

A∩Bn maiztasunak honako hauek dira:

$$f_1, f_2, \cdots, f_m; \quad \sum_{j=1}^{m} f_j = N_{AB}$$

Hain zuzen ere,

f_{A1} aldagaietariko baten balioa A artxiboan agertzeko maiztasuna da
f_{B1} balio hori B artxiboan agertzeko maiztasuna da
f_1 balio hori A eta B artxiboetan agertzeko maiztasuna da

Honako notazio hau egin behar da:

e_A edo e_B aldagai baten balioa bi artxiboetariko bakoitzean txarto erregistratzeko probabilitateak. (Onartu egiten da txarto erregistratzeko probabilitatea independentea dela balio bakoitzetik).

e_{A0} edo e_{B0} aldagai baten balioa bi artxiboetariko bakoitzean ez erregistratzeko probabilitateak. (Onartu egiten da ez erregistratzeko probabilitatea independentea dela balio partikular bakoitzean).

e_T aldagai baten balioa bi artxiboetan modu desberdinean agertzeko probabilitatea, bietan ondo erregistraturik egon arren. (Esate baterako, bi artxiboak une desberdinetan sortuta egon eta pertsonak izena aldatu duenean.)

Azkenik, onartu egiten da e_A eta e_B nahiko txikiak direla, berdin-berdinak (baina akastunak) diren bi sarreren doitasuna egoteko probabilitatea oso txikia izateko eta txarto erregistraturik, erregistratu barik nahiz aldatuta egoteko probabilitateak elkarrekiko independenteak izateko.

Bestalde, m eta u kalkulatzeko arauak ematen dira, eta horiek γ-ren konfigurazio hauei dagozkie: aldagaia bat dator eta zerrendako j-garrena da, aldagaia ez dator bat, artxiboetarikoren batean ez dago aldagairik.

$$m(\text{aldagaia bat dator eta zerrendako j - garrena da}) =$$

$$\frac{f_j}{N_{AB}}(1-e_A)(1-e_B)(1-e_T)(1-e_{A0})(1-e_{B0}) = \tag{19}$$

$$\frac{f_j}{N_{AB}}(1-e_A-e_B-e_T-e_{A0}-e_{B0})$$

$$m(\text{aldagaia ez dator bat}) =$$

$$\left[1-(1-e_A)(1-e_B)(1-e_T)\right](1-e_{A0})(1-e_{B0}) = \tag{20}$$

$$e_A+e_B+e_T$$

$$m(\text{artxiboetarikoren batean ez dago aldagairik}) =$$

$$1-(1-e_{A0})(1-e_{B0}) = \tag{21}$$

$$e_{A0}+e_{B0}$$

$$u(\text{aldagaia bat dator eta zerrendako j - garrena da}) =$$

$$\frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}(1-e_A)(1-e_B)(1-e_T)(1-e_{A0})(1-e_{B0}) = \tag{22}$$

$$\frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}(1-e_A-e_B-e_T-e_{A0}-e_{B0})$$

$$u(\text{aldagaia ez dator bat}) =$$

$$\left[1-(1-e_A)(1-e_B)(1-e_T)\sum_j\frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}\right](1-e_{A0})(1-e_{B0}) = \tag{23}$$

$$\left[1-(1-e_A-e_B-e_T)\sum_j\frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}\right](1-e_{A0}-e_{B0})$$

$$u(\text{artxiboetarikoren batean ez dago aldagairik}) =$$

$$1-(1-e_{A0})(1-e_{B0}) = \tag{24}$$

$$e_{A0}+e_{B0}$$

Bestalde, f_{Aj} / N_A, f_{Bj} / N_B, f_j / N proportzioak, aplikazio askotan, berbera balira bezala har daitezke. Horixe gertatzen da, esate baterako, bi artxiboak biztanleria beretik atera badira. Maiztasun horiek zuzen-zuzen zenbatetsi ahal dira artxiboetatik.

Gogoratu behar da, (18)-ren arabera, aldagai baten osagaiak koefiziente osora egiten duen ekarpena log(m/u) dela eta zenbaki jakina baino koefiziente handiago batekiko konparazioak *link* gisa hartzen direla; zenbaki horretatik beherako koefizientea dutenak, ostera, *ez-link* gisa hartuko dira. Argi dago, (19-24) aintzat hartuta, aldagaiari buruzko adostasunak koefiziente positiboa sortuko duela eta aldagaiaren balioa zenbat eta arraroagoa izan koefizientea hainbat eta handiagoa izango dela. Era berean, aldagairen bati buruzko desadostasunak koefiziente negatiboa sortuko du, eta hori murriztu egiten da e_A, e_B, e_T akatsekin. Erregistroren batean aldagaiaren baliorik agertzen ez bada, koefizientea zero izango da.

Zerrenda estandarizatuen sorrera

Koefizienteen kalkulurako lehen azaldutako metodoa erabiltzeko, akatsik gabeko konfigurazio guztien zerrenda eduki behar da, prozesuan esku hartzen duten aldagai guztiei dagokienez. Hori erraza da, adibidez, sexu motako aldagaien kasuan; izan ere, jakina da bi konfigurazio zuzen baino ez daudela eta bi balio horiekin ados ez dagoen edozein balio akastuna dela.

Izena edo abizena motako aldagaien kasuan, lana ez da hain erraza. Aldagai horien tipografia-akatsen kopurua nahiko handia izaten da, eta beraz, konfigurazioen zerrenda atera baino lehen, zuzendu egin behar dira akats guztiak.

Ondoren, EUSTATek garaturiko metodoa azalduko dugu; metodo horrek, izena edo abizena motako aldagai bakoitzaren konfigurazio guztietan oinarrituz, ahalegina egiten du jatorrizko konfigurazio berekoak izan arren desberdinak diren konfigurazioak zerrendatzeko (desberdinak dira konfigurazioak modu zuzen bat baino gehiago eduki dezakeelako edo baten batean akatsen bat egin delako).

Ideia hauxe da: Jatorrizko C zerrendatik abiatu behar da, batu beharreko bi fitxategietan agertzen diren aldagai baten konfigurazio guztiekin. Helburua kodeen D zerrenda eta *est* izeneko aplikazioa (C-rena D-n) sortzea; horrela, C-ren bi elementuk (balio bera dutenak) irudi bera izango dute *est*-en. Lehen esan dugunez, bi C elementuk irudi bera izan dezakete, bai balio bera konfigurazio desberdinen bidez adierazten delako, bai horietarikoren bat erregistratzean akatsa egin delako.

Hortaz, prozesua, alde batetik, aldagai baten balio bera adierazi arren desberdinak baina zuzenak diren konfigurazioak lotzean oinarritzen da. Egoera horren adibideak dira ENEKOITZ-ENEKOIZ, LEZURI-LEXURI edo JAVIER-XABIER izenak. Beraz, bermatu egin behar da konfigurazio horien *est* bidezko irudia berbera dela.

Horretarako, estandarizazio irizpideak erabili dira, besteak beste, honako hauek:

- Ñ karakterearen ordez ¥ ikurra egotearen eraginezko akatsak zuzentzen dira.
- Marratxoak, puntuak eta komak kentzen dira.
- EZ DAUKA, EZ DAGO, EZ DA AGERTZEN, EZ DAKIGU eta halako balioen ordez missing bat erabiltzen da.

- Azentuz edo dieresiz agertzen diren bokaletatik kendu egiten da azentua edo dieresia.

- Zenbakiak ezabatu egiten dira. (1, 2, 3, 4, 5, 6, 7, 8, 9)

- Karaktere batzuk ezabatu egiten dira. (‘, ´, `, ¨, ^, *, “, º, ª, #, Ç, ¥)

- Zeroen ordez ‘O’ letra sartzen da.

- Partikulak ezabatu egiten dira. (D, DA, DI, DO, L, ETA)

- Txikigarri batzuk itzuli egiten dira.

- Antzekotasun fonetikoa edo grafikoa duten karaktereak edo karaktere taldeak identifikatzen dira.

- Letra batean izan ezik gainerako guztietan bat datozen izen eta abizen bikoteak aztertu eta horiek konfigurazio beretik etor daitezkeen aztertzen da. Horiek beste alde batetik aztertzen dira; izan ere, grafia batzuek, zenbait kasutan antzekotasun fonetikoa edo grafikoa eduki ez arren, karaktere bera adieraz dezakete.

Bestalde, akatsen bat izan duten konfigurazioak aurkitu behar dira. Horiek identifikatu ondoren, *est* aplikazioaren bidez duten irudia jatorrizko konfigurazio zuzenak duenaren berbera izango da. Horrela, FERANNDO, FERNNDO eta FERANANDO konfigurazioak FERNANDOren kode berarekin bat etorriko dira *est* aplikazioan.

Lan hau egitean sortzen den lehenengo oztopoa, izan ere, akatsen bat duten konfigurazioak detektatzeko modua da. Zentzuzko ondorioak ateratzeko, konfigurazioak (batu beharreko artxiboetan) agertzen diren maiztasunetatik abiatu behar da. Horrela, cЄC bakoitzerako, $freq_0(c)$-k aditzera ematen du c bi artxiboetan agertzeko maiztasuna.

Esate baterako, c izatez n luzerako karaktere katea da, batu beharreko fitxategietariko batean agertzen dena, eta, teorian, ondo erregistratu da.

Honelaxe adierazten da:

e = karakterea ondo erregistratzeko probabilitatea

1-e = karakterea txarto erregistratzeko probabilitatea

Karaktereen c kateak n karaktere dituenez, hori ondo erregistraturik egoteko probabilitatea (pentsatu behar da ondo erregistraturik dagoela) hauxe izango da:

$P(0 \text{ akats} \mid \text{luzera}(c) = n) = e^n$.

