

$$fs_{1ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \frac{|x_{ij} - \tilde{x}_{ij}|}{\tilde{x}_{ij}}$$

$$fs_{3ij} = \frac{w_i \left(\frac{\tilde{x}_{ij}}{\tilde{y}_{ij}} \right)}{\left(\frac{\tilde{X}_j}{\tilde{Y}_j} \right)} \times \left[\frac{x_{ij}}{y_{ij}} - \left(\frac{\tilde{x}_{ij}}{\tilde{y}_{ij}} \right) \right]$$

DATU IRAGAZKETA SELEKTIBOA

2014



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

Erakunde Autonomiaduna



EUSKO JAURLARITZA
GOBIERNO VASCO

Lanketa:
EUSTAT
Euskal Estatistika Erakundea
Instituto Vasco de Estadística

Argitalpena:
EUSTAT
Euskal Estatistika Erakundea
Instituto Vasco de Estadística
Donostia-San Sebastián 1
01010 Vitoria-Gasteiz

© **Euskal AEko Administrazioa**

Lehen edizioa
I/2015

Imprimaketa eta Enkoadernazioa:
Eusko Jaurlaritzako Inprimategia eta Erreprografia Zerbitzua

ISBN: 978-84-7749-480-5

Lege-gordailua: VI-792/2014

ARAZKETA SELEKTIBOA

Imanol Montoya Arroniz

imanolmontoya@gmail.com



**EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA**

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

Aurkezpena

Estatistika-erakundeentzako funtsezkoa da datuak arazteko metodo eraginkorrak edukitzea. Izan ere, datuen kalitatea hobetzeko prozesuan, datuak araztea da, hain zuzen, denbora eta diru gehien eskatzen duen alderdia.

Helburua da datu-baseen arazketa selektiboko teknikak aztertzea eta aplikatzea. Arazketa selektiboaren arabera, argitaratu beharreko emaitzetan eragin handia izango duten erregistroak baino ez dira araztu eta zuzendu behar. Horri esker, kostuak eta entregatze-epea murrizten dira.

Dokumentu hau kapitulu batzuetan banatuta dago. Lehengo ataletan, metodologia azalduta dago, simulazio-azterlan bat jaso da, aurretik azaldu den metodologia era praktikoan erakusteko helburuz. Ondoren, azalduta daude SAS makroak, datu-baseak selektiboki arazteko prestatuak. Aurrerago, egiazko adibide bat azalduta dago. Zehazki, datu-baseak arazteko proposatu den metodologia erabili da Euskal Estatistika Erakundearen zerbitzu-estatistika berrian. Bukatzeko, metodologiaren eraginkortasunari eta erabilgarritasunari buruzko ondorio batzuk azaldu dira.

Vitoria-Gasteiz, 2014ko abendua

Josu Iradi Arrieta

Zuzendari nagusia

Aurkibidea

AURKEZPENA	2
AURKIBIDEA.....	3
1. SARRERA	3
2. ARAZKETA SELEKTIBOAREN GAINEKO SARRERA	4
3. MIKROHAUTAKETA.....	6
"SCORE" FUNTZIOA	6
"SCORE" FUNTZIO-MOTAK	7
"SCORE" FUNTZIOA ERAIKITZEKO BESTE ESTRATEGIA BATZUK	11
"SCORE" FUNTZIO GLOBALA	13
ATARIA ZEHAZTEA	14
4. MAKROHAUTAKETA	16
AGREGAZIO-METODOA	16
BANAKETA-METODOA	18
5. SIMULAZIOA	20
DATU-BASE SIMULATUA	20
ERRORE-MOTAK SORTZEA	21
"SCORE" FUNTZIOAK	22
SIMULAZIOAREN EMAITZAK	25
6. SAS MAKROA	27
FUNCION_SCORE MAKROA.....	27
FS_GLOBAL MAKROA	28
7. EZARPEN PRAKTIKOA ZERBITZU-ESTATISTIKA DEITURIKO JARDUERAN	30
ARAZKETA SELEKTIBOAREN EZARPEN PRAKTIKOA	30
7.1. taula. Negozioen kopuruaren zenbateko garbiaren aldagaia araztea.....	31
7.2. taula. Negozioen kopuruaren zenbateko garbiaren aldagaia araztea pertsona bakoitzeko eta establezimendu bakoitzerako	32
7.3. taula. Faktore-kostuari erantsitako balioa araztea.....	33
7.4. taula. Pertsona bakoitzeko faktore-kostuari erantsitako balioa araztea.	33
7.5. taula. Langileen kostuak pertsona bakoitzeko araztea.	34
ARAZKETA SELEKTIBOAREN EZARPENAREN LABURPENA	34

.../...

