

**UNITATE EKONOMIKOEN ERREGISTRO ADMINISTRATIBOAK BATU
PROBABILITATE-TEKNIKAK ERABILITA**

Laura Otero Franco



**EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA**

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Faxa: 945 01 75 01
Helb. elek.: eustat@eustat.es
www.eustat.es

AURKEZPENA

Erregistro administratiboa unitate bati buruzko informazioa duen agiria da. Unitate hori pertsona, establezimendua edo bestelako erakunde bat izan daiteke, administrazioa zerbitzu batek bere xedeetarako informazioa biltzen du. Azken urteetan, erregistro administratiboen ustiaketan oinarritutako estatistikaren potentzialak erritmo handian egin du aurrera.

Erregistro administratiboak tresna indartsuak dira hainbat arlotako errealitatea neurtzeko eta aztertzeko. Erregistro administratiboak direktorioak egiteko eta mantentzeko erabiltzen dira, batez ere, estatistikan, eraginkorrak baitira eta koste arrazionalak baitituzte.

Hala ere, informazio administratiboak zenbait oztopo izaten ditu estatistikaren inguruan erabiltzeko. Adibidez, ohikoa da erregistro administratiboak aurreko aldietan abian jarritakoak izatea edo estatistika-eragiketa egin behar duen unitatearekin inolako zerikusirik ez duten unitateek jarri izana abian. Baina oztopo nagusia beste bat da (erregistro administratibo guztien berezko ezaugarri gisa): ez dira helburu estatistikoetarako sortu eta diseinatu. Beraz, erregistro hauek eskainitako informazioa estatistikaren alorrean erabili aurretik tratatu egin behar izaten da.

Aurrez egin beharreko tratamendu horietako bat erregistroak batzea edo lotzea da. Teknika hauei esker, erregistro administratibo desberdinetan dagoen unitate bat berari buruzko informazioa lotu egiten da.

AURKIBIDEA

AURKEZPENA	1
AURKIBIDEA.....	3
SARRERA	4
SARRERA ETA HELBURUAK	4
PROIEKTUAREN DESKRIPZIOA	4
AURREKARIAK	5
ERREGISTRO ADMINISTRATIBOAK	6
EZAUGARRIAK	6
ABANTAILAK ETA DESABANTAILAK	7
ENPRESEN ETA ESTATISTIKA EKONOMIKOARENTZAKO INTERESEKO BESTE UNITATEEN ERREGISTROAK.....	8
ERREGISTROAK BATZEA.....	11
METODOLOGIA	11
PROGRAMAZIOA	15
PROGRAMA OROKORRA	15
HASIERAKO PARAMETROAK	18
AZTERKETA	19
ESTANDARIZAZIOA ETA HOMOGENEIZAZIOA.	19
PROBABILITATEEN KALKULUA.	25
BLOCKINGA.	26
LINKAK.....	27
EMAITZEN AZTERKETA	29
FITXATEGIEN DESKRIPZIOA.....	29
FITXATEGIEN AZTERKETA	30
BATZEAREN EMAITZA	32
ONDORIOAK.....	33
BIBLIOGRAFIA.....	34

SARRERA

Sarrera eta helburuak

Erregistroak batzea edo lotzea unitate bat bera ordezkatzan duten bi erregistro administratibo desberdinetan (erregistro bakoitzeko elementu bat) elementu bikoteak aurkitzeko prozedura da. Bi erregistro administratiboak bat bera izanez gero, elementu duplikatuak aurkitu beharko lirateke.

Erregistroak batzeko bi teknika orokor daude:

- **Batze determinista:** Elementuak alor guztietan edo aurrez zehaztutako alor kopuruan bat datozenen batzen dira.
- **Probabilitate-batzea:** Elementu bikote bakoitzari probabilitate-pisu jakina ematen zaio eta behar adinako pisua duten bikoteak batutzat hartzen dira.

Koaderno tekniko honen xedea da, unitate ekonomikoei buruzko informazioa duten erregistro administratiboen baturaren azterketa aurkeztea, probabilitate-metodoak erabilia. Zehatz esateko, fitxategi hauen batura du ardatz: EUSTATEko Jarduera Ekonomikoen Direktorioa (JED) eta Gizarte Segurantzako fitxategia.

Proiektuaren deskripzioa

Proiektuak hiru fase ditu:

- I. Fellegi eta Sunterrek proposatutako probabilitatezko batze-metodologia aztertu eta egokitu, unitate ekonomikoen erregistro administratiboak tratatzeko.
- II. SASen aplikazio bat programatu, unitate ekonomikoen bi erregistro administratibo automatikoki batzeko.
- III. Lortutako emaitzak aztertu.

Proiektuko fase desberdinak aurkezteko, koaderno tekniko hau zenbait kapitulutan zatitu da:

1. kapitulan koaderno teknikoko helburuak, proiektuaren faseak eta erregistroak batzeko EUSTATen aurrekariak sartu dira.

2. kapituluaren erregistro administratibo kontzeptua sartu da eta horren ezaugarriak, abantailak eta desabantailak aurkeztu dira. Unitate ekonomikoen informazioa duten erregistro administratiboen kasuan bereizi egin da.

3. kapituluaren erregistroak batzeko metodologia deskribatu da, Fellegi eta Sunterrek 1969an “A theory for Record Linkage” artikuluan aurkeztutako probabilitate-teknika erabilia.

4. kapituluaren unitate ekonomikoen erregistro administratiboak batzeko programaren egitura aurkeztu da, EUSTATEk garatutako probabilitate-tekniken bidez.

5. kapituluaren azterketa sustatu zuten fitxategiekin (EUSTATEN Jarduera Ekonomikoen Direktorioa (JED) eta Gizarte Segurantzako fitxategia) batzeko programa erabilia lortu diren emaitzak aztertu dira.

Azkenik, 6. kapituluaren teknika hauen azterketaren eta garatutako batze programaren aplikazioaren ondorioak daude.

Aurrekariak

1996an berritu zen Biztanleen Udal Errolda azkeneko aldiz eta EUSTATEk erregistro administratiboak lantzen hasia erabaki zuen, Biztanleriaren Estatistika Erregistroa egiteko. Horretarako, urte bakoitzeko abenduaren 31n Udal Erroldetan eta Biztanleriaren Berezko Mugimenduetan (jaiotzak, heriotzak eta ezkontzak) zegoen informazioa erabili zen.

Fitxategi horiek tratatzeko, batzeko teknika deterministak ezartzen hasi ziren identifikatzeko aldagai komunak erabilia (izena, abizenak, NAN/AIZ, jaiotze-data, posta-helbidea, etab.). Ondoren, EUSTATEk sustatzen dituen ikerketa eta metodologia estatistiko-matematikoko beken markoan, Fellegi eta Sunterrek 1969an “*A theory for Record Linkage*” artikuluan aurkeztu zuten eredu teorikoan oinarritutako probabilitate-teknikak aztertzen hasi zen.

Prozedura determinista eta probabilitistiko horiek orokortu beharra zegoenez, batzeko aplikazio independente baten programazioa egin zen, Batzeko Moduluarena (BM).

Banakoaren erregistroak batzean izandako arrakastaren ondorioz, probabilitatezko batze-metodoen aplikazioa aztertzea erabaki zen, unitate ekonomikoen fitxategiak batzeko. Hasteko, banakoak batzeko programa egokitzea pentsatu zen, baina unitate ekonomikoen erregistroetako berezitasunen ondorioz hain gauza arrunta ez zela ikusi zen.

ERREGISTRO ADMINISTRATIBOAK

Erregistro administratiboak datu estatistikoetarako iturriak dira, inkestak eta erroldak bezala. Badu, ordea, beste bi iturriekin alderatuz gero, alde nabarmen bat: erregistro administratiboak erregistroaren jabe den zerbitzu administratiboak ezarritako arauzko irizpideen ondorioak dira, estatistikako irizpide metodologikoen orde.

Erabilera administratiborako hainbat erregistro dago administrazio publikoen esku eta estatistiketarako erabil daitezke. Erregistro horietan mota askotako datuak egoten dira; pertsonen, enpresei, erakundeei eta bestelako populazioei buruzko informazioa. Berez, informatzaileei berriro eskatu beharrik gabeko datuak dira eta, ondorioz, datuak biltzeko gastu ekonomikoak txikitu egiten dira.

Azken urteetan sarriago egin dituzte estatistika-erakundeek estatistikan erabiltzeko moduko erregistroen zerrendak eta urrats handiak eman dira datuak hartzen dituzten erakunde nagusiek informazioa emateko: zerga-administrazioak, hezkuntzak, osasun arloak eta gizarte-segurantzak. Erregistro Zibila eta beste erregistro batzuk aspalditik erabiltzen dira estatistiketarako.

Administrazio estatistikotik, batez ere Eustatetik, ahalegin txikiak egin dira legeetan dagozkien aldaketak eginda herri administrazioetako erregistroen funtzionamendu-araudia egokitu eta estatistikoki aprobetxatzeko.

Ezaugarriak

Jarraian erregistro administratiboen ezaugarrietako batzuk daude:

- 1) Arauzko dokumentuak dira eta ekintza administratiboak erregistratzen dituzte, hau da, ez dira estatistiketarako egindakoak.
- 2) Ez dute urrats estatistikoen logikarik, planifikatzeko, erregistratzeko, jarraipena egiteko, koordinatzeko edota administrazio kontrolerako baitira.
- 3) Informazioa hautemateko tresna esanguratsuak dituzte (txantiloak, galdesortak, txartelak, fitxak, akta koaderno edo liburuak, etab.) eta administrazioaren jarraipena eta kontrola egiteko dira.
- 4) Aldikotasuna dagokion legealdia ezar dezake edo ez. Adibidez, behin erregistratu daitezke (unibertsitateko tituluak, esaterako), aldirokoak izan daitezke (joan-etorri erregistroak eta lizentziak, adibidez) edo ausazkoak izan daitezke (osasun-erregistroak, esaterako).
- 5) Estaldura handikoak dira, erregistro administratiboaren erakunde jabea hala bada.