Beste alde batetik, c balioa hartzen duen aldagaidun jatorrizko biztanleriako kide kopurua freq(c) izanez gero, $freq_0(c)$-k aztertzen diren artxiboetan c konfigurazioa agertzeko kasu kopurua adierazten duenez:

$freq(c) * P(0 \text{ akats} \mid \text{luzera}(c) = n) = freq_0(c)$

Horrenbestez, freq(c) ondoren adierazitakoa izango da:

$freq(c) = freq_0(c)/e^n$

Era berean, C-ren konfigurazio zuzen bakoitzerako, transkripzioan akats bat edo gehiago duten konfigurazioen kopuru osoa kalkula daiteke.

$$P(1\,\text{akats}\,|\,\text{luzera(c)} = \text{n}) = \binom{n}{1} e^{n-1}(1-e) = n \cdot e^{n-1} \cdot (1-e)$$

Horrela, akats bat duen c-tik sortutako konfigurazioen kopuru osoa kalkula daiteke:

$$freq_1(c) = P(1\,\text{akats}\,|\,\text{luzera(c)} = \text{n}) \cdot freq(c) = \frac{freq(c) \cdot n \cdot (1-e) \cdot e^n}{e}$$

Era berean, bi akats dituen c-tik sortutako konfigurazioen kopuru osoa kalkula daiteke:

$$P(2\,\text{akats}\,|\,\text{luzera(c)} = \text{n}) = \binom{n}{2} e^{n-2}(1-e)^2 = \frac{n \cdot (n-1) \cdot e^{n-2} \cdot (1-e)^2}{2}$$

$$freq_2(c) = P(2\,\text{akats}\,|\,\text{luzera(c)} = \text{n}) \cdot freq(c) = \frac{freq(c) \cdot n \cdot (n-1) \cdot (1-e)^2 \cdot e^n}{2 \cdot e^2}$$

Egoera honetan, helburua aldagai baten konfigurazio akastunak zuzenekin erlazionatzea da, ager daitekeen konfigurazio akastunen kopuru osoaren zenbatespena jakin ondoren. 2 akats baino gehiagoko kasuak ez dira kontuan hartzen, gertatzeko probabilitatea oso txikia baita. Nolanahi ere, nahi izanez gero, aintzat har daitezke, akats 1eko edo 2ko kasuetarako egindako arrazoibideari jarraituz.

Oinarrizko ideia akastunak izan daitezkeen konfigurazioak eta zuzenak izan daitezkeenak konparatzea da; horren arabera, zerrenda bat sortu eta, bertan, batzuk nahiz besteak zerrendaturik agertuko dira, konfigurazio batzuk besteetatik banantzen dituzten baliozko akatsen kopurua 2koa edo hortik beherakoa denean.

Intuiziozko ideia, izan ere, artxiboetan maiztasun handiz agertzen diren konfigurazioak zuzenak direla jotzea da. Hortaz, muga bat finkatu eta, horren arabera, C multzoa honako azpimultzo hauetan banatuko da:

$C_1 = \{c \in C \mid freq(c) > lim\}$

$C_2 = \{c \in C \mid freq(c) \leq lim\}$

Gero, C_2-ko balioak C_1-koekin konparatzen dira, horiek erlazionatu ahal izateko, akats kopurua baxua denean. Horrela, lortzen den zerrendan C_2-ko kide bakoitza C_1-ko kide batekin, batzuekin edo ezeinekin erlazionaturik ager daiteke, eta alderantziz.

Ondoren, erlazio horietatik guztietatik euren artean lotuko direnak hautatzen dira. Horretarako, lehen aipaturiko kasuetan oinarrituz, konfigurazio jakinarekin

erlazionaturiko konfigurazioen kopuru osoak ez du inoiz ere gaindituko aldez aurretik zenbatetsitako balioa.

Horrela, C_2-ko elementu bakoitza C_1-ko elementu batekin edo batzuekin erlazionatu da, eta alderantziz. Erlazio multzo horretatik, behin betiko azpimultzoa hautatu behar da; bertan, C_2-ko elementu bakoitza asko jota C_1-ko elementu batekin erlazionaturik egongo da, eta C_1-ko c konfigurazio batekin erlazionaturiko konfigurazioen kopuru osoa ez da inoiz $freq_1$ + $freq_2$ baino handiagoa izango.

$$freq_1 + freq_2 = \frac{freq(c) \cdot n \cdot (1-e) \cdot e^n}{e} + \frac{freq(c) \cdot n \cdot (n-1) \cdot (1-e)^2 \cdot e^n}{2 \cdot e^2}$$

Hautapen hori egiteko, honako irizpide hau finkatzen da: hasierako zerrendatik azpimultzo bat hautatzen da; bertan, C_2-ko elementu bakoitza C_1-ko elementu bakarrarekin erlazionaturik agertuko da; izan ere, akastuna izan daitekeen konfigurazio bat jatorrizko konfigurazio zuzen batetik sortutakoa izan behar da. C_2 azpimultzoko konfigurazio bat C_1-ko batekin baino gehiagorekin erlazionaturik badago, artxiboetan sarrien agertzen dena aukeratzen da, edo akats kopururik txikiena duena.

Halako zerrenda bat sortu ondoren, gerta daiteke C_1-ko elementu bakoitza C_2-ko elementu batekin baino gehiagorekin erlazionaturik egotea. Beraz, horien guztien artean, C_1-ko elementutik etortzeko probabilitaterik handiena dutenak aukeratuko dira (akatsen bat egin ondoren), aintzat hartuta horietariko bakoitzarekin erlazionaturiko konfigurazioen kopuru osoa ez dela handiagoa izango egindako zenbatespena baino.

Prozesua berriro egin behar da, harik eta C_2-ko elementu guztiak C_1-ko elementuren batekin erlazionatu arte, edo C_2-ko elementu guztiak egindako zenbatespenaren berdina den konfigurazio kopuruarekin erlazionatu arte.

Izen edo abizen motako aldagaietan akatsak zuzentzen saiatzeko erabiltzen den metodo honetaz gain, programan aurrerago beste iritzi batzuk ematen dira aldagai horiei buruz, lehenengo ahaleginetan batzen ez diren bikoteetarako; izan ere, horiei buruzko ikerketa sakontxoagoa egiten da.

Lehenengo eta behin, izenetariko bat bestearen txikigarria ote den egiaztatzen da, kasu ezagun batzuk kontuan hartuta.

Halaber, bai izenen eta bai abizenen kasuan, egiaztatu egiten da ea artxiboren batean balio konposatua eta bestean osagaietariko bat soilik agertzen den.

Bi izenen edo bi abizenen arteko erlazioa ezartzeko beste modu bat, izan ere, Jaro-ren karaktere-kateak konparatzekoa da. Horrek bi kateen karaktere komunen kopurua, bien luzera eta transposizio kopurua kalkulatzen ditu, 0.0tik 1.0ra bitarteko antzekotasun neurria ondorioztatzeko.

Transposizioak lekuz kanpo dauden karaktere komunen bikoteen bidez adierazita datoz. Orduan, Jaroren karaktere-kateak konparatzekoaren balioa, α eta β kateetarako, formula honen bitartez adierazten da:

$$S_J = \frac{1}{3}\left(\frac{komunak}{luzera(\alpha)} + \frac{komunak}{luzera(\beta)} + \frac{komunak - transposizioak}{komunak}\right)$$

Ikusten denez, kateak berdin-berdinak badira (komunak = luzera (α) = luzera(β) eta transposizioak = 0), orduan $S_J = 1$.

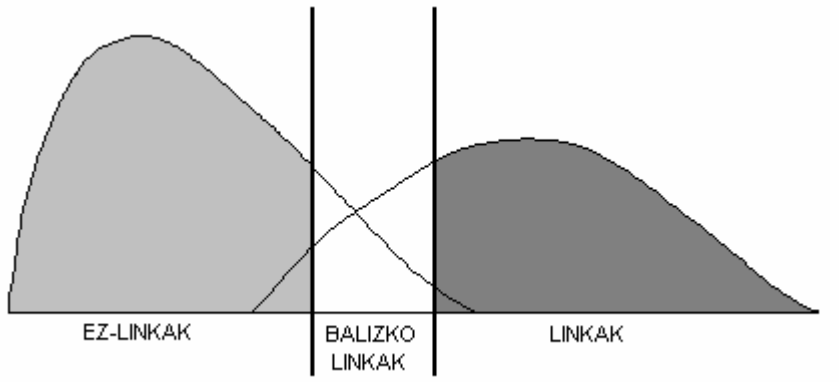
Batzeko aldagaitzat pertsonaren bi abizenak hartzen direnean, egiaztatu egin behar da ea horiek trukatu diren. Aurrerago azalduko dugunez, batze-programa exekutatu ondoren lorturiko emaitzak ikustean, erregistro bikote batzuk pertsona berarenak ziren, baina ez ziren batzen, abizenen ordena trukaturik zegoelako.

Azken kontrastea geure arloaren berezitasunari lotuta dago. Euskal Autonomia erkidegoan bi hizkuntza ofizial daude, gaztelania eta euskara. Beraz, bi fitxategitan konfigurazio desberdinak sor ditzaketen baliokidetasun onomastikoak daude. Horrenbestez, gaztelania-euskarazko izenen baliokidetasun batzuk dituen datu-basea sortzen da, eta, izen bikote bakoitzerako, horiek zerrenda horren barruan ote dauden egiaztatzen da. Zerrendaren barruan badaude, euren arteko erlazioa dagoela esaten da.

Horren ostean, koefizienteak berriro kalkulatu eta orain erabiltzaileak ezarritako mugatik gorako koefizientea dutenak hautatzen dira.

Muga ezartzeko metodoa

Konfigurazio garrantzitsu guztiak (γ_j^k) adierazi ondoren, eta horiei loturiko koefizienteak ($w_j^k; k = 1,2,\ldots,K; j = 1,2,\ldots,n_k$) zehaztutakoan, orain μ eta λ osagaiei dagozkien T_μ eta T_λ mugako balioak ezarri behar dira. Muga horiek (T_μ eta T_λ) hiru zonatan banatzen dituzte erregistro bikoteak.