8. ONDORIOAK	36
ARAZKETA SELEKTIBOAREN GAINEKO LABURPENA ETA ONDORIOAK	36
BIBLIOGRAFIA	38

1. Sarrera

Koaderno tekniko honetan bildutakoa Euskal Estatistika Erakundeak, 2012. urtean, datu-baseen arazketarako estatistika- eta matematika-metodologietan prestatzeko eta ikertzeko emandako bekari esker egindako lanaren emaitza da.

Dokumentua honako kapitulu hauetan banatuta dago:

Bigarren kapitulan sarrera bat egin da, eta koaderno tekniko hau egitean izandako helburuak aipatu dira.

Hirugarren kapitulan, “score” funtzioa zer den zehaztu, eta mikrohautaketa egiteko metodologia azalduta dago. Horrez gain, honako gai hauek ere bildu dira: zenbait funtzio-mota, funtzioak osatzeko estrategiak, nola konbina daitezkeen “score” funtzio global batean eta, azkenik, nola finkatzen den ataria.

Laugarren kapitulan, makrohautaketa-kontzeptua garatu da, eta agregazio-metodoa eta banaketa-metodoa bereizi dira.

Bosgarrenean, simulazio-azterlan bat jaso da, aurretik azaldu den metodologia era praktikoan erakusteko. Halaber, honako hauek ere azalduta daude: datu-base simulatua nola sortu den, “score” funtzioak nola kalkulatu diren eta zer emaitza lortu diren.

Seigarren kapitulan labur azalduta daude SAS makroak, datu-baseak selektiboki arazteko prestatuak.

Zazpigarren kapitulan, egiazko adibide bat azalduta dago. Zehazki, datu-baseak arazteko proposatu den metodologia erabili da Euskal Estatistika Erakundearen zerbitzu-estatistika berrian.

Amaieran, metodologiaren eraginkortasunari eta erabilgarritasunari buruzko ondorio batzuk azaldu dira.

Zinez eskertu nahi dut Metodologia, Berrikuntza eta I+G Ataleko kide guztiek emandako laguntza eta, oro har, Eustateko langile guztien adeitasuna.

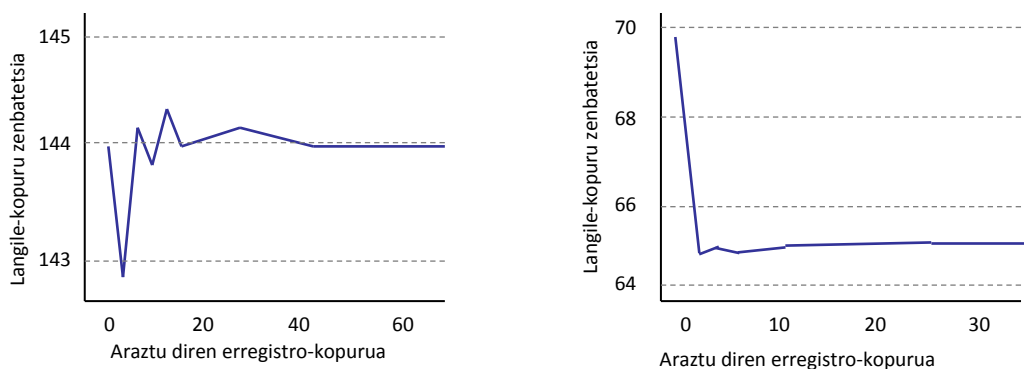
GAKO-HITZAK: arazketa selektiboa, “score” funtzioa, “score” funtzio globala, mikrohautaketa, makrohautaketa.

2. Arazketa selektiboaren gaineko sarrera

Datuen kalitatea hobetzeko prozesuan, datuak eskuz edo interaktiboki araztea da, hain zuzen, denbora eta diru gehien eskatzen duen alderdia. Iraganean, erregistro guztiak eskuz araztu ohi ziren, baina beharrezkoak ziren langile eta lanordu asko. Azken hamarkadotan, eskuzko arazketaren gaineko azterketak erakutsi du eskuz araztu beharreko erregistroen kopurua askoz ere txikiagoa izan daitekeela. Izan ere, eskuzko arazketak ia ez du eraginik hainbat erregistrotan, behintzat, intereseko parametro nagusien zenbatesleei dagokienez.