- 6) Erregistratzeko unitate batzuk definitzen dituzte: pertsonak, eraikinak edo establezimenduak, ekintzak (gertaerak edo jarduerak), zerbitzuak, baliabideak, merkatal transakzioak, etab. Ondorioz, aldagai ugari dago.
- 7) Erregistratutako unitateak eta neurtutako aldagaiak aldagai estatistikoetara bihurtu daitezke, nahiko metodologia errazak erabilita.
- 8) Formatu desberdinetako artxiboetan biltegitratzen dira: paperean, fitxategietan edo espedienteetan eta baita euskarri digitalean ere.

Abantailak eta desabantailak

Erregistro administratiboak beste iturriekin alderatuz gero, abantaila hauek dituzte:

- 1) Datuak ekoizteko gastu txikia. Gehienetan erregistro administratiboak erakunderen baten arauzko prozeduraren batengatik egiten dira, ekintza administratiboren bat kontrolatu beharra duelako; beraz, erakundeko bulegoetan hartzen da, normalean, informazioa eta horrela ez dago arloko eragileen gasturik.
- 2) Informatzaileek galdesortak betetzeko karga txikiagoa. Gehienetan inkesta nazionalak luzeak izaten dira, baina kontuan izan behar da badirela erregistro administratibo korapilatsuak ere, hala nola, zergen aitortpenak.
- 3) Gobernuko organoen artean ibili daiteke informazioa eta administrazio publikoak ez dauka ahalegin bikoitza egin beharrik. Teorian abantaila denak administrazio publikoko erakundeak sinkronizatzen behartzen ditu, datuak eskatzeko terminoetan. Prozesu hau oso konplexua da, erakunde bakoitzeko txantiloak independenteak baitira edo tresnak erabiltzen dituzten populazioak aldatu egin baitaitezke.
- 4) Populazio xedearen estaldura osoa lortzen du. Kasu askotan erregistro administratiboek estaldura osoa lortzen dute, adibidez, osasunean, justizian edo hezkuntzan. Baina biktimizazioa bezalako kasuetan erregistro administratiboek ez dute estaldurarik lortzen, kriminaltasuneko salaketa gutxi egon daitekeelako.
- 5) Erantzun ezen akatsak beste iturrietan baino txikiagoak dira, ez dago laginketako akatsik. Ez dago laginketako akatsik, ez baita laginik hartzen; horren ordez estaldurako akatsa dago.
- 6) Azpi-populazioetan desagregatu daiteke. Abantaila hau oso garrantzitsua da, erregistro administratiboak datu interesgarri ugari izan baitezake, eta horiekin azpi-populazioak lortu. Horregatik garrantzitsua da erregistro administratiboak aztertzea, egokitasuna ebaluatzeko.
- 7) Herrialde bateko lurralde eremu guztietako informazio-sistemak indartzen ditu. Argi dago Estatuak bere informazio-sistemak indartu nahi dituenean, abantaila hau aukera bat gehiago dela.

- 8) Informazioaren kalitatea hazi egiten da, intereseko gaiak eskatzen dituen zehaztasunekiko inprimakiak eraiki baitaitezke. Hala ere, inprimakiak aldatzea zaila da, aukera badagoen arren.
- 9) Datu-serieak eraikitzeke egiazko oinarria dira. Erregistro administratiboek prozesu administratiboaren historia eraman dezakete, eta horrek denborazko seriea eraikitzea indartzen du.

Bestalde, erregistro administratiboek desabantailak ere badituzte, esaterako:

- 1) Unitate administratiboen eta estatistikoen arteko korrespondentzia gabezia. Azterketako unitateak litekeena da bat ez etortzea, erregistro administratiboa banako batena izan baitaiteke eta ez etxe edo merkataritzako establezimendu batena.
- 2) Aldagaien definizioetan aldeak. Oro har txantiloiek eta formatuek ez dute aldagaiak modu operazionalen definitzeko metodologiarik edo identifikatzeko edo deskribatzeko aldagaiak izan daitezke, azterketa deskriptibo batek behar duen sakontasunik gabe.
- 3) Kode administratiboen eta estatistikoen arteko konbertsio falta.
- 4) Xede estatistikoarekin bat ez datozen datuak eta erreferentziazko aldiak. Hau gerta daiteke, erregistro administratiboen xedea ez baita estatistikoa, badira, ordea, erregistro administratiboak datu estatistiko bihurtzeko prozesuak.
- 5) Aldaketa politikoen eraginak, erregistro administratiboen jarraitutasunean. Txantiloiei eta prozedura administratiboak aldatu egin daitezke, gorabehera politikoekin. Alderdi hau oso garrantzitsua da eta erregistro administratiboaren ahultasuna jartzen du agerian.
- 6) Erregistroetan identifikatzaile komuna falta da datuak uztartzeko. Era berean, zeregin baterako langile egonkorak ere falta dira.
- 7) Epe luzerako ikuspegia falta da, sistema estatistikoa garatu eta koiunturari lehentasuna emateko.
- 8) Erregistroak ematen dituzten organoen artean ez dago kooperazio politikaririk, parte hartzen duten eragile eta erakundeek ez dute akordiorik eta lurraldeko desagregazioentzako lege estatistikorik ere ez.

Enpresen eta estatistika ekonomikoarentzako intereseko beste unitateen erregistroak

Ekoizpen ekonomikoko unitate estatistikoak hauteman eta eguneratzeko erregistrorik garrantzitsuenetakoa batzuk Jarduera Ekonomikoen gaineko Zerga (IAE), Gizarte Segurantzaren kotizatzeak unitateak eta Merkataritza Erregistroak dira.

IAE udal zerga da eta urtero kargatzen du enpresa jardueren eta jarduera profesional eta artistikoen ekitaldia. Bere kudeaketa zentralizatua Foru Ogasunek egiten dute gure kasuan eta lurralde partekatuan Zerga Administrazioako Espainiako Agentziak. Jarduera

bakoitzak lizentzia behar du toki batean, beraz, subjektu pasibo batek hainbat lizentzia aurkez ditzake helbide batean edo gehiagotan. Subjektua eta helbidea identifikatzeko beste datu batzuen artean, lizentzian egoten dira alta eta bajako datak, kodea edo jardueraren epigrafea (garai bateko CNAE-74an oinarritua) eta jarduera deskribatzeko literal bat.

IAEn enpresen, pertsona fisiko eta juridikoen, profesionalen eta artisten ia merkataritza jarduera guztiak sartzen dira (gure kasuan EAEko lurraldeen eremuan). Balio Erantsiko Zergaren (BEZ) subjektu guztiek IAEko matrikulan aitortu beharra daukate.

Baina egungo araudiak zergak ordaintzetik salbuesten ditu fakturazio maila batez azpitik dauden pertsona fisiko eta juridiko guztiak. Honek ondorio zuzen bat izan du: subjektu hauek jarduera utzi arren, ez dute baja ematen eta errealitate ekonomikoaren irudia puztutakoa izaten da.

Gainerakoan, IAEk nekazariei, abeltzainei eta lehen sektoreko jardueri buruzko informaziorik ez du ematen, ezta administrazio publikoen eta irabaz asmorik gabeko erakundeen ez merkataritzako jardueri buruzkorik ere. Foru Ogasunek ematen dituzten fitxategiek ez dute jarduerarako baimen nazionala duten enpresa handien daturik ere sartzen, horiekin akordio bereziak baitituzte, euskal udalei dagokien zergaren zatiaren banaketa proportzionalerako.

Enpresa eta unitate juridikoen erregistroetarako intereseko Gizarte Segurantzaren erregistroetan daude Kotizazio-kontuaren Kodeak (CCC), langile autonomoen araubide bereziko (RETA) erregistroak eta nekazaritzako erregimen berezikoak (REA, galzorian RETAn sartu delako) eta itsas langileen erregimen berezia (REA). Gaur egun hiruhileroko fitxategiak daude data jakin bateko altako unitateekin, gure kasuan gure autonomia erkidegoko eremu geografikora mugatuta beti.

CCC horiek enpresari pertsona fisikoaren edo juridikoaren funtsezko datuak ditu, identifikazio fiskala eta izena eta enpresariak kontratatutako soldatapekoen kotizazio-kontu bakoitzaren datuak. Enpresari bakoitzak CCC bat edo gehiago izan ditzake toki jakin batean, Gizarte Segurantzaren kotizatze mota desberdinetan kontratatuta baldin baditu. CCC bakoitzak kotizazio-unitateko helbidearen datuak ditu, jardueraren EJSN kodea, alta eta egoera datak, lan harreman mota eta CCC hori duen langile kopurua.

RETAeko erregistroetan erregistro horretako pertsonen datu deskriptiboak daude, baita jardueraren helbidearenak eta, gutxitan, EJSN kodea ere bai. Ez ditu enpresariak eta beste langileak bereizten, bazkide nagusiak, kooperatibistak, mendeko autonomoak eta bestelakoak, Gizarte Segurantzaren erregimen honetan sarrarazten baititu, erregimen orokorraren orde.