T_μ mugatik ezkerrerako zonan, *ez-link* gisa sailkatuko diren erregistro bikoteak sartuko dira. T_λ mugatik eskuinetara dagoen zonan, *link* gisa artxibatuko diren erregistro bikoteak daude. Eta bitarteko zonan, azkenik, *balizko link*ak izeneko erregistro bikoteak

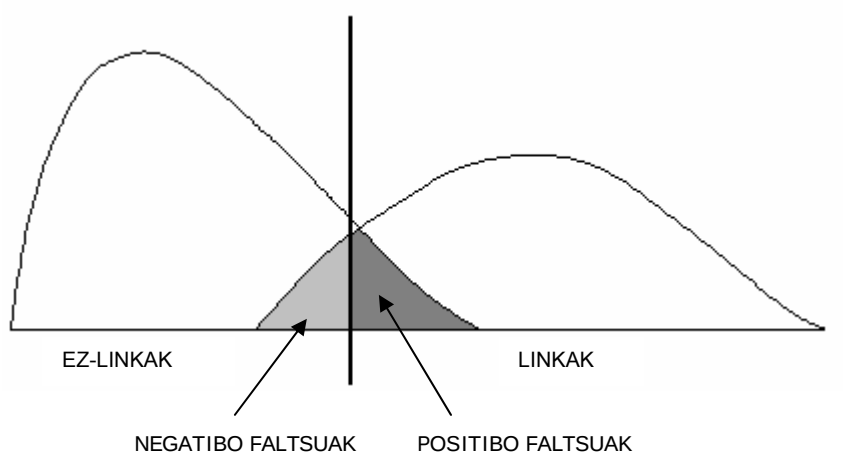
sartuko dira, eta horiek eskuzko ikuskapena behar dute ea zein taldetan sailkatu behar diren jakiteko.

Halako erabakiak hartzean bi akats egiten ari gara. Lehenengo akatsa pertsona berarenak ez diren bi erregistro *link* gisa finkatzean egiten da. Bigarren akatsa, berriz, izatez pertsona berarena den erregistro bikotea *ez-link* gisa finkatzean. Akats horiek ondoren adierazitakoaren arabera kuantifikatu ahal dira:

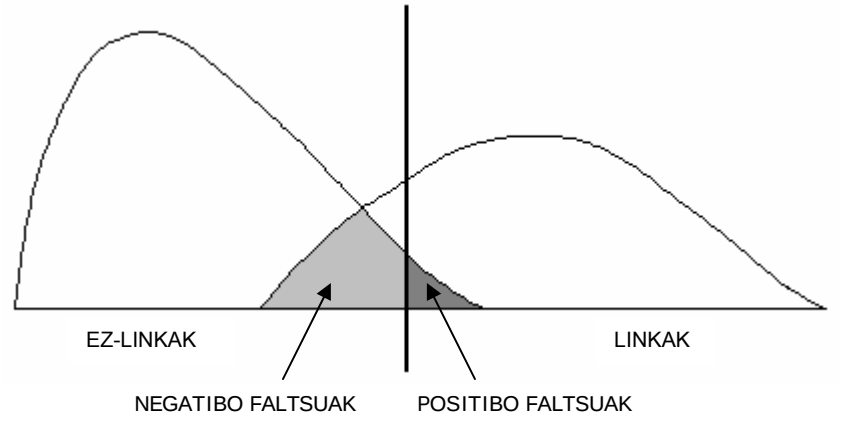
$$\mu = P(A1 \mid U) = \sum_{\gamma \in A1} u(\gamma)$$ *Link* bikoteen proportzioa U-n.

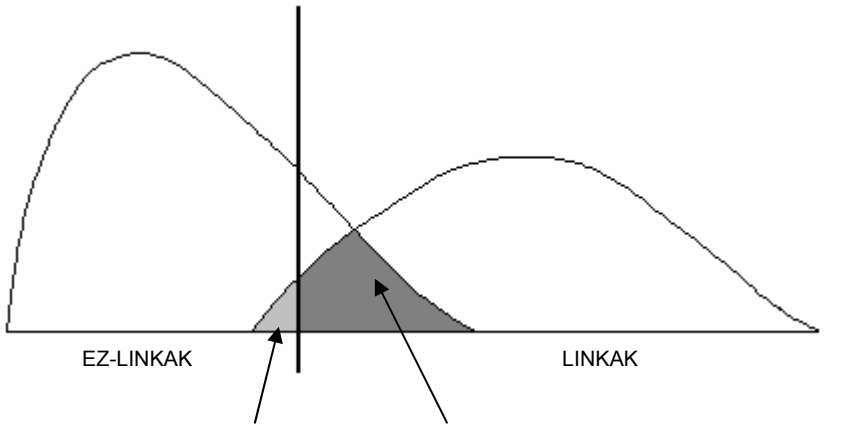
$$\lambda = P(A3 \mid M) = \sum_{\gamma \in A3} m(\gamma)$$ *Ez-link* bikoteen proportzioa M-n.

Bi mugen artean dauden erregistro bikoteak eskuz prozesatu behar ez izateko, EUSTATen muga bakarra hartzea erabaki zen, $T = T_\mu = T_\lambda$.



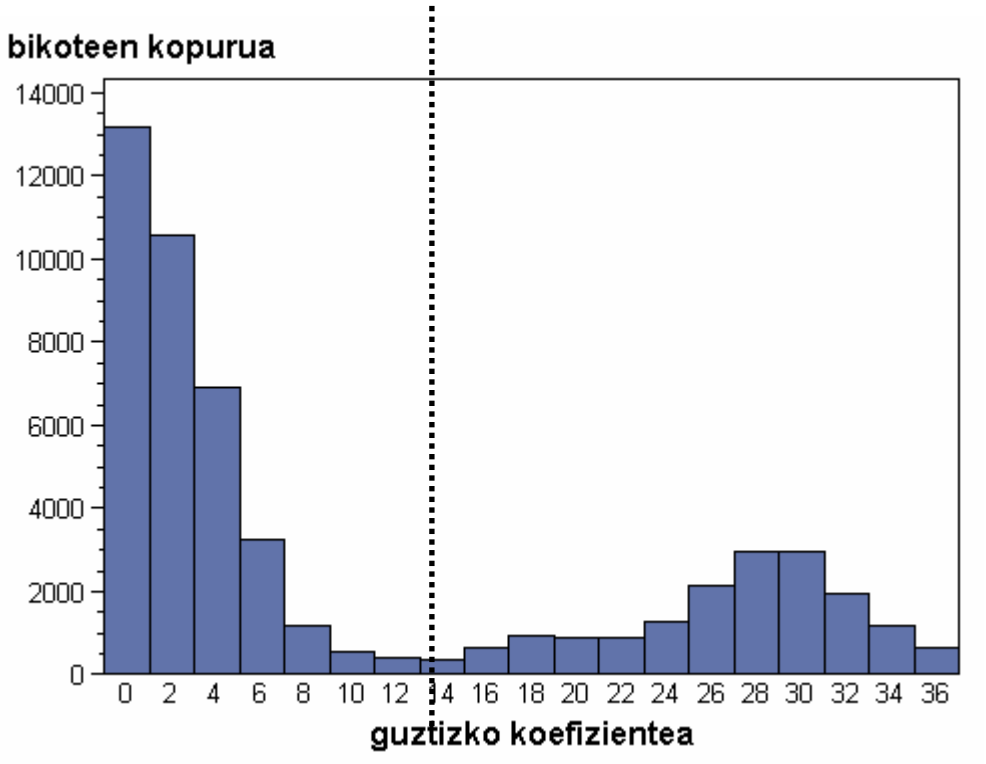
Bi mugak berdinduta, muga bakarraren balioa zenbat eta handiagoa izan, negatibo faltsuen kopurua ere hainbat eta handiagoa izango da, baina positibo faltsuen kopurua txikiagoa izango da. Era berean, muga bakarraren balioa zenbat eta txikiagoa izan, negatibo faltsuen kopurua ere hainbat eta txikiagoa izango da, eta positibo faltsuen kopurua, berriz, handiagoa.





Muga hori ondo ezartzeko metodo erraza dago eta metodo hori kalkulaturiko guztizko koefiziente guztien maiztasunen histogramaren behaketan oinarritzen da. Histogramak informazio baliotsua ematen du erabakia hartzen laguntzeko, horrek koefizienteen balio desberdinetarako hartzen duen forman oinarrituz. Hau da, pertsona berari dagozkion bikoteek koefiziente handia izan behar dute, eta pertsona bakarrari ez dagozkion erregistro bikoteek koefizientea txikia izan behar dute (are gehiago, koefiziente negatiboa).

Praktikan hori ez da horrela gertatzen, aldagaietan akatsak eta kasualitatezko doitasunak egotearen ondorioz. Nolanahi ere, maiztasunen histogramaren forma ondoren adierazitakoaren antzekoa izango da:



Horrela, mugako balioa aukeratzen da, eta balio horretatik gorako koefizientea duten erregistro bikoteak *link* gisa ezartzen dira; koefiziente txikiagoko bikoteak, berriz, *ez-link* gisa adierazten dira.

*Blocking*a

Bi fitxategi (A eta B) batzeko orduan, onena A-ren erregistro bakoitza B-ren erregistro guztiekin konparatzea da; hala eta guztiz ere, horretarako, neurri ertaineko bi fitxategiren kasuan, |AxB| konparazioa egin beharko litzateke, eta kopuru hori handiegia da.

Horrenbestez, *match* batekin bat etor daitezkeen erregistro bikoteak soilik ateratzen dituen eskema erabiltzen da. Prozesu horren izena *blocking* da.

Horren helburua programak egin beharreko konparazio kopurua murriztea da; horretarako, fitxategiak elkarrekiko baztertzaileak eta zehatzak diren blokeetan banatzen dira; bloke horiek ikusitako *match* bikoteen proportzioa handitzeko diseinatzen dira, konparatu beharreko erregistro bikoteen kopurua murriztu ahal izateko.

*Blocking*a, oro har, bi fitxategien sailkapenen bidez inplementatzen da, aldagai bat edo batzuk aintzat hartuta. Esate baterako, bi fitxategiak posta-kodearen arabera sailkatzen badira, posta-kode bereko bikoteak erabiliko dira. Posta-kode bera erregistraturik ez duten bikoteak ez dira aztertuko eta, hortaz, *ez-match* gisa hartuko dira berez.

Jakina, erregistro bakoitzak berorrek fitxategi osoan duen erlazioa bilatuz gero, benetako *match* bat aurkitzeko aukera handiagoa izango litzateke, azterketaren barruan erregistro guztiak sartuko bailirateke.