Azterlan batzuek (besteak beste, Granquist, 1995; Granquist and Kovar, 1997; eta Hoogland, 2000) erakutsi dute, oro har, intereseko parametroen gaineko zifra fidagarriak lortzeko errore guztiak ez direla zuzendu behar, baizik eta eragin handiena sortzen dutenak.

Honako grafiko honetan irudikatu da zer eragin duen errore guztien segidako zuzenketak intereseko parametroen kalkuluan. Adibide honetako parametroa langileen kopurua da (Hoogland, 2000).



1. irudia: Langileen kopuru zenbatetsia, araztu diren erregistroen kopuruaren arabera. Azkeneko zenbateslean duten eragina zein den, erregistroak antolatu dira. Ezkerreko grafikoaren oinarria eraikuntza-enpresak dira, eta eskuinekoa ingeniariarritza zibileko enpresak. (Hoogland, 2000).

Arazketa selektiboaren arabera, argitaratu beharreko emaitzetan eragin handia izango duten erroreak baino ez dira araztu eta zuzendu behar. Estrategia horri jarraituta, kostuak eta entregatze-epea murriztuko dira.

Datu-base batean zer erregistro araztu erabakitzeko, zenbait metodo erabil daitezke. Datu-bilketaren lehenengo faseetan erabiltzen direnean, bilketa oraindik bukatuta ez

dagoenean, metodo horiei mikrohautaketa esaten zaie. Metodo horiek, or har, banaka ezartzen zaizkie erregistroei. Aurreko aldietako datuak edo azpitalde homogeneoei buruzko zenbatespenak hartzen dituzte oinarri. Bestalde, makrohautaketa-metodoen erabilera pentsatuta dago ia datu guztiak bilduta daudenean gauzatzeko. Metodo horiek eskuragarri dauden datu guztien informazioa aztertzen dute, balio eragingarrien bila.

Koaderno tekniko honetan azalduta daude araztu beharreko erregistro eragingarriak hautatzeko metodo batzuk. Horretarako, honako iturri hauek erabili dira, batik bat: *Handbook of Statistical Data Editing and Imputation* (Wall, Pannekoek and Scholtus, 2011) liburuko “Selective Editing” kapitulua; Holandako Estatistika Institutu Nazionalak argitaratu dituen koaderno teknikoak (Hoogland, van der Loo, Pannekoek and Scholtus, 2011; eta Wall, 2008) eta europar proiektuetan jasotako gomendioak (EUREDIT Project, 2004; eta EDIMBUS, 2007).

3. Mikrohautaketa

Hauxe da mikrohautaketaren helburu nagusia: datuak biltzen bukatu ez bada ere, araztu beharreko erregistroak hautatu ahal izatea.

Atal honetan, honako hauek azalduko ditugu: “score” funtzioa; funtzioa osatzeko era ohikoenak; funtzioa osatzeko beste estrategia batzuk; balio global bat kalkulatzeko, funtzioak nola konbinatu; eta atariko balioa, araztu beharreko erregistroak hautatzeko bide emango duena, nola zehaztu.

“Score” funtzioa

Erregistroak arazterakoan, mikrohautaketarako prozesuko tresna nagusia “score” funtzioa da. Funtzio horrek puntuazioa (score) esleitzen dio erregistro bakoitzari, aztertu den aldagai bakoitzeko. Puntuazio horrek adierazten du erregistro hori araztuz gero, zer eragin izango duen zenbatetsi beharreko parametroan. Lehenik, puntuazio handiko erregistroak araztu beharko dira.

“Score” funtzio lokalak neurtzen du erregistro baten aldagai zehatz bat arazteak zer eragin izango duen. “Score” funtzio horrek bi osagai izan ohi ditu: arriskua eta eragina. Arriskuak balizko errore baten tamaina eta probabilitatea barne hartzen ditu; eta eraginak adierazten du erregistro horren ekarpena azterketa-parametroaren zenbatespenean. Honako bi osagai hauek biderkatuta, puntuazio lokalak lortzen ditugu:

$$s_{ij} = F_{ij} \times R_{ij} = \text{influencia}_{ij} \times \text{riesgo}_{ij}$$

s_{ij} “score” funtzioa da i erregistrorako, j aldagaian. Eraginari lotutako osagaiak, oro har, hauxe neurtzen du: aurreikusitako edo itxarondako balioak zenbatesle osoari egiten dion ekarpen erlatiboa. Arrisku-osagaiak balio gordinaren eta aurreikusitako edo itxarondako balioaren arteko konparazioaren bidez neurtu ohi da. Balio bataren eta bestearen arteko desbiderapenak txikiak baldin badira, ez dago zertan pentsatu errorerik dagoenik; desbiderapen handiek, ordea, errore handia egon daitekeela adierazten dute.