Beraz, ez dugu enpresa sozietario ugariri buruzko gizarte segurantzatik datorren daturik, bereziki sozietate mugatuei buruzkorik, horietako langileak (bi edo hiru gehienetan) jabeak edo akziodun nagusiak ere badiren arren eta RETAen kotizatzen dute erregimen orokorreko CCCn kotizatu orde.

Gizarte Segurantzako datuak hobeto eguneratuta egoten dira, enpresariak isunak izaten baitituzte aitopenak eta kotizazioen ordainketak atzeratzeagatik edo jarduera utzitakoan kotizazio unitateei baja ez emateagatik. Irekita egon baina zaharkituta dauden kotizazio-unitateak jabeek iruzurrean utzitako elkartearenak ziaten dira ia erabat eta erregistroko agintariak ofiziozko baja ematen diete.

Merkataritza Erregistroak dira enpresei eta merkataritza trafikoan interesa duten gainerakoei segurtasun juridikoa eskaintzeko tresna nagusiak. Erregistratu eta gertakari batzuk aitortu beharra dute merkataritza elkarte guztiek, eta borondatez egin dezakete hori pertsona fisikoek eta merkataritza-jarduerak garatzen dituzten beste pertsona ez juridikoek. Gainera, elkarrekin beraien kontuak erregistroetan utzi behar izaten dituzte, jendeak horien berri izan dezan. Urteroko depositu hori falta bada, isuna jar dakiokete elkarteari.

MEko fitxategiei esker, merkataritza elkarteei buruzko oinarritzko datuak ezagutu ditzakegu, beraien helbide soziala gure autonomia erkidegoko hiru probintzietako batean baldin badago: Erregistro identifikazioko zenbakia eta identifikazio fiskalekoa, izena, egoitzaren helbidea, xedea eta interes txikiagoko beste datu batzuk. Aldi berean, kontuen gordailu horiek inkesta eta estatistika ekonomikoentzako interes handia dute.

Gizarte Segurantzarekin bezala, MErekin ere “baztertutako” elkarrekin eduki ditzakegu, jabeek likidatu gabekoak edo, aktiboan egon arren, hainbat urtez konturik utzi ez dutenak.

Aipaturiko hiru iturri administratiboen kasuan (IAE, GZ eta ME) bakoitzaren aldagaien edukiak oso desberdinak dira, bakoitzak bere kasuistikaren, gabezien, akatsen... ondorioak jasaten ditu, zerbitzu administratiboak oso desberdinak eta geografian oso sakabanatuta egotearen ondorioz.

ERREGISTROAK BATZEA

Ikusi denez, erregistro administratiboak ez dira berez estatistika iturriak, erroldak eta inkestak diren bezala, izan ere, batetik, xede administratiboa dute eta, bestetik, arauzko kontrolerako balio dute, hau da, erakunde bati buruzko gertaera bat edo banako ekintza bat erregistratzen dute eta zuzenean erasaten dio.

Gainera, eragiketa estatistikoak ikerketaren helburuekin bat datozen definizioak eta sailkapenak behar ditu, aldiz, erregistro administratiboek ez dute zertan bat etorri alderdi metodologiko hauekin.

Erregistro administratiboak estatistikarako iturri bezala erabiltzeko, erregistroak batu beharra dago. Teknika honen bidez iturri administratibo desberdinetako informazioa egoki erabili daiteke estatistikan. Kapitulua honetan EUSTATEN batzeko probabilitate-teknikak garatzeko erabili den metodologia deskribatzen da.

Metodologia

Batze automatikoko programa egiteko, Fellegi eta Sunterrek 1969an "A theory for record linkage" artikuluan aurkeztutako eredu teorikoa erabili da. Eredu honen oinarriak hauek dira:

Eredu teorikoa

Batu beharreko erregistro administratiboak A eta B izenak ematen zaizkie eta a eta b izango dira erregistro administratiboetako kide generikoak, hurrenez hurren.

Bi artxibo horiek elementu komunak izango dituzte, eta, beraz, batzearen helburua sor daitezkeen $A \times B$ bikote guztien artean unitate ekonomiko berari buruzkoak zein diren jakitea da. Hau da, helburua multzoa banatzea da

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

batuta ez dauden multzoen batasunean

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

eta

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

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

Aztertutako biztanleriaren unitate bakoitzak ezaugarri jakinak ditu lotuta, esaterako, izena, helbidea, enplegua, etab. Unitate ekonomiko berari buruzko kideak identifikatu behar dira. Hala eta guztiz ere, erregistro administratiboak sortzeko prozesuak akatsak eta zehaztugabetasunak sar ditzake (kodifikazio, transkripzio eta teklatze akatsak, aldaketa tipografikoak edo fonetikoak, datuen galera, etab.), sortutako elementuetan. Akats horien ondorioz, A-ren eta B-ren bi kidek, unitate ekonomiko berari buruzkoak ez direnek, erregistro berdin-berdinak sor ditzakete, eta, sarriago, A-ren eta B-ren bi kide berdin-berdinek elementu desberdinak sor ditzakete. A-ren eta B-ren kideei dagozkien elementuak $\alpha(a)$ eta $\beta(b)$ dira, hurrenez hurren.

Bi erregistro administratiboko elementuak bikoteka jartzeko lehenengo urratsa horiek konparatzea da. Konparazioaren emaitza kode multzoa da, eta horiek ondorengoak bezalako baieztapenetan kodifikatzen dira: “izena bat dator bi elementuetan”, “izena bat dator eta Almacenes Garrido da”, “izena ez dator bat”, “elementuetako batek ez du izenik” edo “izenaren zati bat berdina da, baina osoa ez”. Formaren aldetik, **konparazio bektorea** $\alpha(a)$ eta $\beta(b)$ erregistroen funtzioko bektorea da:

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

Ikusten denez, γ funtzioa definitzeko oinarria $A \times B$ da. Idatzi daiteke $\gamma(a, b)$, $\gamma(\alpha, \beta)$ edo soilik γ . **Konparazio-espazioa** esaten zaio γ -ren balizko errealizazioen multzoari eta Γ erabiltzen da izendatzeko.

Batze eragiketaren prozesuan ikusten da $\gamma(a, b)$ eta hau erabaki behar izaten da:

- (a, b) bikotea den, $(a, b) \in M$ (erabaki hau **link** izaten da eta A_1 deitzen zaio)
- (a, b)
- ez-bikotea den, $(a, b) \in U$ (erabaki hau **ez-link** izaten da eta A_3 deitzen zaio) Baina litekeena da egoera batzuetan bi erabaki hauetako bat ezin hartu izana berariazko akats maila batzuentzat, beraz, hirugarren erabaki bat har daiteke, A_2 deitzen da eta **balizko link** esaten zaio.

Baldintza hauetan L **batze-araua** definitzen da Γ konparazio-espazioa $D = \{d(\gamma)\}$ ausazko erabaki-funtzioen multzoan aplikatzea bezala, eta formula horretan:

$$d(\gamma) = \{P(A_1 | \gamma), P(A_2 | \gamma), P(A_3 | \gamma)\}; \quad \gamma \in \Gamma$$

eta

$$\sum_{i=1}^3 P(A_i | \gamma) = 1$$

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. Onartzen da elementu bikote bat $[\alpha(a), \beta(b)]$ ausaz hautatzea, probabilitate-prozesu baten

arabera konparatzeko. Ondoriozko konparazio-bektorea $\gamma[\alpha(a), \beta(b)]$ ausazko aldagaia da beraz. γ -ren **balizko probabilitatea esaten zaio** $(a, b) \in M$ $m(\gamma)$ baita eta izango da:

$$m(\gamma) = P\{\gamma[\alpha(a), \beta(b)] \mid (a, b) \in M\} = \sum_{(a,b) \in M} P\{\gamma[\alpha(a), \beta(b)]\} \cdot P[(a, b) \mid M]$$

Era berean, γ -ren **balizko probabilitatea esaten zaio** $(a, b) \in U$ bider $u(\gamma)$ baita. Horrenbestez,

$$u(\gamma) = P\{\gamma[\alpha(a), \beta(b)] \mid (a, b) \in U\} = \sum_{(a,b) \in U} P\{\gamma[\alpha(a), \beta(b)]\} \cdot P[(a, b) \mid U]$$

Batze-arau honi loturiko bi akats mota daude. Lehenengoa, *matchekin* bat ez datozen elementu bikoteak konparatzean horiek *link* gisa hartzen direnean gertatzen da, eta horren probabilitatea hau da:

$$P(A_1 \mid U) = \sum_{\gamma \in \Gamma} u(\gamma) \cdot P(A_1 \mid \gamma)$$

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

$$P(A_3 \mid M) = \sum_{\gamma \in \Gamma} m(\gamma) \cdot P(A_3 \mid \gamma)$$

Γ espazioko batze-araua μ, λ ($0 < \mu < 1, 0 < \lambda < 1$) **akats mailako batze-araua** dela esaten da eta honela izendatzen da $L(\mu, \lambda, \Gamma)$ baldin eta

$$P(A_1 \mid U) = \mu \quad y \quad P(A_3 \mid M) = \lambda$$

Esaten da $L(\mu, \lambda, \Gamma)$ batze-araua hobe zina dela, baldin eta

$$P(A_2 \mid L) \leq P(A_2 \mid L')$$

edozein $L'(\mu, \lambda, \Gamma')$ deritzonerako mantentzen bada, lehengo erlazioak egiaztatzen dituzten batze-arau guztien artean.