Hala eta guztiz ere, prozesu horren kostua handiegia izango litzateke. Horrela, *blocking*a kostu konputazionalaren (erregistro bikote gehiegi aztertzea) eta *ez-match* faltsuen proportzioen (erregistro bikoteak *ez-match* gisa sailkatzea, erregistroak bloke berekoak izan ezean) arteko konpentsazioa da. Oztopo hori arintzeko, komenigarria da *blocking* aldagai bat baino gehiago erabiltzea eta horiek elkarrekiko independenteak izatea.

Halaber, *blocking*aren irizpidetzat aukeratzen den aldagaia ere oso garrantzizkoa da. *Blocking* idealeko osagai batek banaketa orekatuko balio kopuru nahiko handia izan behar du. Hau da, artxiboa banatzen duten blokeetariko ezeinek ere ez du handiegia izan beharko, eta bloke kopurua ez da txikiegia izango, exekuzioan denbora aurreztu ahal izateko. Era berean, aldagaiek akatsak izateko probabilitate baxua izango dute (hau da, koefiziente altua); izan ere, aldagai horretan akatsen bat egin bada, horren eraginpeko erregistroak bloke desberdinetan jarriko dira, eta beraz, ez dira konparatuko eta ezin izango dira batu.

Orain, geure batze-programan inplementaturiko *blocking* motak azalduko ditugu. Esan behar da irizpide horiek modu modularrean programatu direla, etorkizunean beste irizpide batzuk ere erabili ahal izateko.

Jaiotze-dataren irizpidea.

Irizpide honek erregistroko jaiotza-data aldagaiaren araberako blokeak egiten ditu. Hau da, jaiotza-data bereko erregistro bikoteak soilik ikertuko ditu.

Erregistroak ez dira aintzat hartuko aldagai honetarako baliorik ez dutenean; izan ere, ez dago arrazoirik pentsatzeko fitxategi batean aldagai hori ez agertzeak aditzera ematen duela bigarren fitxategian ere agertuko ez denik.

Horrenbestez, komenigarria da *blocking* irizpide hau bakarrik ez erabiltzea, batez ere jaiotza-data aldagaia hutsik egoteko probabilitatea handia denean.

Gako irizpidea.

Irizpide honek, artxiboen sailkapenerako, gako erako aldagaia erabiltzen du, esate baterako, *etxebizitzaren identifikatzailea* aldagaia. Aldagai horrek ez du erregistro bakoitza identifikatzen; izan ere, *etxebizitza identifikatzaile* bereko pertsona bat baino gehiago egon daiteke, guztiak etxebizitza berean bizi direlako.

1. testu-kodeko irizpidea.

Irizpide hau pertsonaren abizenetan oinarrituz osaturiko karaktere kode jakinean oinarritzen da. Fitxategi bakoitzerako bi aldagai berri sortzen dira, horiek abizen bakoitzeko lehenengo letraren eta hurrengo bi kontsonanteen konbinazioa biltzen dute. Azken *blocking*aren aldagaia aurreko bi aldagaien bilketa da, eta horrela, bi abizenetan oinarrituz gako alfabetikoa sortzen da. Karaktere serie hau abizen osoen ordez erabiltzen da, aldagai horren balioak erregistratzean egin daitezkeen akatsak ekiditeko.

2. testu-kodeko irizpidea.

Irizpide hau pertsonaren abizenetan oinarrituz eraturiko karaktere-kodean oinarritzen da. Fitxategi bakoitzerako bi aldagai berri sortzen dira eta horiek abizen bakoitzaren lehenengo hiru letrak biltzen dituzte. Azken *blocking*aren aldagaia aurreko bi aldagaien bilketa da, eta horrela, bi abizenetan oinarrituz gako alfabetikoa sortzen da. Aurreko kasuan bezala, karaktere batzuk erabiltzen dira, abizen osoen ordez, aldagai honen balioak erregistratzean egin daitezkeen akatsak ekiditeko.

Kapitulua
3

SAS programazioa

Ondoren, eta orokorrean, EUSTATen garaturiko batze-programa azaltzen da, berori osatzen duten SAS makroak zehaztuz.

Input aldagaiak

Programa exekutatu baino lehen, erabiltzaileak beharrezko datu batzuk sartu behar ditu, horren jarduketa egokia izateko. Zehaztu behar diren datuak ondoren adierazitakoak dira:

Karpeten kokalekua.

Erabiltzaileak hiru aldagai zehaztu behar ditu, honako karpeta hauen kokalekuarekin

1 – Batu beharreko bi fitxategiak dituen karpeta. Karpeta horretan programaren emaitzen fitxategiak ere gordeko dira. Hiru izango dira: bata baturiko erregistro bikoteak dituena eta beste biak bi fitxategietariko bakoitzean batu ezin izan diren erregistroak dituztenak

2 – Programa exekutatu ahala sortzen doazen datu-fitxategiak biltzeko karpeta.

3 – Programa exekutatu ahala sortzen eta ezabatzen doazen datu laguntzaileen fitxategiak biltzeko karpeta.

Fitxategien izena.

Erabiltzaileak batu behar diren bi fitxategien izenak ere adierazi beharko ditu.

Batzeko aldagaien izenak eta mota.

Erabili nahi den batzeko aldagai bakoitzerako, aldagai horrek fitxategi bakoitzean zer izen duen eta zer aldagai mota den adierazi behar da programan; honako mota hauek egon daitezke:

0 – Erregistroa identifikatzeko gakoa.

1 – Izenaren aldagaia.

2 – Abizenaren aldagaia.

3 – Sexuaren aldagaia.

4 – Jaiotze-dataren aldagaia.

5 – Jaiotze-urtearen aldagaia.

6 –Zenbakiekin soilik osaturiko gako aldagaia.

7 – Zenbakiekin eta letrekin osaturiko gako aldagaia.

Gutxienez, identifikazioko gako aldagai bat zehaztu behar da erregistroak bereizteko, eta izenaren aldagai bat, abizenaren bat eta sexuaren beste bat ere bai.

Sexuaren adierazleak.

Gizonezko sexua eta emakumezko sexua adierazten duten balioak zein diren ere adierazi behar zaio programari.

***Blocking* irizpide motak.**

Programa *blocking*aren zein irizpiderekin exekutatu nahi den ere erabaki behar da.

Ahalezko motak honako hauek dira:

1 – Jaiotze-data.

2 – Gakoa.

3 –1. testu-kodea.

4 –2. testu-kodea.

Irizpide horietariko bakoitza aurreko kapituluko atal egokian azalduta dago.

Parametroak.

Batze-aldagai bakoitzerako, koefizienteak kalkulatzeko behar diren parametro batzuen balioa ezarri behar da. Honako hauek dira:

- Aldagai horren balio bat lehenengo artxiboan txarto erregistraturik egoteko probabilitatea.
- Aldagai horren balioa bigarren artxiboan txarto erregistratuta egoteko probabilitatea.
- Aldagai horren balioa bi fitxategietan desberdina izateko probabilitatea, bietan ere ondo idatzita egon arren.

Gainera, honako aldagai orokor hauei ere balio bat esleitu behar zaie:

- Karaktere bat ondo erregistratuta egoteko probabilitatea.
- Izen edo abizen bat ohikoa den ala ez erabakitzeko erabiltzen den balioa.

Muga.

Halaber, mugako balio bat ere ezarri beharko da, batu beharreko erregistro bikoteak bereizteko. Une jakinean, programaren exekuzioa gelditu, histograma agertu eta, bertan oinarrituz, mugako balioa ezarri eta sartu egiten da.

SAS makroak

Makroaren izena	Azalpena
ejecución	Programako makro nagusia da. Makro horri bi dei egiten zaizkio. Lehenengo deian, fitxategiak eta batze-aldagaiak aztertu, aldagaiak estandarizazioa nahiz homogeneizazioa egin, balioak esleitzeko prozedura gauzatu (lehen azaldutako zerrenda estandarizatuak aintzat hartuta), koefizienteak kalkulatu, blokeak banatu eta blokeren bateko erregistro bikoteen koefizienteen banaketa aztertzen da. Banaketa hau histograma forman agertzen da. Programak itzultzen dio histograma erabiltzaileari, eta, horretan oinarrituz, *linkak* eta *ez-linkak* bereizteko erabiliko duen mugako balioa ezartzen du. Makro horri bigarren dei bat egingo zaio, azken pausoa egin dezan; bertan, mugatik beherako baina bertatik oso hurbileko koefizientea duten erregistro pareak ikertzen dira.
paso1_análisis	Lanerako erabiliko diren liburutegiak sortzen dira. 1. artxiboa 2. artxibotik bereizten da (1. artxiboa txikiena izango da eta erregistro gutxiago izango du, fitxategi hau guztiz batu nahi dela jotzen baita). Aldagaiak aztertu eta erabiltzaileak sarturiko datuetan inkongruentziarik ez dagoela egiztatu behar da.
recuento	Datu-fitxategi bateko erregistro kopurua kalkulatu behar da.
paso2_confección_listas	Makro honek beste makro batzuei dei egiten die, izen eta abizen motako aldagaiak homogeneizatzeko eta estandarizatzeko prozesua egiteko; izan ere, halako aldagaiek akats asko eduki ditzakete eta, gainera, konfigurazio bera modu desberdinetan adierazita egon daiteke.