Puntuazio globala puntuazio lokalak konbinatzen dituen funtzio bat da; eta erregistro osorako neurraketa osatzea du helburu.

$$S_i = f(s_{i1}, \dots, s_{iJ})$$

Mikrohautaketa-metodoak ezarri ahal dira, datuen bilketa bukatu ez bada ere. Erregistro bat erabilgarri dagoenean, puntuazio globala lortzen da, eta aurretik zehaztu den atariko balio batekin erkatzen da. Puntuazioak atari hori gainditzen baldin badu,

erregistroa ez-onargarritzat joko da. Erregistro horiek araztu beharreko erregistroen artean sartuko dira.

Formalki, hautaketa horren oinarria onarpen-mailaren adierazlea da, eta hau da haren definizioa:

$$\text{Indicador plausibilidad}_i = \begin{cases} 1 & (\text{plausible}) \text{ si } S_i \leq C, \\ 0 & (\text{no plausible}) \text{ en otro caso} \end{cases}$$

Catariko balioa da.

Mikrohautaketa-estrategia honako hiru urrats hauen bidez labur daiteke:

- “Score” funtzio lokalak kalkulatzeko, intereseko aldagai nagusietarako. Horretarako, erreferentziatzen har daitezke aurreikusitako edo itxarondako balioak, aurreko aldiak edo azpitalde homogeneoetako datuetan oinarrituta.
- Puntuazio lokal horiek puntuazio global batean nahasteko funtzio bat zehaztea.
- Atariko balioa, araztu beharreko balioak zehaztuko dituen, zehaztea.

“Score” funtzio-motak

o “Score” funtzio-motak guztizkoetarako

“Score” funtzio batek zenbatu behar du erregistro bat arazteak zer eragin izango duen intereseko zenbateslean. Demagun x_{ij} dela x_j aldagaiaren balioa i erregistroan. Gure intereseko zenbateslea guztizkoa baldin bada, honela zehatz daiteke:

$$\hat{X}_j = \sum_{i \in D} w_i \hat{x}_{ij}$$

Bertan, D datuen multzoa da; eta i erregistroak. Horrez gain, w_i pisuek zuzentzen dute inklusio-probabilitate desberdinaren arabera, edota erantzun-ezaren arabera. Araztutako datuak $\hat{x}_{ij}$ dira. Horrek esan nahi du datu gordinen erregistro batzuk, x_{ij} , arazketa-prozesuan sartu, eta zuzendu egin direla. Erregistro gehienei dagokienez, x_{ij} , ontzat joko da, eta hura eta $\hat{x}_{ij}$ berdinak direla hartuko da. Hortaz, erregistro bat arazteak guztizkoaren gainean duen efektua (efektu batukorra) esan dakioke i araztutako erregistroa kontuan hartzeak ala baztertzeko guztizko kopuru zenbatetsian sortuko duen aldeari edo diferentziari. i erregistroa ez araztuz gero, guztizko kopuru zenbatetsia $\hat{X}_j - w_i(\hat{x}_{ij} - x_{ij}) = \hat{X}_j^{(-i)}$ da. Hortaz, honela adieraz daiteke diferentzia:

$$d_i(\hat{X}_j) = \hat{X}_j^{(-i)} - \hat{X}_j = w_i(\hat{x}_{ij} - x_{ij})$$

$d_i(\hat{X}_j)$ diferentzia $\hat{x}_{ij}$ balio ezezagunaren eta zuzenduaren mende dago; beraz, kalkulazina da. “Score” funtzio batek oinarri hartzen du $\hat{x}_{ij}, \tilde{x}_{ij}$ balio ezezagunerako hurbilketa. Balio ezezagun horri itxarondako balioa esaten zaio. Oro har, hauexek dira itxarondako balioak:

- Erregistro beraren balio araztuak, aurreko denboraldietan inkesta berean lortuak, bi denboraldien arteko bilakaeraren gaineko zenbatespen batez biderkatuta.
- Erregistro beraren aldagai antzeko baten balioa, beste datu-iturri batetik eskuratuta.
- Aurreko denboraldi bateko antzeko erregistroen azpitalde homogeneo bateko intereseko aldagaiaren batez bestekoa edo erdiko balioa.