Ikusten denez, definizio honen arabera, erabakitze arau hobe zina maximizatu egiten ditu konparazio positiboak hartzeko aukerak (hau da, A_1 edo A_3 erabakiak), akats maila finkoekin. Zentzuzko erabakia dela dirudi, kontuan hartuta A_2 erabakia hartzeko eskuzko batze-eragiketak egin behar direla eta horiek kostu handia dutela. Gainera, A_2 aukera txikia ez bada, batze-prozesuaren erabilgarritasuna zalantzazkoa izango dela dirudi.

Egileek batze-arau hobe zina proposatzen dute (μ, λ) akats-mailetan, eta honela litzateke:

$$d(\gamma) = \begin{cases} (1,0,0) & \text{si } T_\mu \leq \frac{m(\gamma)}{u(\gamma)} \\ (0,1,0) & \text{si } T_\lambda < \frac{m(\gamma)}{u(\gamma)} < T_\mu \\ (0,0,1) & \text{si } \frac{m(\gamma)}{u(\gamma)} \leq T_\lambda \end{cases}$$

$$\text{non } T_\mu = \frac{m(\gamma_n)}{u(\gamma_n)}, \quad T_\lambda = \frac{m(\gamma_{n'})}{u(\gamma_{n'})} \text{ eta } n, n' \text{ bi oso, non } 0 < n \leq n' < N_\Gamma.$$

Aplikazio askotan, akats maila nahiko altuak onartu ahal lirateke, A_2 ekintzaren aukera ezabatzeko. Kasu honetan n eta n' hartzen dira kontuan edo T_μ eta T_λ , eta γ -ren batez besteko multzoa aurreko forman hutsa litzateke. Hau da, (a,b) bikote bakoitza M -n edo U -n aurkitzen da. Are gehiago, EUSTATen hala erabaki da, batzeko programa automatikoa garatzerakoan, beraz, muga bakarra ezartzen da $T_\mu = T_\lambda$.

Batze-arau hobezina eraikitzeko zehaztasun gehiago, $m(\gamma)$ eta $u(\gamma)$ pisuen kalkulua eta eredu teorikoaren beste xehetasun batzuk ikusteko, Fellegi eta Sunterren aipaturiko artikulua [\[1\]](#) eta "Erregistroak batzeko metodo automatikoak eta EUSTATen duten erabilera" [\[2\]](#) koaderno teknikoa kontsulta daitezke.

PROGRAMAZIOA

Aurreko kapituluaz azaldutako metodologia erabili eta SAS programa bat egin da, unitate ekonomikoen bi fitxategiak batzeko. Programa bi fitxategi zehatz batzeko da, batez ere, eta bi horiek EUSTATen Jarduera Ekonomikoen Direktoria (JED) eta Gizarte Segurantzako fitxategia dira.

Bi fitxategien berezitasuna dela-eta, ezin izan da edozein bi fitxategi batzeko programa generikorik egin, baizik eta bi fitxategi zehatzentzako *ad hoc* programatu da. Baina batzeko programa modularra eta sekuentziazkoa da, beraz, programak deituriko makro asko fitxategi desberdinekin etorkizunean egingo diren batzeetan erabil daitezke.¹ Kapitulu honetan batzeko programaren egitura deskribatzen da. Hurrengo kapituluaz azterketaren xede diren fitxategiak azpimarratuko dira, baita horietan batzea ezartzeko behar diren prozedura laguntzaileak ere.

Programa orokorra

Batze programak SAS fitxategi hauek ditu:

- Programa nagusia, erabiltzaileak programa burutzeko beharrezko argudioak definitzen dituena.
- Programa nagusiak deitzeko makroak dituen programa, makro horiek egiten dituzten batzeko etapak.

Batzeko berezko makroak ditu aurreko atalean deskribatutako Fellegi eta Sunter eredu teorikoan oinarrituta, baina horiez gain blocking prozedurarako makroak ere baditu. Prozedura hau ez da ezinbestekoa teorian, baina bai praktikan. Erregistro administratibo bateko elementu guztiak beste erregistro administratibo bateko elementu guztiekin konparatzea ezin da hartu maila konputazionalan, beraz, beharrezkoa da blocking irizpideren bat erabiltzea batu daitezkeen elementu bikoteen azpimultzoak aukeratzeko eta horrela elementu guztien konparazio masiboak saihesteko.

Fitxategi hauez gain, programa laguntzaile bat sortu da, batze programa zuzen egiteko behar diren kanpoko bi taula eraikitzen dituena. Taula hauek eraikitzen ditu:

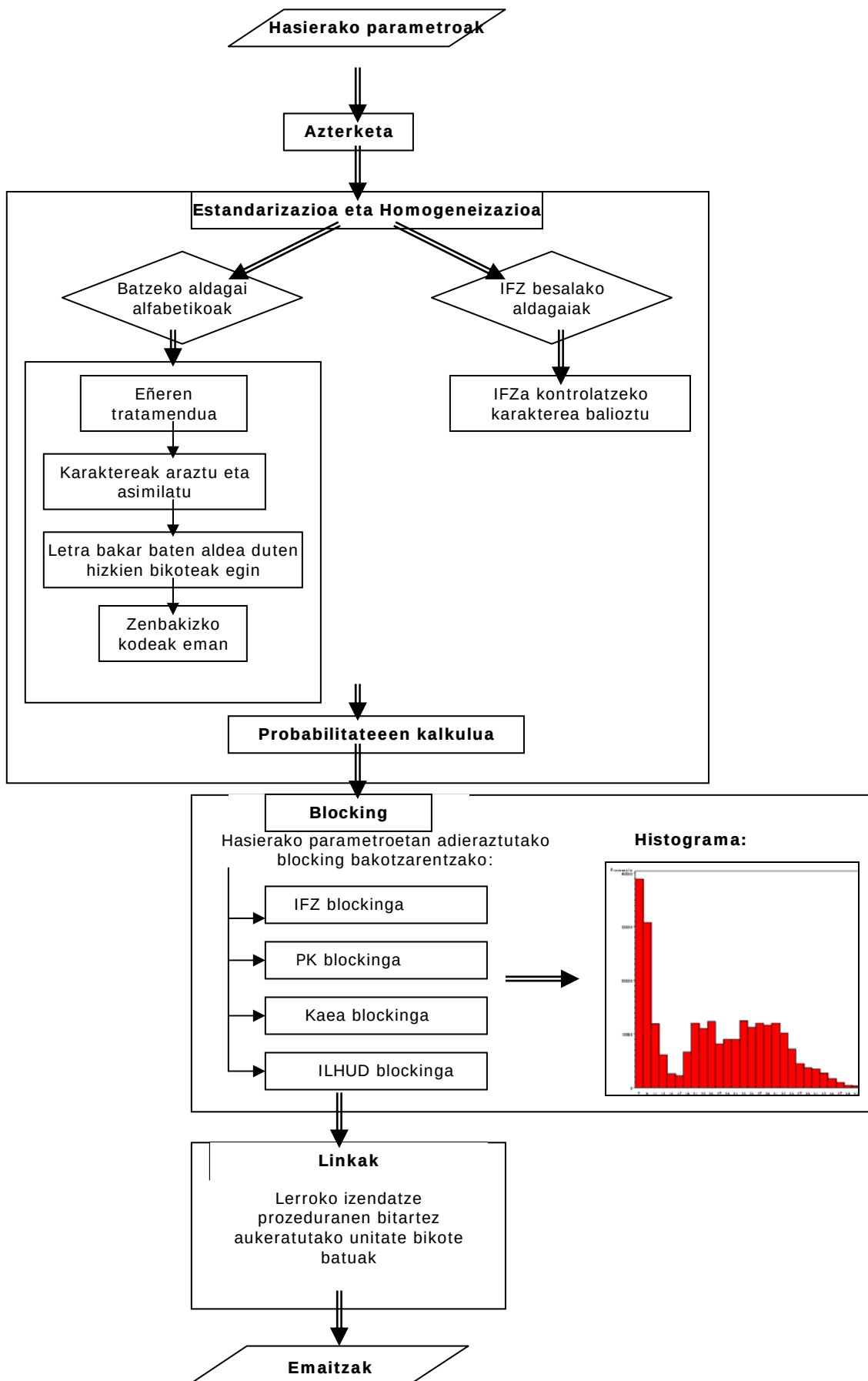
- Partikula batzuk (preposizioak, artikulak, etab.) dituen datu-multzoa. Partikula horiek estandarizazio eta homogeneizazio urratsean egiten den batze alfabetikoen aldagaietatik kentzen dira.
- Izen sozialaren sigla batzuk dituen datu-multzoa. Horiek estandarizazio eta homogeneizazio urratsean egiten den batze alfabetikoen aldagaietatik kentzen dira.

Programa laguntzaile hau batze programa egin aurretik behin bada ere egin behar da, programa erabilitakoan taulak egon daitezkeen.

¹ Proiektu honetan garatutako SAS makroak interesatuta egon eta ekoizpen estatistikoan erregistro administratiboak batzen jarduten direnen esku daude.

Taula hauek erregistro kopuru jakin batekin eraiki dira, baina komenigarria da taulei erregistroak gehitzea, emaitzak lortu ahala ahalik eta partikula eta izen sozialaren sigla gehien eduki dezaten eta ondorioz, estandarizazio eta homogeneizazio urratsa ahalik eta eraginkorrena izan dadin.