enies	Ñ letraren ordez ¥ ikurra duten konfigurazioak zuzentzen ditu.
estandarizar	Izen edo abizen motako aldagaiak estandarizatzen ditu, irizpide batzuen arabera, adibidez, zeinu ortografikoak nahiz puntuazioak ezabatu edo antzekotasun fonetikoa zein grafikoa duten karaktereak identifikatu.
emparejar	Antzekotasun fonetiko edo grafiko handirik eduki ez arren, karaktere batzuk esangura berarekin erabil daitezke batzuetan. Beraz, makroak karakter bakarrean bereizten diren konfigurazioak aztertu eta jatorri berekoak ote diren ebaluatzen du.
comunes_y_no_comunes	Konfigurazio bakoitza batu beharreko bi artxiboen multzoan zein maiztasunarekin agertzen den ikertzen du. Ikerketa horren ostean, maiztasun horren arabera balizko komunak eta ez-komunak bereizten dira. Balioak konfigurazio berekotzat hartuko dira, baldin eta, lehen aipaturiko estandarizazio pausoak eman ondoren, kodigo berean gertatu badira.
buscar_errores	Maiztasun baxuko balioak (ez-komunak) ohiko balioren batetik (balio komunak) sortutako konfigurazio akastunak direla suposatzen da. Horrenbestez, distantziaren batez bestekoa ezarri eta balio komun batekin eta balio ez-komun batekin osaturiko balio bikoteak hautatzen dira, eta horiek distantzia nahiko laburra eduki beharko dute, konfigurazio beretik etorri ahal izateko.
confección_listas	Aurreko makroan ateratako bikote guztien artean, akats probabilitateen eta maiztasunen irizpideak aintzat hartuta jatorrizko konfiguraziotik datozen izen edo abizen bikoteei dagozkiela jotzeko zantzurik izanez gero.
indicador_sexo	Izen motako aldagaia estandarizatu ondoren, balio bakoitza gizon batena edo emakume batena den idetifikatzen du makro honek.

paso3_asignar_estandarizados	Gizonen nahiz emakumeen zerrendan agertzen diren konfigurazioak ikertu eta erabakia hartzen du hori taula batean edo bestean sartzeari buruz, konfigurazio horrek bi zerrendetan dituen maiztasunak aintzat hartuta. Lan hori egin ondoren, gizonen izena, emakumeen izena eta abizenak aldagaien konfigurazio bakoitzari esleituriko kodigoak hasten ditu berriro. Bi serie kode egongo dira, bata abizenentzat eta bestea bai gizonen izenentzat eta bai emakumeen izenentzat, baina ordena horrexetan, hau da, kode zenbaki jakina egongo da, aldagai batean bilduko dena, eta horretarako, kode hori baino txikiagoak diren kode guztiak gizonen izenenak izango dira, eta handiagoak diren guztiak emakumeen izenena.
est_final	paso3_asignar_estandarizados makrotik deitzen den makro laguntzailea; kodeak berriro hasten ditu, horiek zenbaki naturalen elkarren atzeko jarraipena izateko.
paso4_cálculo_pesos	Metodologian azaldutako m eta u koefizienteak kalkulatzen dira.
paso5_blocking	Erabiltzaileak ezarritako *blocking* irizpide guztiak banan-banan exekutatzen dira. *Blocking* prozedura horietatik berreskuraturiko koefiziente guztiak bildu eta histograma marrazten da (5etik gorako koefizienteekin, histograma hori argiago ikusteko), eta, bertan oinarrituz, mugako aldagaiaren balioa erabakiko da.
blocking_fecha_nacimiento	Erregistroen *blocking*a egiten du, jaiotze-dataren aldagaiaren doitasunerako.
blocking_clave	Erregistroen *blocking*a egiten du, gakoaren aldagaiaren doitasunerako.
blocking_codigo_texto1	Erregistroen *blocking*a egiten du, lehenengo letrarekin eta hurrengo bi kontsonanteekin osaturiko karaktere kateen doitasunerako.
blocking_codigo_texto2	Erregistroen *blocking*a egiten du, abizen bakoitzaren lehenengo hiru letrekin osaturiko karaktere-kateen doitasunerako.
fusión	*Blocking*eko irizpide bati dagozkion makroetako bakoitzetik deituriko makro laguntzailea da; m eta u koefizienteen kalkulua egiten du, aldagai bakoitzaren konfigurazio guztietarako.

paso6_posibles_links	*Match* gisa hartzen diren erregistro bikoteak aukeratzen dira. Lehenengo eta behin, ezarritako mugatik gorako koefizientea dutenak hartzen dira eta, gero, azterketa apur bat zehatzagoa egiten da, mugatik hurbileko baina hortik beherako koefizienteak dituzten bikoteetariko batzuk aukeratzeko.
asignación_lineal	Jatorri berekoak izan daitezkeen erregistro bikoteen talde bati dagokionez, esleipen unibokoak hartzen ditu, aukera bat baino gehiago dagoenean koefizienterik handienekoa hartzeko.
iml	Esleipen linealeko prozedura egiten du (lsap), 1-1 esleipenerako.
comparador_strings	Jaroren karaktere kateak konparatzekoa kalkulatzen du, bi konfigurazioren arteko distantzia neurtzeko.

.

Kapitulua
4

Aplikazioak

Ondoren, EUSTATen egin diren batzearen aplikazio batzuk azaltzen dira[2], probabilitate-batzeko metodoak erabiltzen dituztenak. Lehenengoak kasu bereziak dira; horietan, probabilitate-metodoaren xehetasunen bat sartu zen. Emaitzen arabera, eta gaiari buruzko literatura aintzat hartuta, aurreko kapituluan azaldutako batze-programa garatu zen. Gainerako aplikazioak, izan ere, programa hori exekutatzean lorturiko emaitzen ondoriozkoak dira.

Ezkontzen Estatistika eta Biztanleriaren Estatistika Erregistroa

Erregistroak batzeko metodologien ezarkuntzan egindako lehenengo proben helburua (Euskal Estatistika Erakundean probabilitate-tekniken bidez egindako probak), izan ere, ezkontzen titularren datu pertsonalen informazioa jasotzen duen fitxategia eta Oinarrizko Biztanleria Unitateko datuak dituena (Biztanleriaren Estatistika Erregistroa) biltzea izan zen.

Ordura arteko batzeek batze-ehuneko txikia lortzen zuten, halako fitxategian ez baitago NAN bezalako identifikatzaile bakarrik edo identifikaziorako bestelako gakorik.

Lehenengo eta behin, Oracle datu-basetik datuen bi fitxategiak hartzen dira; batean, 1996ko urtarriletik aurrerako ezkontzen erregistroak daude, eta bestean, berriz, Oinarrizko Biztanleria Unitate bakoitzeko erregistro guztien datuak.

Erabiliko diren batze-aldagaiak zehaztu behar dira. Horretarako, aldagai horien formatuak ikusten dira. Erraz egiazta daiteke aldagai batzuk ez direla inondik inora ere erabilgarriak, bikoteak sortzeko prozesurako, aintzat hartuta horietariko asko informazio gutxirekin agertzen direla edo fitxategi bakoitzean modu desberdinean erregistratzen direla. Beste kasu batzuetan, adibidez, sexuaren aldagaiaren kasuan, aurretiazko birkodifikazioa behar da, aldagai hori prozesuan erabili ahal izateko, fitxategi batean “H” nahiz “V” eta bestean “6” zein “1” agertzen baita.

Kasu honetan, batze-aldagaitzat erabiliko diren aldagaiak honako hauek dira:

Izena
Lehenengo abizena
Bigarren abizena
Sexua
Jaiotze-data

[2] Testu honetan aipaturiko batzearen aplikazio guztiak EUSTATen eremuaren barruan egin dira, horren garapenean inplikaturiko langile guztiengan eragina duen sekretu estatistikoaren babesean (Euskal Autonomia Erkidegoko Estatistikari buruzko 4/1986 Legean araututa).

Guztira, ezkontzen 64.384 erregistro zegoen eta horietatik, lehen aipaturiko prozedurak erabiliz, 30.802 batu ziren, hau da, erregistroen %47,84.

Lehenengo eta behin, batze-aldagai guztietan zuzeneko adostasuna zuten erregistroak SAS prozedura baten bidez batu ziren, nekezago justifikatzeko moduko adostasuna zuten erregistroei probabilitate-batzea aplikatzeko. Ahalegin horretan batu ez ziren erregistroak, probabilitate-batzeko programatik igarotzen dira, *blocking* irizpide desberdinak erabiliz. Lehenengo irizpidea jaiotze-data izan zen eta bigarrena, berriz, izenaren, lehenengo abizenaren eta bigarren abizenaren hasierako letrak. Guztira, 60.055 erregistro batu ziren, hau da, jatorrizko fitxategiaren %93,28.

Merkataritzako Erregistroa eta Jarduera Ekonomikoen Gidazerrenda

Probabilitate-batzeko prozedurak aplikatzeko beste kasuetariko bat, izatez ere, bi fitxategi batzeko programa orokorraren egokitzapena izan zen, Merkataritzako Erregistroko enpresak Eustaten Jarduera Ekonomikoen Gidazerrendarekin batzeko. Batze horren helburua hauxe izan zen: Merkataritzako Erregistroan jasotako informazio ekonomikoa Jarduera Ekonomikoen Gidazerrenda gaurkotzeko iturritzat erabiltzea, eragiketaren estaldura handitzeko, Jarduera Ekonomikoen Gidazerrenda informazio hori emateko ondorengo eremua baita.

Lan hori egiteko lehenengo pausoa Merkataritzako Erregistroko fitxategi bakarra lortzea izan zen. Hori egitura bereko 6 fitxategiren batuketa da (lurralde historiko bakoitzeko 2). Lurralde bakoitzean, fitxategi batek jatorrizko digitala dauka eta bestea, berriz, eskaneatuta eta karaktereak ezagututa (OCR) lortzen da. Ondoriozko fitxategitik ezabatu egiten dira bikoizturiko erregistroak (berriena kontserbatuz).

Hurrengo prozesuan, Jarduera Ekonomikoen Gidazerrendarekin gurutzatu behar da, IFK/NANa eremutik. Horren ondorioz, gidazerrendarekin zuzeneko adostasunik gabeko erregistroak dituen fitxategia lortzen da. Zuzeneko adostasunik gabeko erregistroak dituen fitxategi horri aplikatu behar zaio batze-prozesua, IFK/NAN eta izena estandarizatu ondoren. Azterturiko guztien artean, irizpide batzuk betetzen dituzten eta gidazerrendarekiko adostasuna duten erregistro berriak berreskuratzen dira.

Batzerako aldagaiak honako hauek dira:

IFK/NANa
Izena
Posta-kodea
Telefonoa.