$d_i(\hat{X}_j)$ diferentzia w_i pisuen mende ere badago. Izan ere, pisu horiek inklusio-probabilitate desberdinaren arabera egiten dute zuzenketa. Baina, horrez gain, erantzun-eza ere, datuak biltzen amaitu arte ezezaguna, zuzentzen dute. “Diseinuaren pisuak”, v_i , dira kasu horietan erabiltzen den hurbilketa. Pisu horiek inklusio-probabilitate desberdinaren arabera soilik egiten dute zuzenketa.

Hurbilketa horiek erabilia, i erregistroa arazteagatik sor daitekeen efektua zenbatu daiteke “score” funtzioaren bidez:

$$s_{ij} = v_i |x_{ij} - \tilde{x}_{ij}| = v_i \tilde{x}_{ij} \times \frac{|x_{ij} - \tilde{x}_{ij}|}{\tilde{x}_{ij}} = F_{ij} \times R_{ij} = \text{influencia}_{ij} \times \text{riesgo}_{ij}$$

Hortaz, esan daiteke “score” funtzio hori eragin-faktore baten eta arrisku-faktore baten biderkadura dela. Arrisku-faktorea, $R_{ij} = |x_{ij} - \tilde{x}_{ij}| / \tilde{x}_{ij}$, balio gordinaren eta itxarondako balioaren arteko diferentziaren neurketa erlatiboa da. Aldea oso handia baldin bada, balioa okerra izan daitekeela adierazten du. Eragin-balioa, $v_i \tilde{x}_{ij}$, erregistroak guztizko balio zenbatetsian egiten duen ekarpena da.

Arriskua eragin-faktoreaz biderkatzen badugu, erregistroaren arazketak guztizko zenbatetsian izango zuen efektuaren neurketa bat lortuko dugu. Balioak oso handiak baldin badira, baliteke erregistro horrek errore eragingarria edukitzea; hortaz, komeni litzateke erregistroa berrikustea. Balioak, aldiz, txikiak direnean, litekeena da erregistro horiek errore eraginkorrik ez izatea. Hortaz, ez da ezinbestekoa zehatz-mehatz araztea.

Aldagaiak ez-negatiboak baldin badira, ekonomiari buruzko inkesta gehienetan gertatzen den bezala, arrisku-faktoreak oinarri har dezake balio gordinaren eta

itxarondako balioaren arteko ratioa, balio horien arteko diferentzia absolutua hartu beharrean.

$$(x_{ij} - \tilde{x}_{ij}) / \tilde{x}_{ij} = \frac{x_{ij}}{\tilde{x}_{ij}} - 1$$

Horrela eginez gero, arriskuaren adierazpena izango da balio gordinaren eta itxarondako balioaren arteko ratioa; eta -1 gehituko zaio, bi balioak berdinak izanez gero, arriskua zero izango dela ziurtatzeko. Dena den, adierazpen horrek oraindik ez du kontuan hartzen diferentzia handiek eta txikiek itxarondako baliotik aldentzen diren desbiderapenak jakinarazten dituztela. Beraz, puntu hori zuzentze aldera, honako arrisku-funtzio hau, ratioan oinarrituta, zehaztu da:

$$R_{ij} = \max\left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}}\right) - 1$$

Funtzio horrek ziurtatzen digu zenbateko bereko gehikuntza biderkatzaileek, gorakoek edo beherakoek, puntuazio bera izango dutela. Arrisku hori eragin-faktoreaz biderkatzen baldin badugu, “score” funtzio bat lortuko dugu, batuketa-funtzioaren ordezkia:

$$s_{ij} = v_i \tilde{x}_{ij} \times \left(\max\left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}}\right) - 1\right)$$

Azkenik, oro har, “score” funtzioaren bertsio eskalatua erabiltzen da. Horretarako, F_{ij} eragin-faktorearen ordezkia $F_{ij} / \sum_i F_{ij}$ eragin erlatiboa erabiltzen da, honela:

$$\sum_i F_{ij} = \sum_i v_i \tilde{x}_{ij} = \tilde{X}_j$$

Hori dela eta, lortu den balio eskalatua jatorrizko balioa zati guztizko zenbatetsia da. Guztizkoa itxarondako balioetan oinarrituta dago. Balioa eskalatzen baldin bada, balio hori ez da aztertutako aldagairen tamainaren eta unitatearen mende egongo. Eta hori baliagarria da balio global bat lortzeko, “score” funtzio batzuk konbinatzen direnean.