Jarraian batze programaren algoritmoa grafikoki adierazten duen fluxu-diagrama bat dago, eta ondoren deskribatzen da:



Hasierako parametroak

Batze programak ongi funtzionatzeko, parametro batzuk hasi behar izaten dira programa burutzen hasi aurretik. Erabiltzaileak sartu beharreko input datuak daude jarraian labur-labur:

- Batzeko aldagaiak adierazi behar dira, hau da, bi fitxategietako bakoitzean aldagaiak duen izena eta zer aldagai mota den. Batzean erabil daitezkeen balizko aldagai tipoak eta dagozkien kodeak dira hauek:

KODEA	BATZEKO ALDAGAIA
0 TIPOA	Erregistroa IDENTIFIKATZEKO GAKOA
1 TIPOA	IZENA
2 TIPOA	KALEA
3 TIPOA	UDALERRIA
4 TIPOA	PROBINTZIA
5 TIPOA	TELEFONOA
6 TIPOA	POSTA-KODEA
7 TIPOA	IFZ

Esan beharra dago 0 TIPOA (Erregistroa IDENTIFIKATZEKO GAKOA) ez dela batzeko aldagai bat, baizik eta elementu bakoitza bere erregistro administratiboaren barruan modu unibokoan identifikatzeko kodea. Hala ere, ezinbesteko aldagaia da, batutako elementu bikoteak dagozkien gako bidez identifikatuko baitira erregistro administratiboetan.

- Gutxienez blockingeko irizpide bat adierazi beharra dago. Aurrez esan den bezala, blockinga berez batzeko teknika ez den arren, ezinbestekoa da neurri bateko fitxategiekin aritzerakoan. Jarraian daude erabil daitezkeen balizko blocking irizpideak eta dagozkien kodea.

KODEA	BLOCKING MOTA	DESKRIPZIOA
1. irizpidea	IFZ blockinga	IFZrekin bat etortzea
2. irizpidea	PK blockinga	Posta-kodea edo kodeak (bat baino gehiago badaude) bat etortzea
3. irizpidea	Kalea blockinga	Kalearen edo kaleen (bat baino gehiago badaude) hitzez hitzeko kodea bat etortzea
4. irizpidea	ILHUD Blockinga	Enpresaren izenaren lehenengo hitzaren, lurralde historikoaren eta enpresaren udalerraren kodea bat etortzea

- Haslerako parametro teoriko batzuk ezarri beharra dago (ikus bibliografia [1]):

e_A, e_B	Fitxategi bakoitzean batzeko aldagaien balioa gaizki erregistratu izanaren aukera. Onartu egiten da gaizki erregistratzeko probabilitatea independentea dela balio bakoitzetik.
e_T	Batzeko aldagaiaren balioa bi artxiboetan modu desberdinean agertzeko probabilitatea, bietan ondo erregistraturik egon arren. Esate baterako, bi artxiboak une desberdinetan sortuak izan eta unitate ekonomikoak izena aldatu badu.

Uste da e_A eta e_B parametroen balioak nahikoa direla, bi sarrera berdin, nahiz eta okerrak izan, bat etortzeko probabilitatea oso txikia izateko.

- Une jakin batean programak balio bat eskatzen dio erabiltzaileari, unitate ekonomikoen bikoteak zein pisutatik aurrera dauden bat eginda zehazten duen mugarako.

Azterketa

Programaren lehenengo urratsa aurrez aitortutako batzeko aldagaien eta gako mota aldagaiaren azterketa da.

Horretarako, zenbait egiaztapen egiten dira, hala nola, gutxienez batzeko aldagai bat eta blockingeko irizpide bat ezartzea.

Akatsen bat egonez gero edo daturen bat zuzen sartu ez bada, programa amaitu egiten da eta pantailako *log*ean mezu bat agertzen da, egindako akatsa identifikatuaz; erabiltzaileak konpondu behar du akats hori.

Aldiz, beharrezko informazio guztia sartu bada, programa egiten jarraitzen da.

Estandarizazioa eta homogeneizazioa.

Estandarizazio eta homogeneizazio lanak oso garrantzitsuak dira batze prozesuetan, aldagaiak aurrez ongi tratatuz gero batutako bikote kopurua asko hazi baitaiteke eta, are garrantzitsuagoa, kalitate handiagoko bikoteak batuko dira, hau da, unitate berekoak izateko segurtasun handiagoa dago.

Batzeko bi aldagai mota estandarizatu behar dira: batzeko aldagai alfabetikoak (unitate ekonomikoaren izena eta kalea) eta IFZ. Mota hauetako bakoitzak bere estandarizazioa du. Batzeko aldagai mota hauetako bakoitzaren estandarizazio eta homogeneizazio prozesuak deskribatuko dira jarraian.

Batzeko aldagai alfabetikoak.

Batzeko aldagai alfabetikoak estandarizatzeko eta homogeneizatzeko, zenbait etapa daude. Azken helburua bi alor berri lortzea da: *est_var** eta *cod_var**, estandarizatu beharreko batzeko aldagaia *var** izango dela suposatuta.

*est_var** alor berriak jatorrizko *var** batze aldagaiaren literala estandarizatuta dauka, aldiz, *cod_var** alor berriak zenbakizko kode bat du eta horrek identifikatzen du batzeko aldagaiaren balioa. Horrela, bada, unitate ekonomiko edo kale izen multzo batzuk jatorrian desberdinak izan arren, modu berdinean estandarizatu dira eta *cod_var** kode berak adierazten ditu.

Jarraian daude labur-labur deskribatuta batzeko programan garatzen diren estandarizazio eta homogeneousazio tratamendu desberdinak:

▪ ***Eñeren tratamendua***

Testu editoreetan edo programetan karaktereak kodifikazio desberdinekin sortu ahal izan direnez erregistro administratiboetan, ohikoa da tileta duten bokalak edo Ñ gaizki kodifikatuta agertzea. Makro honek Ñ ordeztu /, # eta ¥ sinboloak agertzean zuzentzen du.

Makroak /, # eta ¥ sinboloak dituzten hitz guztiak aztertzen ditu eta Ñ letraren akats tipografikoak izan ote daitezkeen aztertzen du. Kontu handia izan behar da, zenbait kasutan /, # eta ¥ sinboloak ez baitira Ñ letraren ordeztuak.

Akats hauek saihesteko, /, # eta ¥ sinboloen ordeztu jatorrizko hitzean Ñ jarrita beste edozein elementutan dagoela ikusi behar da.

▪ ***Karaktereak araztu eta asimilatu***

Etaparen honetan ahalik eta gehien homogeneousatu nahi izaten dira batzeko balizko aldagaiak, balio berari dagozkion bi literal ez daitezkeen desberdin agertu akats tipografikoengatik.

Horretarako, arazketa lan hauek egiten dira:

- Puntuazioko zeinuak kentzen dira, hala nola, puntua, koma edo gidoia.
- Batze alfabetikoko aldagaia enpresaren edo establezimenduaren izena bada, gizarte siglaren karaktereak kendu egiten dira.
- Bokaletako tiletak kendu egiten dira.
- Alfamerikoak ez diren karaktereak kendu egiten dira: parentesiak, izartxoak, traolak, etab.
- Batze alfabetikoko aldagaiari informazio garrantzitsurik eskaintzen ez dioten artikulak eta preposizioak kendu egiten dira.
- Grafia bakarrarekin hartzen dira karaktere edo karaktere multzo hauek, antzekotasun fonetiko edo grafikoaren ondorioz akatsa sor badezakete.

Grafia berarekin ordezkatzeko diren karaktereak edo karaktere taldeak		
Y	I	
TX, TS, CH	TZ	
K	C	Ondoren A, O, U bokalak baditu
K	QU	Ondoren E, I bokalak baditu
N	M	Ondoren B, P kontsonanteak baditu
V	B	
Ñ	N	
GU	G	
Z	C	Ondoren E, I bokalak baditu, Z aurretik T ez dagoenean

▪ **Letra bakar baten aldea duten hitzekin bikoteak egin**

Fase honetan letra baten aldea besterik ez duten hitzen bikoteak aztertzen dira, eta aldatzen den letra hori zehatza izaten da. Taula honetan daude programan aztertzen diren letra bikoteak.

C	K	K	C	K	QU	QU	K
C	TZ	TZ	C	L	LL	LL	L
C	Z	Z	C	M	N	N	M
C	X	X	C	Q	QU	QU	Q
C	Q	Q	C	R	RR	RR	R
C	QU	QU	C	S	X	X	S
G	J	J	G	S	TZ	TZ	S
I	J	J	I	S	Z	Z	S
I	LL	LL	I	TZ	X	X	TZ
J	X	X	J	TZ	Z	Z	TZ
K	Q	Q	K	X	Z	Z	X

Hitz batean letra baten ordezkari dagokion bikotea jarri eta fitxategi multzoan dagoen beste hitz batekin bat etorri gero, berdinak direla esaten da eta, ondorioz, zenbakizko kode bera ematen zaie.

- **Zenbakizko kodeak eman**

Batze alfabetikoko aldagaiko hitz guztien zerrenda bat edukitzen da, bere estandarizazioarekin eta zenbakizko kodearekin tratatzen ari dena. Azkeneko fase honetan batze alfabetikoko aldagaiaren jatorrizko balioei batu beharreko fitxategi bakoitzean estandarizatuak eta dagozkien zenbakizko kodeak ematen zaizkie.

IFZ

Estandarizazio fase honetan Espainian erabiltzen diren kode fiskal guztiak behar bezala sailkatzen eta baliozkotzen dira. 9 karaktere alfanumerikoko alor bat aztertzen da eta aztertutako kode mota bakoitzerako zenbakizko balio bat itzultzen da eta balio positibo guztiek (zero baino handiagoek) kode fiskala zuzena dela adierazten dute.