Batze desberdinak egiten dira, honako aldagai hauek kontuan hartuta:

- IFK/NANa, izena, posta-kodea eta telefonoa
- IFK/NANa eta izena
- Izena
- IFK
- Telefonoa
- Posta-kodea

IKFren arabera eta telefonoz egindako batzeen kasuan, izenetan nolabaiteko antzekotasuna egotea ere eskatzen zen, ez baitzen nahikoa IFK estandarizatuaren eta telefonoaren balioak bat etortzea. Antzekotasun hori kalkulatzeko, berariaz diseinaturiko SAS makro bat exekutatzen zen eta horrek erlazioko lau kasu hartzen zituen kontuan:

1. Konprimaturiko izenak bat datoz.

2. Jaroren karaktere kateak konparatzekoa 0.85etik gorakoa da.

3. Enpresako izenetariko bat beste izenaren laburdura da, edo izen horren siglak.

4. Bi izenek berdin dituzten letren kopurua balio jakina baino handiagoa da.

Makroak, irizpide horiek betetzen zituzten izendun erregistro bikoteak hautatzeaz gain, establezimenduen erregistro bikote bakoitzerako lau irizpide horietariko zein betetzen zen ere adierazten zuen.

Ustiapen honetarako orain arte azaldutakoa ez da probabilitate-metodoei buruzkoa. Horiek azken batzean egiten ziren, posta-kodearen batzean. Jakina, bi enpresa posta-kode bera edukitzeagatik batzeak ez dauka inolako logikarik, posta-kode bereko enpresa asko baitago. Kasu honetan, enpresen izenen artean nolabaiteko antzekotasuna dagoela egiaztatzeaz gain (lehengo batzeetan bezala), koefiziente jakina esleitzen zaio erregistro bikote bakoitzari, izenaren maiztasunaren arabera.

Hau da, posta-kode berekoak izateaz gain izenetan ere erlazioa duten erregistroak hautatzen dira, lehen adierazitako irizpideren bat erabiliz. Bikote horietariko bakoitzari koefiziente bat esleitzen zaio. Koefiziente hori zuzenean erlazionaturik dago bi izenen hitz berdinak bi fitxategietan agertzeko maiztasunarekin.

Azkenik, honako hau betetzen duten erregistro bikoteak batzen dira:

(1) IFK berbera (estandarizatua).

(2) Izen bera (estandarizatua).

(3) Telefono bera eta izenen arteko erlazioa.

(4) Koefizientea aldagaia balio jakin bat baino handiagoa da.

(5) Izenetariko batean agerturiko hitz guztiak beste izenean agertzen diren.

Emaitzak honako hauek izan ziren: hasierako 30.000 enpresa-erregistroetatik 22.500 batu ziren, prozedura zehatzaren bidez. Batu barik gelditu ziren 7.500 erregistroak, berriz, azaldutako prozeduraren bidez ikertu eta horietako 900 inguru erlazionatu ziren. Hau da, guztira, batze-ehunekoa %75etik %78ra handitu zen, prozedura horri esker.

Aipagarria da bi etapetariko ezeinetan batu ezin izan ziren erregistroetariko asko nekazaritza sektorekoak izan zirela eta, horien ezaugarri bereziak aintzat hartuta, ezin direla ezein metodoren bidez batu.

Errenta Pertsonalaren eta Familia-Errentaren Estatistika

EUSTATen beste esperientzia batean ere probabilitate-batzeari loturiko prozesuetarikoren bat erabili zen, eta esperientzia hori Errenta Pertsonalaren eta Familia-Errentaren Estatistikari buruz egin zen; estatistika hori 2001erako egindakoa da.

Errenta Pertsonalaren eta Familia-Errentaren Estatistika, egin ere, Foru Ogasunetatik jasotako fitxategien eta Biztanleriaren Estatistika Erregistroaren artean egiten da.

Batze hori iturri desberdineko eremuak konparatzeko prozesutzat hartzen da; gainera, balio berdinak aurkitzeko eta doitasun bakoitzeko batze-koefiziente bat esleitzeko ere baliozkoa da; horrela, gero, koefiziente osoa kalkulatzen da eta, horren arabera, kasuan kasuko Oinarrizko Biztanleria Unitatearen gakoa esleitzen da (Biztanleriaren Estatistika Erregistroari dagokionez). Hau da, batze zehatzaren prozesua aplikatzen da, koefiziente jakinak aintzat hartuta.

Aplikazio honen berritasuna, izan ere, batze-prozedura baino lehen *Izena eta abizenak* aldagaiaren homogeneizazio tratamendua egitea da, garaturiko probabilitate-metodoetan oinarrituz.

Ondoren, batze-prozeduran ezinbestekoa den *Izena eta abizenak* aldagaiaren homogeneizazio tratamenduak azaltzen dira.

Foru ogasunetako hiru fitxategi jasotzen dira, erregistro diseinu desberdina dutenak, eta aldagaien arteko homogeneizazio tratamendua egin behar da. *Izena eta abizenak* eremuan hiru azpieremuko eskemara egokitu behar da: izena, lehenengo abizena eta bigarren abizena, asteriskoekin bereizita.

Jarraian, aldagai hori tratatzean aurkituriko arazoak azaltzen dira.

- Abizen konposatuetarikoren bat duten zenbait erregistrotan *Izena eta abizenak* eremua akastuna da. Lehenengo abizena konposatua izanez gero, abizen horren azken partikula bigarren abizen bihurtzen da, eta azken hori izentzat doa, benetako izenetik asteriskoz bereizi barik. Bigarren abizena konposatua bada, horren azken partikula izenaren eremura igarotzen da, benetako izenetik asteriskoz bereizi gabe.

- Beste erregistro batzuk izena, lehenengo abizena eta bigarren abizena (espazio zuriz bereizita) eskemara egokitzen dira.

Homogeneizaziorako oinarritzat, honako hauetan erregistraturiko pertsonen izenetan eta abizenetan oinarrituz sorturiko datu-baseak erabiltzen dira: Biztanleriaren eta Etxebizitzen 2001eko Zentsuan eta Biztanleria Jardueraren Arabera sailkatzeko Inkestan.

Fitxategi horietako izenen eta abizenen datuak hartu, horietariko bakoitza bi fitxategietan agertzeko maiztasuna kalkulatu eta, maiztasun horren arabera, honako datu-base hauek sortu dira:

Datu-basearen izena	Azalpena
listado_final aldagaiak: *variable* *resultado*	Biztanleria eta Etxebizitzen 2001eko Zentsuko eta Biztanleria Jardueraren Arabera sailkatzeko Inkestako pertsonen izenen eta abizenen datuetan oinarrituz sortzen da. Izena eta abizenak eremuetako osagai bakoitza hartu eta, horiek agertzeko maiztasunaren arabera, balio bat esleitu zaie resultado eremuan, hori ondokoetarikoren bat izango da. 1 = Izena 2 = Maiztasunik txikieneko izena 3 = Abizena 4 = Maiztasunik txikieneko abizena 5 = Izena eta abizena
datos_2 aldagaiak: *variable* *var1* *var2* *resultado* *n_datos*	Hau ere Biztanleria eta Etxebizitzen 2001eko Zentsuko eta Biztanleria Jardueraren Arabera sailkatzeko Inkestako pertsonen izenen eta abizenen datuetan oinarrituz sortzen da. 2 osagaiko izenei edota abizenei dagozkien datuak aurkitzeko sortu da. var1 eta var2 osagaiak izen edota abizen konposatuko osagaiak dira; n_datos 2 balio du, bi osagai direla adierazteko, eta ondoriozko eremuak balioetariko bat hartzen du: 1 = Izen konposatua 3 = Abizen konposatua
datos_3 aldagaiak: *variable* *var1* *var2* *var3* *resultado* *n_datos*	Aurreko datu-basearen antzekoa, baina 3 osagaiko izenei eta abizenei dagozkien datuekin. var1, var2 eta var3 eremuak osagai horiek izango dira, n_datos 3 balio du, eta ondoriozko eremuak balioetariko bat hartzen du: 1 = Izen konposatua 3 = Abizen konposatua
datos_4 aldagaiak: *variable* *var1* *var2* *var3* *var4* *resultado* *n_datos*	Aurreko datu-basearen antzekoa, baina 4 osagaiko izenei eta abizenei dagozkien datuekin. var1, var2, var3 eta var4 eremuak osagai horiek izango dira, n_datos 4 balio du, eta ondoriozko eremuak balioetariko bat hartzen du: 1 = Izen konposatua 3 = Abizen konposatua

Sarrerako fitxategia Ogasuneko fitxategia da *antes* (*lehen)*, *variable* (*aldagaia)* eta *después* (*gero)* eremuekin. Tratamendua *variable* eremuari soilik aplikatzen zaio eta hori izenei eta abizenei dagokie.

Lehenengo eta behin, aldaketa txiki batzuk egiten dira izenak eta abizenak estandarizatzeko (kendu egiten dira azentuak, dieresiak, ikurrak, zenbakiak eta elkarren atzeko letra errepikakorrak, RR eta LL izan ezik, eta izenen eta abizenen uzkurdura batzuk itzultzen dira, adibidez, M-ren ordez MARIA edo PZ-ren ordez PEREZ).

Aldaketa horien helburua Ogasuneko fitxategietako izenak eta abizenak honako hauetan lorturikoekin errazago identifikatzea da: Biztanleriaren eta Etxebizitzen 2001eko Zentsuan eta Biztanleria Jardueraren Arabera sailkatzeko Inkestan. Izan ere, aldaketa horiek azken horiei ere egin zaizkie.

Ondoren, Ogasuneko fitxategietako *Izena eta abizenak* eremua hitzez hitz banantzen da. Izen eta abizen konposatu batzuk lotzen dituzten hitzak ezabatzen dira (adibidez, DE, LA, LOS, SAN,…). Hitzen taldeak hartu eta konparatu egiten dira lehen azaldutako listado_final, datos_2, datos_3 eta datos_4 datu-baseetan bildutako izen eta abizen konposatuekin. Doitasunik aurkituz gero, *variable* aldagaiaren eta *n_datos* aldagaiaren zenbakia biltzen da.