Kode fiskala zuzena dela esaten da, kontroleko karakterea dagokiona denean. Kontroleko karakterea zifren edo letren funtzio bat da eta identifikazio fiskala eta, noizik eta behin, dituzten kokapenak osatzen dituzte. Horiek sartuta egin nahi izaten dena da, identifikazio horren transkripzio eta digitazio akatsak saihestu. Beraz, kodeari dagokion kontrol-karakterea kalkulatzeko da eta bat al datorren ikusten da.

Jarraian daude IFZ kontrolatzeko karaktereak baliozkotzeko programatutako funtzioak itzuli ditzakeen balioak:

Mota	Ezezaguna	IFZ	IFK	AIZ	Behin behineko IFZ
Zuzena:		1	2	3	4
Okerra:	0	-1	-2	-3	

Makro honek espainiar legeen zehaztapen hauek betetzen dituzte:

2423/1975 Dekretua, irailaren 25ekoa

338/1990 Errege Dekretua, martxoaren 9koa

1624/1992 Errege Dekretua, abenduaren 29koa, 338/1990 aldatzen duena

155/1996 Errege Dekretua, otsailaren 2koa

1998ko uztailaren 3ko Agindua, 2423/1975 Dekretuko Eranskina aldatzen duena

1065/2007 Errege Dekretua, uztailaren 27koa

EHA/451/2008 Agindua, 2008ko otsailaren 20koa

INT/2058/2008 Agindua, 2008ko uztailaren 14koa

Jarraian dago kode fiskal bakoitza baliozkotzeko erabiltzen den prozedura:

- **Pertsona fisikoen IFZ**

Oro har, espainiar pertsona fisikoentzako identifikazio fiskalerako erabiltzen den zenbakia nortasun agiri nazionalakoa izaten da, aldiz atzerritarren kasuan atzerritarren identifikazio zenbakia (AIZ) erabiltzen da; Barne Ministerioak ematen ditu biak.

Atzerriko pertsona fisikoen kasuan, identifikazioa X, Y edo Z letrekin hasten da.

Gainera, 14 urtez azpiko espainiarrek NAN ez badute edo atzerrian bizi badira eta AIZ ez badute, identifikatzeko gako bat ematen die Ogasunak eta K, L edo M letrarekin hasten da, kasuaren arabera.

NAN zenbakiak zortzi zifra izaten ditu gehienez, aldiz, gainerako espainiar edo atzerritar pertsona fisikoenak zazpi izaten ditu gehienez (hasierako K, L, M, X, Y edo Z kendu ondoren). Kasu guztietan kontroleko karakterea letra bat izaten da beti.

Taula honetan dago pertsona fisikoen IFZ desberdinen deskripzioa:

Mota	Formatua	Iruzkina
NAN	Zortzi zenbaki + kontroleko digitua	Barne Ministerioak emandako nortasun agiri nazionala duten espainiarrak.
IFZ K	K + 7 zenbaki + kontroleko digitua	14 urtez azpiko espainiarrak
IFZ L	L + 7 zenbaki + kontroleko digitua	NAN gabeko atzerrian bizi diren espainiarrak
IFZ M	M + 7 zenbaki + kontroleko digitua	AIZ ez duten atzerritarrei Zerga Agentziak ematen dien IFZ
IFZ X	X + 7 zenbaki + kontroleko digitua	Poliziak identifikatutako atzerritarrak, 2008ko uztailaren 15era arte emandako atzerritarren identifikazio zenbakia dutenak
IFZ Y	Y + 7 zenbaki + kontroleko digitua	Poliziak identifikatutako atzerritarrak, 2008ko uztailaren 16az geroztik emandako AIZ dutenak (INT/2058/2008 Agindua, BOE uztailaren 15a)
IFZ Z	Z + 7 zenbaki + kontroleko digitua	Poliziak identifikatutako atzerritarren AIZ zifretan Y agortzen denerako gordetako zenbakia.

Kontroleko karakterea nola kalkulatzeko den deskribatzen da jarraian:

NAN

NANeko 8 zenbakiez osatuta zenbakia zati 23 eginda lortzen den emaitza hartzen da eta taula honen arabera dagokion letra ematen zaio:

0	T	8	P	16	Q
---	---	---	---	----	---

1	R	9	D	17	V
2	W	10	X	18	H
3	A	11	B	19	L
4	G	12	N	20	C
5	M	13	J	21	K
6	Y	14	Z	22	E
7	F	15	S	23	T

IFZ X, Y, Z

X ordez 0 jartzen da, Y ordez 1 eta Z ordez 2 eta NAN estandarrekin egiten den gauza bera egiten da.

IFZ K, L, M

Bere kontrol karakterea, pertsona juridiko baten IFZ balitz bezala kalkulatzen da (aurrerago deskribatua).

• **Pertsona juridiko eta, oro har, erakundeen IFZ**

Identifikazio Fiskaleko Zenbakia Ogasunak ematen die pertsona juridikoei eta nortasun juridikorik gabeko erakundeei -merkataritza elkarteei, erakundeei, taldeei, etab.- eta bederatzi karaktere izaten ditu. Bederatzigarrena kontroleko karakterea izaten da (digitu edo letra bat).

Lehenengo karakterea forma juridikoari dagokio eta A, B, C, D, E, F, G, H, J, N, U, V eta W izan daiteke merkataritza elkarte eta erakundeen kasuan eta P, Q, R eta S kongregazio erlijioso eta administrazio publikoko erakunde eta organismoen kasuan.

Taula honetan daude lehenengo letrak har ditzakeen balioak, bere izaera juridikoaren arabera.

Letra	Izaera juridikoa	Kontroleko karakterea
A	Sozietate anonimoak	Zenbakizkoa
B	Erantzukizun mugatuko sozietateak	Zenbakizkoa
C	Sozietate kolektiboak	Zenbakizkoa
D	Sozietate komanditarioak	Zenbakizkoa
E	Ondasun erkidegoak eta jaso gabeko jarauntsiak	Zenbakizkoa
F	Sozietate kooperatiboak	Zenbakizkoa
G	Elkartek	Zenbakizkoa
H	Jabetza horizontaleko jabeen erkidegoak	Zenbakizkoa
J	Nortasun juridikodun edo gabeko sozietate zibilak	Zenbakizkoa
P	Tokiko korporazioak	Alfabetuzkoa
Q	Erakunde publikoak	Alfabetuzkoa
R	Kongregazio eta erakunde erlijiosoak	Alfabetuzkoa
S	Estatuko administrazioaren eta autonomia erkidegoetakoan organoak	Alfabetuzkoa

U	Enpresen aldi baterako elkartzeak	Zenbakizkoa
V	Gainerako gakoetan definitu gabeko tipoak	Zenbakizkoa
N	Atzerriko erakundeak	Alfabetuzkoa
W	Espanian ez dauden erakundeen establezimendu egonkorak	Alfabetuzkoa

Kontroleko letra edo digitua kalkulatzeko, urrats hauek ematen dira (gogoan izan kalkulu honek pertsona juridikoen eta, oro har, erakundeen IFZ baten kontroleko zenbakia kalkulatzeko eta K, L eta M letrekin hasten diren pertsona fisikoen IFZ kalkulatzeko balio du). Horretarako, zazpi zifrak hartzen dira oinarri 2. kokapenetik 8.erakoak, idatzita dauden ordenan. Horiek $a_1, a_2, \dots, a_7$ izanda, orduan

1. $A = a_2 + a_4 + a_6$ izanda, azpiindize bikoitiko digituen batura.
2. $B = b_1 + b_3 + b_5 + b_7$ non $b_i = \text{digituen batura } 2 \times a_i$, $i = 1, 3, 5, 7$
3. $C = A + B$, $E = C$ eta $D = 10 - E$ ($E = 0$ bada, orduan $D = 0$)
4. Kontroleko karakterea digitu bat bada, orduan D da. Aldiz, kontroleko karakterea alfabetuzkoa bada, taula honetan D-k duen balioari dagokion letra hartzen da:

D-ren balioa	1	2	3	4	5	6	7	8	9	0
Kontroleko karakterea	A	B	C	D	E	F	G	H	I	J

Probabilitateen kalkulua.

Erregistro bikote baten batze aldagaien balioak alderatzean, hiru adierazle mota desberdin eman daitezke:

$$\gamma = \begin{cases} \gamma_1 & \equiv \text{los valores coinciden y son el } j - \text{ésimo} \\ \gamma_2 & \equiv \text{los valores no coinciden} \\ \gamma_3 & \equiv \text{algún valor está ausente} \end{cases}$$

Non $j = 1, \dots, m$ balio zehatz bat da, m balio desberdinena. Eta balio horietako bat har dezake batze aldagaiak.

Fase honetan probabilitate hauek kalkulatzeko dira:

- Batzeko aldagaien balioak bat etorri eta batzeko aldagaien balizko j -garrenaren berdinak izateko probabilitatea, erregistro bikoteak unitate ekonomiko bera adierazten baitu ($m(\gamma_1)$).
- Batzeko aldagaien balioak bat etorri eta batzeko aldagaien balizko j -garrenaren berdinak izateko probabilitatea, erregistro bikoteak ez baitu unitate ekonomiko bera adierazten ($u(\gamma_1)$).

- Batzeko aldagaien balioak bat ez etortzeko probabilitatea, erregistro bikoteak unitate ekonomiko bera adierazten baitu ($m(\gamma_2)$).
- Batzeko aldagaien balioak bat ez etortzeko probabilitatea, erregistro bikoteak unitate ekonomiko bera adierazten ez duelako ($u(\gamma_2)$).
- Batzeko aldagaien balioetakoren bat ez egoteko probabilitatea, erregistro bikoteak unitate ekonomiko bera adierazten baitu ($m(\gamma_3)$).
- Batzeko aldagaien balioetakoren bat ez egoteko probabilitatea, erregistro bikoteak unitate ekonomiko bera adierazten ez duelako ($u(\gamma_3)$).

Probabilitate hauen kalkulu zehatzak bibliografian aipatutako dokumentazioan kontsulta daitezke.

Blockinga.

Jarraian erabiltzaileak programan aitortutako blocking irizpideak daude. Gainera, blocking irizpideak aukeratzen dituen elementu bikote multzo bakoitzaren amaierako pisua kalkulatzen da.

Erabiltzaileak aitortutako blocking irizpideen arabera, dagokion makroa burutzen da, blockingeko irizpideren bat betetzen duen elementu bikoteekin datu-multzoa lortzeko.

- **%blockingIFZ.** Enpresaren IFZrekin bat datozen elementu bikoteak aukeratzen ditu.
- **%blockingPK.** Posta-kode bera duten elementu bikoteak aukeratzen ditu. Kasu honetan posta-kode bat baino gehiago egon daitezke. Beraz, posta-kode motako batze aldagai bat baino gehiago egonez gero, mota horretako aldagai guztiak bat etortzeko egiten da blockinga. Hau da, posta-kode motako aldagai guztiak bat datozen elementu bikoteak lotzen dira.
- **%blockingKalea.** Kalearen izenari lotutako kodearekin bat datozen elementu bikoteak aukeratzen ditu. Kasu honetan kalearen izen motako batze aldagai bat baino gehiago egon daitezke. Beraz, kalearen izen motako batze aldagai bat baino gehiago egonez gero, mota horretako aldagai guztiak bat etortzeko egiten da blockinga. Hau da, kalearen izen motako aldagai guztiak bat datozen elementu bikoteak lotzen dira.
- **%blockingLHUD.** Enpresaren izen kodifikatuaren lehenengo hitza, lurralde historikoa eta enpresaren udalerrria bat datozen elementu bikoteak aukeratzen ditu.

Blockingerako nahi adina irizpide aitortu daitezke. Bat baino gehiago aitortzea gomendatzen da, unitate ekonomiko bera adierazten duen elementu bikote bat ez dadin batu blocking aldagaiko akatsengatik.

Blocking irizpideren bat betetzen duen elementu bikotea eraikitakoan, bakoitzaren pisu osoa honela kalkulatzen da:

$$w = \sum_{k=1}^K w_k \quad \text{non} \quad w_k = \log \frac{m(\cdot)}{u(\cdot)} = \log m(\cdot) - \log u(\cdot), \quad k = 1, \dots, K$$

K da programaren erabiltzaileak aitortutako batzeen aldagai kopurua.

Blocking irizpideren bat betetzen duten elementu bikote guztientzako osoko pisua kalkulatutakoan, erabiltzaileari histograma bat erakusten zaio bikote horien maiztasunarekin, pisu muga zehazteko.

Linkak

Elementu bikote bat emanda bi gauza gerta daitezke: Elementuek unitate ekonomiko bera ordezkatzea edo unitate ekonomiko desberdinak izatea.

Egoera teoriko idealean bi elementuek unitate ekonomiko bera ordezkatzen badute, batze aldagai guztiek bat etorri beharko lukete eta, ondorioz, osoko pisu oso handia izan. Baina hau ez da kasu guztietan gertatzen, batzeko aldagai batzuetan aldeak egon baitaitezke, akats tipografikoengatik edo egoera aldaketengatik edo batzeko aldagairen baten balioen bat ez dagoelako.

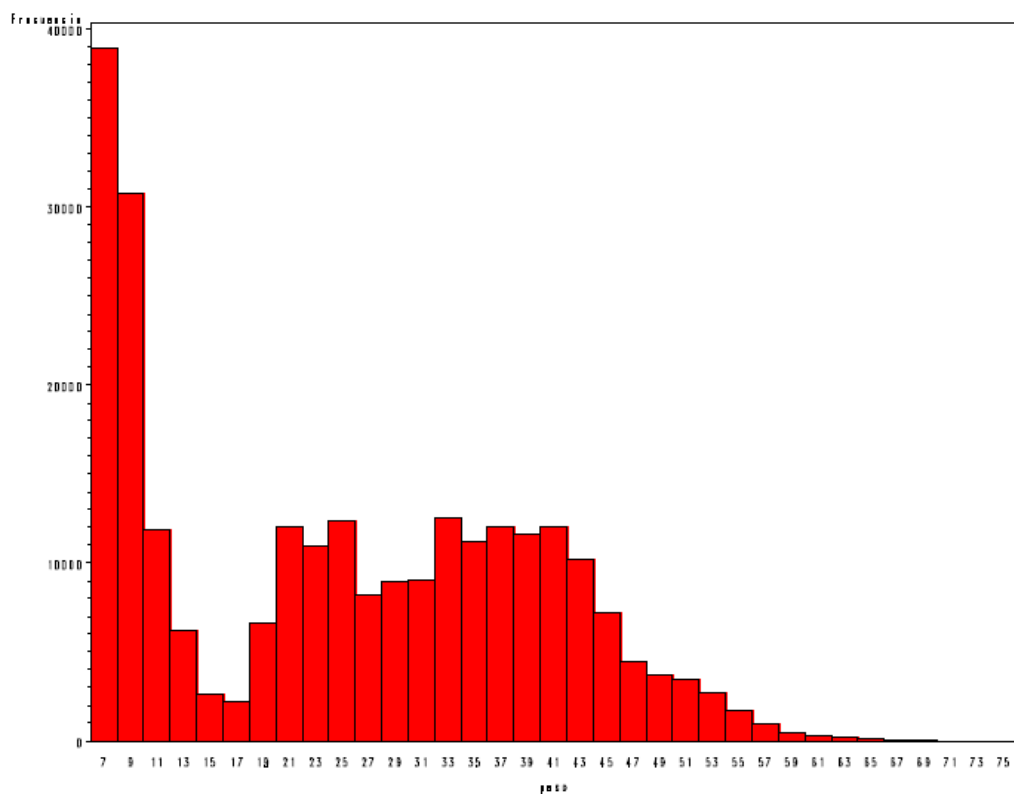
Aldiz, bi elementuek unitate ekonomiko desberdinak ordezkatzen badituzte batzeko aldagaiek ez lukete bat etorri behar eta elementu bikoteak pisu oso txikia izan behar luke (baita negatiboa ere). Baina ausazko bat etortzeak egon daitezke batzeko aldagai batzuetan eta pisuak gora egin.

Ondorioz, egoera teoriko idealean histograman bi atal argi leudeke, bata eskuinean, pisu altu baten inguruan, unitate ekonomiko bat bera ordezkatzen duten elementu bikoteena eta bestea ezkerrean, pisu baxu baten inguruan, unitate ekonomiko desberdinen elementu bikoteena.

Praktikan atal hauek ez daude hain bereizita eta tarteko pisuetan daude balio batzuk, bihurgune bat eratuaz.

Beraz, erabiltzaileak histogramari begiratu beharra dauka eta pisu bat ezarri, pisu berdineko edo handiagoko elementu bikote guztiak batutzat hartzeko. Mugako pisu bezala histogramako ibarra hartzea gomendatzen da.

Jarraian dago pisuen histograma baten adibidea:



Erabiltzaileak mugako pisuaren balioa erabakitakoan, lerroko izendatze prozedura egiten da blocking fasean aukeratutako unitate bikoteekin, batutakotzat hartutako elementuak ezartzeko.

EMAITZEN AZTERKETA

Atal honetan erregistroen batze programako aplikazioa deskribatuko da, fitxategi administratiboetara bideratuta dago eta aurrez fitxategi hauetan deskribatutako unitate ekonomikoen informazioa dute: EUSTATEko Jarduera Ekonomikoen Direktorioa (JED) eta Gizarte Segurantzako fitxategia.

Fitxategien deskripzioa

Jarduera ekonomikoen direktorioa (JED)

Jarduera Ekonomikoen Direktorioko informazioa hiru datu-multzo hauek ematen dute:

- Tokiko jarduera ekonomikoen unitateen JEDEko datuak dituen fitxategi batek.

Tokiko jarduera ekonomikoen unitatea jarduera irizpideen araberrako tokiko unitatearen azpizati bat da, soilik JEDEkoa.

- Unitate juridikoen JEDEko datuak dituen fitxategi batek.

Unitate juridikoak pertsona juridikoak dira. Legeak aitortzen du badirela, dituzten edo kide diren erakunde edo pertsonak dena delakoak izanda ere. Pertsona fisikoak ere izan daitezke unitate juridiko, jarduera ekonomikoren bat egiten badute modu independentean.

- Tokiko unitateen JEDEko datuak dituen fitxategi batek.

Tokiko unitatea enpresa edo enpresa zati bati dagokio (tailerra, lantegia, biltegia, bulegoa, meatzea, gordailua) eta topografikoki mugatutako tokian egongo da. Toki horretan edo hortik abiatuta egiten dira enpresa bat beraren kontura lagun bat edo gehiago (lanaldi partzialean ere bai) aritzen diren jarduera ekonomikoak –salbuespenetan izan ezik.

Gizarte Segurantza

Gizarte Segurantzako fitxategiaren informazioa bi datu multzotan ematen da:

- Enpresariei dagozkien Gizarte Segurantzako datuak dituen fitxategian.
- Kotizazio-unitateei dagozkien Gizarte Segurantzako datuak dituen fitxategian.