Aintzat hartu behar da konbinazio bat baino gehiago aurkitu ahal dela, konposatuak aztertzeko orduan (esate baterako, PEDRO JOSE RUIZ OTALORA honelaxe deskonposatu ahal da: PEDRO*JOSE*RUIZ-OTALORA edo PEDRO-JOSE*RUIZ*OTALORA). Horrenbestez, erregistro horiek nabarmendu egiten dira, markagailutzat jarduten duen aldagai baten bidez, eta aldagai horrek aditzera emango du erregistro hori zehatzago aztertu beharreko kasu berezia dela.

Kode egokiak esleitzen dira, Ogasuneko erregistroetako izenak eta abizenak bilatzean lorturiko emaitzaren arabera (erreferentziako datu-baseekin). Konposatuetan batzuetan erabiltzen diren loturak, lehen ezabatu direnak, horiek ere kodifikatu egiten dira, baina ez zenbakien bidez, ikurren bidez baizik.

Hona hemen esleipenak:

Partikula	Kodea
DE, DEL, Y	&
1 luzerako edozein hitz Y izan ezik	+
DA, DI, DO, SAN, DOS, SA, SANTA, SANTO, SAO, BON	*
LA, LAS, EL, LOS	$

Azkenik, markagailu aldagai baten bidez berezitzat hartu eta datu-base batean banatu diren (zehatzago aztertzeko) erregistroak ikuskatzen dira. Horien balioen arabera, mantendu beharrekoari buruzko erabakiak hartzen dira.

Helburua *_asignados (_esleituak)* eta *_no_asignados (_ez-esleituak)* fitxategiak *Izena eta abizenak* eremuko azken balioekin lortzea da.

Izenak eta abizenak homogeneizatzeko tratamendua egin ondoren, fitxategien batzea egiten da; kasu honetan, batze determinista izango da.

Koefizienteen esleipenean honako aldagai hauek hartzen dute parte (PFEZn eragina dutenak):

- IZENA (izenaren alfagakoa, 1. abizenaren alfagakoa eta 2. abizenaren alfagakoa).

- HELBIDEA (Lurraldea, Udalerria, Barrutia, Atala, Erakundea, Kalea, Zenbakia, Solairua eta Eskua).

- NANa (8, 7, 6, 5 eta 4 digitu).

- JAIOTZE DATA (Eguna, Hilabetea eta Urtea).

Oinarrizko Biztanleria Unitatearen gakoa koefiziente guztiak zenbatu ondoren esleituko da; horrela, funtzioak ezabatu egingo ditu, esleituriko koefizientearen arabera (domeinu multzoa aintzat hartuta), koefizienterik txikieneko erregistroak, eta gero bitan agertzen direnak. Horrela, koefizienterik handiena duen eta, aldi berean, baxan ez dagoen Oinarrizko Biztanleria Unitatea aukeratuko da. Azaldutako homogeneizazio prozesuaren eta batze deterministako prozeduraren ostean, erregistroen %98 baino gehiago batzen dira.

Biztanleriaren eta Etxebizitzen Zentsua eta Biztanleria Jardueraren Arabera sailkatzeko Inkesta

Batze-teknika hauek EAEko Biztanleriaren eta Etxebizitzen 2001eko Zentsuaren balioztatzeari egindako aplikaziotzat ere erabili dira, Biztanleria Jardueraren Arabera sailkatzeko Inkestan lorturiko datuetan oinarrituz. Eragiketa hau hiruhilekoan behin egiten da eta helburua estatistika-informazio jarraitua planteatzea da, EAEko talde nagusien bolumenari eta horien ezaugarriei buruz, jarduera ekonomiko desberdinetan duten partaidetza kontuan hartuta. Horrenbestez, batze-prozesua aplikatzeko bi fitxategi hartzen dira oinarritzat. Bata, EAEko Biztanleriaren eta Etxebizitzen 2001eko Zentsua (2.082.587 erregistro), eta bestea, Biztanleria Jardueraren Arabera sailkatzeko Inkesta (Zentsuaren erreferentziako datatik hurbilen dagoen hiruhilekoari buruzkoa, eta 10.831 erregistro ditu).

Bi fitxategien eremu komunak eta batze-aldagaitzat erabiltzen direnak honako hauek dira:

Izena
Lehenengo abizena
Bigarren abizena
Sexua
Etxebizitzaren identifikatzailea
Jaiotze-data

Blocking-fitxategitzat honako hauek erabiltzen dira:

Jaiotze-data
Etxebizitzaren identifikatzailea
Testu-kodea, bi abizenen lehenengo letrak eta hurrengo bi kontsonanteek osatua
Beste kode bat, bi abizenen lehenengo hiru letrek osatua.

Batze-prozesua aplikatu ondoren, Biztanleriaren eta Etxebizitzen Zentsuko fitxategian 10.535 erregistro aurkitu ziren, Biztanleria Jardueraren Arabera sailkatzeko Inkestako 10.831 erregistroetatik. Hortaz, batze-ehunekoa %97,27 da.

Vitoria-Gasteizko Biztanleriaren Udal Errolda eta Biztanleriaren Estatistika Erregistroa

Vitoria-Gasteizko Biztanleen Udal Erroldako eta Biztanleriaren Estatistika Erregistroko fitxategien arteko batzea ere egin da. Fitxategi horietariko bakoitzaren erregistro kopurua hauxe da:

Vitoria-Gasteizko Biztanleen Udal Errolda	233.670
Biztanleriaren Estatistika Erregistroa	2.456.831

Helburua Vitoria-Gasteizko Biztanleriaren Udal Erroldan erregistraturiko pertsona guztiak Biztanleriaren Estatistika Erregistroan aurkitzea da.

Erabilitako batze-aldagaiak honako hauek izan dira:

Lehenengo abizena
Bigarren abizena
Izena
Jaiotze-data
Sexua

Eta *blocking*-irizpidetzat honako hauek erabili dira:

Jaiotze-data
1. testu-kodea
2. testu-kodea

Lehenengo hurbiltzean honako emaitza hauek lortu ziren:

Baturiko erregistroen bikote kopurua	231.646
Vitoria-Gasteizko Biztanleen Udal Erroldako batu gabeko erregistro kopurua	2.357
Biztanleriaren Estatistika Erregistroko batu gabeko erregistro kopurua	2.225.188

Horrenbestez, baturiko erregistro kopurua fitxategi bakoitzean batu gabe gelditu direnekin batuz gero, jatorrizko fitxategien erregistro kopuru osoa lortu beharko genuke. Hala eta guztiz ere, zenbakiei erreparatuta ikusten denez ez da halakorik gertatzen.

Izan ere, Vitoria-Gasteizko Biztanleen Udal Erroldako erregistro batzuk Biztanleriaren Estatistika Erregistroko erregistro batekin baino gehiagorekin batu dira, horrek bikoiztuak dituelako.

Alderantziz egitean ere gauza bera gertatzen da, baina neurri txikiagoan. Hau da, Vitoria-Gasteizko Biztanleen Udal Erroldako erregistro batzuk Biztanleriaren Estatistika Erregistroko erregistro berarekin batu dira.

Guztira, Vitoria-Gasteizko Biztanleen Udal Erroldako 231.313 erregistro desberdin batu dira Biztanleriaren Estatistika Erregistroko erregistroren batekin. Hortaz, batze-ehunekoa %98,99 da.

"Aurrekariak" atalean azaldu dugunez, lehenago ere fitxategi horien batzea egin da, probabilitatekoak ez diren teknikak erabiliz. Batze horren emaitzak edukitzean, bi metodoen bidez lorturiko emaitzak konparatzeko erabakia hartu zen.

Vitoria-Gasteizko Biztanleen Udal Erroldako 635 erregistro batu dira Biztanleriaren Estatistika Erregistroko erregistro desberdin batekin, batze-metodo bakoitzean. Erregistro horiek eskuz ikuskatu eta baturiko bi erregistroak konparatu dira. Oro har, 5 erregistrotik 4tan egokiagoa da probabilitate-metodoaren bidez baturiko erregistroa. Eta metodo determinista 25 erregistrotik 1ean soilik dabil hobeto. Gainerako erregistroetan, edo ez da irizpiderik aurkitu bi metodoetatik hobea zein den erabakitzeko, edo biak batu dira pertsona berarekin, baina gako desberdinak erabiliz.

Bestalde, 535 erregistro probabilitate-batzearen bidez batu dira, baina lehengo prozeduraren bidez batu barik egon dira; era berean, lehenengo prozeduraren bidez baturiko 1.126 erregistro ez dira batu agiri honetan azaldutako programaren bidez.

Azken erregistro horiek aztertzean ikusi diren gauza batzuk ondoren adierazitakoak dira:

- Metodo deterministak arazoak sortzen ditu atzerriko pertsona-izenak batzeko orduan, kasu batzuetan izen bereko zenbait erregistro euren artean batzen baitira, abizen desberdinekoak izan arren. Izan ere, atzerriko pertsona askoren bigarren abizena ez dago erregistraturik eta metodo deterministak ezartzen du hutsik dauden bi eremuren arteko konparazioa positiboa dela, eta izena bera denez *match* bat sortzen da, lehenengo abizena berdina izan ez arren. Probabilitate-metodoaren bidez konpondu egiten da arazo hori, metodo horrek eremu guztien doitasuna edo antzekotasuna aztertzen baitu.

- Bi metodo horietan akatsak ikusi dira, neba-arrebak bereizteko orduan, batez ere bikien kasuan.

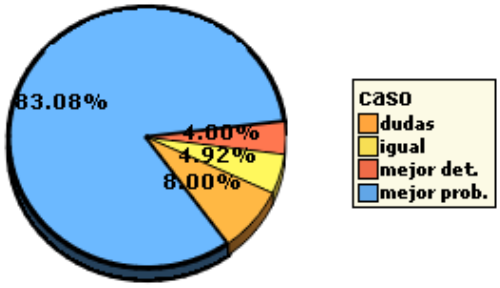
- Kasu batzuetan, pertsonaren abizenak tokiz aldatuta agertzen dira. Hau da, fitxategi batean lehenengo abizentzat agertzen dena bigarren abizena da beste fitxategian, eta alderantziz. Abizen truke horren ondorioz, probabilitate-metodoak ez ditu *match* gisa identifikatzen, *match* izan arren (metodo deterministak ondo identifikatzen ditu).