Gizarte Segurantzako fitxategia hiruhilero eguneratzen da, beraz, urte natural batean 4 fitxategi jasotzen dira. Erregistroak batzean erabilitako fitxategiak erregistro baten urtebeteko aldagai guztiak zenbatzen ditu. Hau da, establezimenduren baten ezaugarriren bat aldatzen bada, erregistro hau gehitzen da fitxategira eta lau aldiz dauden establezimenduak egon daitezke.

Fitxategien azterketa

Fitxategi administratiboak batu aurretik, batuko diren datu multzoak eraiki behar dira. Horretarako, iturri bakoitzerako datu multzo bakarra eraikitzen da batzeko beharrezko informazio guztiarekin; hau da, erregistroan unitate bakoitza identifikatzen duten gakoak eta batzeko balio duten bi erregistroen aldagai komun guztiak. Gure kasuan **ula.sas7bdat** datu multzoak eraiki ziren JEDeko tokiko unitate bakoitzaren informazioarekin eta **uco.sas7badat** Gizarte Segurantzako kotizazio-unitateetako informazioarekin.

Jarraian batzean kontuan izango diren aldagaien taula bat dago:

JED			GIZARTE SEG.		
UNITATE JURIDIKOA	UJA_CIF	IFZ	ENPRESARIA	EMP_COFDNI	IFZ/NAN
	UJA_NOMBRE	Izena		EMP_NOMBRE	Izena
	UJA_PROV	Probintzia		EMP_TH	Lurralde Historikoa
	UJA_MUN	Udalerria		EMP_MUN	Udalerria
	UJA_CP	Posta-kodea		EMP_CP	Posta-kodea
	UJA_CALLE	Kalearen izena		EMP_T_CALLE	Kalearen izena
TOKIKO UNITATEA	ULA_TH	Lurralde Historikoa	KOTIZAZIO UNITATEA	UCO_TH	Lurralde Historikoa
	ULA_MUN	Udalerria		UCO_MUN	Udalerria
	ULA_CP	Posta-kodea		UCO_CP	Posta-kodea
	ULA_CALLE	Kalearen izena		UCO_T_CALLE	Kalearen izena

Ikusten denez, batzeko aldagai gutxi dago. IKZ/NAN aldagaiak enpresa adiera bakarrean identifikatzen du. *Izena* aldagaiak ere baztertzeko gaitasun handia du, nahiz eta akats ugari izan dezakeen. Aldiz, kokatzeko beste aldagaiak (*probintzia*, *udalerria*, *posta-kodea* eta *kalearen izena*) ez dira oso adierazgarriak.

Batzeko aldagai bakoitzaren baztertzeko gaitasunaz gain, garrantzitsua da kontuan izatea erregistro administratiboetako elementu guztiek ez dutela aldagaientzako informazioa izaten. Ondorengo taulan daude aldagai bakoitza ez duten elementuak:

JED (ula.sas7bdat)			GIZARTE SEG. (uco.sas7badat)		
200.675 erregistro			554.177 erregistro		
Aldagaia	MISSK	MISSEH	Aldagaia	MISSK	MISSEH

UJA_CIF	0	% 0	EMP_COFDNI	265	% 0.05
UJA_NOMBRE	0	% 0	EMP_NOMBRE	1	% 0.00
UJA_PROV	0	% 0	EMP_TH	143673	% 25.93
UJA_MUN	0	% 0	EMP_MUN	143673	% 25.93
UJA_CP	18	% 0.01	EMP_CP	140996	% 25.44
UJA_CALLE	179	% 0.09	EMP_T_CALLE	143673	% 25.93
ULA_TH	0	% 0	UCO_TH	112027	% 20.22
ULA_MUN	1437	% 0.72	UCO_MUN	112027	% 20.22
ULA_CP	5958	% 2.97	UCO_CP	101856	% 18.38
ULA_CALLE	1653	% 0.82	UCO_T_CALLE	104213	% 18.81

Aurreko taulan ikusten denez, JED osoa den arren, Gizarte SEgurantzako fitxategiko kokapen aldagaietan balio asko falta da, eta horrek batzea zaildu egiten du.

Blockingeko irizpide desberdinak erabilita probak egin dira. Kasu guztietan IFZ blockinga erabili da, informazio asko ematen baitu eta nahiko ongi erregistratuta baitago.

Kontrolako karakterearen baliozkotasuna aztertzeko programatutako makroari esker dakigu hori. Azterketa horretako emaitzen taula da honakoa:

ValidarNIF.s as emaitza	JED	Gizarte Segurantzza
-3	8	1
-2	38	0
-1	55	6
0	140	265
1	102109	424931
2	92616	112602
3	5706	16372
4	3	0
Guztira	200675	554177

Gogoan izan balio negatiboek kontrolako karakterea ez dela zuzena adierazten dutela eta 0k adierazten du IFZ formatua ez dela estandarra (hau da, ez dituela 9 karaktere). Beraz, taulan ikus daiteke:

- JEDeko tokiko unitateetan establezimenduen % 0,05ek okerreko IFZ duela eta % 0,07ren IFZ formatua ez dela osoa. Gainera, 3 establezimenduk behin behineko IFZ dute (IFZren balidazio kodea = 4) eta hori ezin dela balidatu.
- Gizarte Segurantzako kotizazio unitateei dagokienez, % 0,001ek IFZ okerra du eta % 0,05ek IFZ formatu ez osoa.

Batzearen emaitza

Blockingeko irizpide desberdinak erabilia batze programa burutzean lortutako emaitzen laburpen-taula bat dago jarraian.

Blocking irizpideak	Mugako pisua	Egiteko denbora	Batutakoak		
			1. fasea	2. fasea	Guztira
IFZ	20	10 ordu	14748	111948	126696
IFZ + Kalea	28	39 orduren ondoren bertan behera utzi	-----	-----	-----
IFZ + NTHMUN	32	6 ordu	14748	80040	94788

1. fasean batzeko aldagai guztiak bat datozen elementu bikote guztiak batu ziren, hau da, zuzeneko batzea izan zen. 2. fasean esleipen linealeko algoritmoak hautatutako elementu bikoteak batzen dira. Mugako pisua baino handiagoa dutenak hautatzen ditu, hau da, probabilitate prozedurak batutako elementu bikoteak dira.

ONDORIOAK

Informazio administratiboak oztopo ugari izaten ditu estatistikan erabiltzeko. Erregistro administratiboak ez dira estatistikarako sortu eta diseinatu eta ematen diguten informazioa erabiltzeko aurrez tratatu egin behar da.

Proiektu honekin gizarte segurantzako erregistroa tresna indartsu bihurtu dugu, JED eraiki eta mantentzeko erabilera estatistikorako. JED funtsezko erreferentziazko direktorioa dugu eta hori gabe ezingo genuke estatistika fidagarria eta kalitatezkoa egin. Gizarte Segurantzako informazioa aldagai garrantzitsuak eguneratzeko erabiltzen da, adibidez, langile kopurua eguneratzeko.

Probabilitatezko batze metodoa eta aurreko tratamenduak ezartzearen ondorioz, erregistroak batzeko ehuneko onargarriak eman dira, batzeko aldagaien kalitatea kontuan izanda. Koadernoan deskribatu dugun bezala, batzean erabilitako aldagai lagungarri askok missing balioak dituzte eta horrek tratamendu oro zailtzen du. "Zuzenean" batuz gero, establezimenduen % 10era ere ez litzateke iritsiko.

Proiektuaren ideia orokorra establezimenduen edozein bi erregistro batzeko planteatu zen, banakoekin egin zen bezala. Eskura zeuden establezimenduen erregistroak aztertu ondoren, bakoitzaren berezitasunen ondorioz batera tratatzea ezinezkoa zela ikusi zen eta makro sekuentzial eta modular bat programatzea erabaki zen, moduluak berriro erabili ahal izateko eta egokitzeko, establezimenduen etorkizuneko edozein batzetan.

Azterketa hobetu egingo da, erregistro administratiboak hobetzen diren heinean eta blocking eraginkorra egiteko bide berriak aztertzen diren heinean. Horiei esker fitxategia bloke txikietan zatitu ahal izango da burutzeko denbora (alderaketa kopurua) onargarria izan dadin eta kalitatea eta balizko datu bikoteak ez galtzeko bezain zabala izan dadin.

Lortutako emaitzak ebaluatzeko, kalitate-kontrola egin zen batze programako bigarren burutzapen fasean batutako unitate bikote batzuen laginak ausaz hartuta. Lehenengo fasean unitate ekonomikoen bikoteak batzen dira, eta batzeko aldagai guztiek bat etorri behar izaten dute. Bigarren fasean erabilitako probabilitate-metodologiak zehaztutako erregistro bikoteak batzen dira. Batutako erregistro gehienetan egiaztatu zen batzea zuzena edo gutxienez logikoa zela. Adibidez, enpresa bereko establezimendu bakarrak batzea lortu da, batzeko aldagai garrantzitsu batzuk ez dauzkaten arren.

Proiektu honen emaitza oso ona izan da eta hemendik aurrera EUSTATEko jarduera ekonomikoen direktorioko kalitatean eragina izango duen tresna bat edukiko dugu.

BIBLIOGRAFIA

[1] E. E. ERAKUNDEA

ERREGISTROAK BATZEKO METODO AUTOMATIKOAK ETA EUSTATEN DUTEN ERABILERA http://www.eustat.es/document/datos/ct_15_e.pdf

[2] I.P. FELLEGI eta A.B. SUNTER

A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183-1210, 1969

[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. *International Journal of Epidemiology* (2002).

[5] AYESTARAN, MARINA AND LEGARRETA, LEIRE.

Applying methods of record linkage for census validation in the Basque Statistics Office. Instituto Vasco de Estadística (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).