Azken ohar hori aintzat hartuta, egoera hori aztertzeko batze-programan kodea sartzea planteatu zen. Hori egin eta emaitzak berriro kalkulatu ziren. Bigarren hurbiltze honetan, emaitzak ondokoak izan ziren:

Baturiko erregistroen bikote kopurua	231.818
Vitoria-Gasteizko Biztanleen Udal Erroldako batu gabeko erregistro kopurua	2.185
Biztanleriaren Estatistika Erregistroko batu gabeko erregistro kopurua	2.225.016

Lehen bezala, Vitoria-Gasteizko Biztanleen Udal Erroldako erregistro batzuk Biztanleriaren Estatistika Erregistroko erregistro batekin baino gehiagorekin batzen ziren, eta beraz, batze-bikote bat baino gehiago sortzen zuten. Guztira Vitoria-Gasteizko Biztanleen Udal Erroldako 231.485 erregistro desberdin daude batuta, hau da, guztizkoaren %99.06. Aldaketa horren bidez, 172 erregistro gehiago batu dira.

Programaren bigarren exekuzioaren ostean, berriro ere modu desberdinean baturiko erregistroak bilatu ziren, aintzat hartuta erabilitako metodoa. Kasu honetan, kasu guztien %83an funtzionamendua hobea zen probabilitate-metodoan, eta %4an hobea da metodo deterministan; kasu guztien %5ean batzea berdina izan zen bietan eta, azkenik, kasuen %8ko hondarra gelditzen da, eta horietan ezin izan da erabakirik hartu.



Biztanleria Jardueraren Arabera sailkatzeko Inkesta eta Biztanleriaren Estatistika Erregistroa

Biztanleriaren Estatistika Erregistroarekin baturiko beste fitxategi bat Biztanleria Jardueraren Arabera sailkatzeko Inkesta izan da. Hala eta guztiz ere, kasu honetan helburua ez da Biztanleria Jardueraren Arabera sailkatzeko Inkestako erregistro guztiak aurkitzea, Oinarrizko Biztanleria Unitatearen gakoa ez zutenak aurkitzea baizik.

Biztanleria Jardueraren Arabera sailkatzeko Inkesta probabilitate-lagin jarraituan oinarritzen da, hau da, etengabe berriztatzen den etxebizitza-panel batean. Etxebizitza-lagina modu ausazkoan ateratzen da Etxebizitzen Gidazerrendatik, Lurralde Historikoaren arabera mailakatuta.

Etxebizitzen lagina lortu ondoren, Biztanleriaren Estatistika Erregistrotik hautaturiko etxebizitzetako pertsonen datuak batzen dira. Horrela, etxebizitza bakoitzeko pertsona guztien lagina ateratzean, guztiek dute Oinarrizko Biztanleria Unitatearen gakoa.

Hala ere, batzuetan, etxebizitza bati inkesta egitean, ikusten dira bertan bizi diren pertsonak ez direla laginean agertzen direnak (esate baterako, lehen bertan bizi zirenak beste bizileku batera joan direnean).

Kasu honetan, etxebizitzan orain bizi diren pertsonei egiten zaie inkesta. Horrenbestez, pertsona berri horiek ez dute beteta edukiko Oinarrizko Biztanleria Unitatearen gako eremua, eta gero Biztanleriaren Estatistika Erregistroan pertsona horiek aurkitu beharko dira, batze-programa erabiliz.

Biztanleria Jardueraren Arabera sailkatzeko Inkestako fitxategian 41.568 pertsona desberdin zeuden, eta horietatik %10ek baino apur bat gehiagok ez dute Oinarrizko Biztanleriaren Unitateko gakorik. Beraz, fitxategi hauek honako erregistro kopuru hauekin batzen dira:

Biztanleria Jardueraren Arabera sailkatzeko Inkesta, Oinarrizko Biztanleriaren Unitateko gako barik	4.117
Biztanleriaren Estatistikaren Erregistroa	2.456.831

Erabilitako batze-aldagaiak honako hauek izan dira:

Lehenengo abizena
Bigarren abizena
Izena
Jaiotze-data
Sexua

Eta blocking-irizpidetzat honako hauek erabili dira:

Jaiotze-data
1. testu-kodea
2. testu-kodea

Honako emaitza hauek lortu ziren:

Baturiko erregistro bikoteen kopurua	3.277
Biztanleria Jardueraren Arabera sailkatzeko Inkestako batu gabeko erregistro kopurua (Oinarrizko Biztanleria Unitatearen gako gabe)	855
Biztanleriaren Estatistika Erregistroko batu gabeko erregistro kopurua	2.453.556

Biztanleria Jardueraren Arabera sailkatzeko Inkestako 3.262 erregistrok ez zuten Oinarrizko Biztanleria Unitatearen gakorik, eta horiek probabilitate-batzearen bidez aurkitu dira. Kopuru hori erregistro guztien %79.23 da.

Batze-ehunekoa aintzat hartuta, badirudi emaitza ez dela Biztanleriaren Udal Erroldakoa (lehen azaldutakoa) bezain ona. Kasu honetan, gauza bat argitu behar da; hain zuzen ere, Biztanleria Jardueraren Arabera sailkatzeko Inkestan elkarrizketaturiko pertsonetariko batzuk ezin dira aurkitu Biztanleriaren Estatistika Erregistroan, agian bertan ez daudelako. Esaterako, etxebizitza batean aurkitu eta jatorrizko laginean agertzen ez diren pertsonak (ez dute Oinarrizko Biztanleria Unitatearen gakorik); Euskal Autonomia Erkidegotik kanpoko pertsonak dira eta, beraz, ez daude erregistraturik Biztanleriaren Estatistika Erregistroan.

Biztanleria Jardueraren Arabera sailkatzeko Inkestakoak izan eta Oinarrizko Biztanleria Unitatearen gakorik ez duten erregistroetatik batu ez direnak (Biztanleriaren Estatistika Erregistroan agertzen ez direlako) egiaztatzea lan oso zaila da. Dena den, erregistro horietatik zenbat ezin izan diren batu Biztanleriaren Estatistika Erregistroaren azken

gaurkotzea baino geroago jaio direlako egiaztatu ahal izan da. Kasu honetan, Biztanleria Jardueraren Arabera sailkatzeko Inkestako 187 erregistroren jaiotza-data Biztanleriaren Estatistika Erregistroan erregistraturiko azkena baino geroagokoa da.

Kapitulua
5

Ondorioak

Orain arte lorturiko emaitza guztiak aztertuta ondorioztatzen denez, probabilitate-batzearen bidezko metodoa oso erabilgarria da aldagai komuntzat erregistroaren identifikazio-gakorik ez duten fitxategiak batzeko.

Probabilitate-metodoak metodo deterministaren aldean duen lehenengo hobaria kasu zailetan eraginkortasun handiagoa izatea da, eta bestea, berriz, errazagoa eta azkarragoa da. Horrela, kasu batzuetan eskuzko tratamendua ezabatu egiten da eta, hortaz, kostuak murriztu egiten dira.

Probabilitate-metodoak gaitasun konputazional handia behar du, baina horrek gaur egun gero eta garrantzi txikiagoa dauka.

EUSTATek metodo horrekin jarraitu nahi du lanean, baina ez pertsonen arteko batzeak egiteko soilik; izan ere, enpresen fitxategiak batzeko ere erabili nahi du.

Enpresen fitxategiei dagokienez, pertsonak batzearen eta enpresak batzearen aldeko desberdintasun nagusia kasu bakoitzeko aldagai mota da. Enpresen erregistroetan, *Enpresaren izena* aldagaia ezin daiteke pertsonaren izena balitz bezala tratatu, kasuistika oso desberdina baita. Enpresen izenetan dibertsitate handiagoa dago eta, gainera, akatsak egoteko probabilitatea ere handiagoa da, konplexutasuna handiagoa izatearen ondorioz. Beste desberdintasun bat, izan ere, enpresen fitxategietan sarritan agertzen den aldagai bat helbidearena da, eta pertsonen kasuan hori ez da tratatu. Aldagai hori modu oso desberdinetan erregistraturik ager daiteke eta, berori estandarizatzeko, lan gehigarria egin behar da.

Bestalde, probabilitate-metodoetan erabilgarritasun handia ikusteak ez du esan nahi Erakundeak metodo deterministen erabilera guztiz baztertuko duenik. Izan ere, batzearen emaitzak, neurri handian, fitxategien kalitatearen araberakoak dira, eta beraz, kasu batzuetan egokiagoa izango da probabilitate-metodoak barik beste metodo batzuk erabiltzea. Gainera, fitxategi batzuen kasuan, onena metodo desberdinak konbinatzea izan daiteke.

Gaur egun, EUSTAT metodo horien guztien erabilera konputazionala optimizatzeko eta errazteko batze-modulu baten sorreran ari da lanean, kasu bakoitzean metodorik egokiena aplikatu ahal izateko.

Kapitulua
6

Bibliografia

[1] FELLEGI, IVAN P. AND SUNTER, ALAN B.

A theory of Record Linkage. Journal of the American Statistical Association, December, 64. lib., 328. zenb., 1183-1210. or. (1969).

[2] JARO, M.A.

Record Linkage research and the calibration of record linkage algorithms. U.S. Bureau of the Census (1984).

[3] JARO, M.A.

Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association.

[4] BLAKELY, T. AND SALMOND, C.

Probabilistic record linkage and a method to calculate the positive predictive value. Internation Journal of Epidemiology (2002).

[5] AYESTARÁN, MARINA AND LEGARRETA, LEIRE.

Applying methods of record linkage for census validation in the Basque Statistics Office. Euskal Estatistika Erakundea (2004).

[6] WINKLER, WILLIAM E.

Matching and Record Linkage. Bureau of the Census (1993).

[7] CHRISTEN, PETER AND CHURCHES, TIM.

Febrl – Freely extensible biomedical record linkage. Australian National University (2003).

[8] YANCEY, WILLIAM E.

An Adaptive String Comparator for Record Linkage. U.S. Bureau of the Census, Statistical Research Division (2004).