Laburbilduz, guztizkoetarako eskalatuta dauden bi “score” funtzio lokal, bata batukorra eta bestea biderkatzailea, honako hauek dira, hurrenez hurren:

$$s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \frac{|x_{ij} - \tilde{x}_{ij}|}{\tilde{x}_{ij}} \text{ eta } s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \left(\max\left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}}\right) - 1\right)$$

o Itxarondako balioetarako ereduak

Oro har, itxarondako balio bat aldagai laguntzaileen eta koefizienteen funtzio bat da:

$$\tilde{x}_{ij} = f(\hat{\mu}_1, \dots, \hat{\mu}_K, z_{i1}, \dots, z_{iK})$$

Aldagai laguntzaile horiek batzuetan ateratzen dira uneko datu-basetik, baina are ohikoagoa da araztuta izan diren aurreko erregistroetatik edo inkestetatik ateratzea.

Azpitalde zehatz baten intereseko aldagaiaren batez bestekoaren edo erdiko balioaren balio zenbatetsia izan daiteke itxarondako balio bat, aldagai laguntzaileetan oinarritua. Adibidez, ekonomia-erregistroetan azpitalde bat osa dezakete industria-motak eta haren tamainak.

Aldagai laguntzaile bat intereseko aldagaiarekin oso korrelazionatuta baldin badago, zatiketa egin ohi da intereseko aldagaiaren eta aldagai laguntzailearen artean. Gero, ratio hori erkatu beharko da ratio horretarako itxarondako balioarekin. Adibidez, demagun gure aldagai laguntzailea langileen kopurua dela, eta intereseko aldagaia fakturazioa dela. Ratioa fakturazioa langile bakoitzeko izango da. Fakturazioa oso aldakorra izan daiteke establezimenduen artean, baita industria-mota berean ere, baina langile bakoitzaren araberrako ratioak aldakortasun askoz ere txikiagoa izan ohi du.

“Score” funtzioen bi aldagaien ratioa erabiltzen denean, arrisku-faktorean x_{ij} eta $\tilde{x}_{ij}$

ordeztu, eta balio gordina eta itxarondako balioa $\left(\frac{x_{ij}}{y_{ij}}\right)$ eta $\left(\frac{\tilde{x}_{ij}}{y_{ij}}\right)$ ezartzen dira.

Berrir, ratorako itxarondako balioa izan daiteke lehengoko aldi bateko batez bestekoa edo erdiko balioa, eta hobe azpitalde homogeen batena baldin bada.

Oro har, itxarondako balioetarako praktikan erabiltzen diren ereduak ez dira oso sofistikatuak. Horregatik, iragarpenak ez dira guztiz zehatzak. Hala eta guztiz ere, iragarpen horiek baliagarriak izan ohi dira. Izan ere, mikrohautaketaren helburua, hain zuzen, arazteko bidaltzen diren erregistroak behar bezala hautatzea da; eta ez, ordea, erregistroen iragarpen zehatza lortzea (Lawrence and McKenzie, 2000).

o “Score” funtzioak, luzera-datuekin

Aldizka jasotzen diren inkestetan eta erregistroetan, ohikoa da aurreko denboraldietako balioak aldagai laguntzaile gisa erabiltzea. Honako formula honek erakusten du aurreko denboraldietako balioak itxarondako balio gisa erabiltzen dituen arrisku-faktorea, ratioan oinarritua. Hidioglou-k eta Berthelot-ek proposatu zuten (Hidioglou and Berthelot, 1986):

$$R_{ij} = \max \left(\frac{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}} \right)}{\left(\frac{\tilde{x}_{ij,t}}{\hat{\tilde{x}}_{ij,t-1}} \right)}, \frac{\left(\frac{\tilde{x}_{ij,t}}{\hat{\tilde{x}}_{ij,t-1}} \right)}{\left(\frac{x_{ij,t}}{\hat{x}_{ij,t-1}} \right)} \right) - 1$$

Honako hauek dira osagaiak: x_{ij} intereseko aldagaiaren balioa $x_{ij,t}$ da, i erregistrarako t uneko denboraldian; eta unitate berak aurreko denboraldian izan zuen balio araztua $\hat{x}_{ij,t-1}$ da. Aldaketarako itxarondako balio gisa proposatzen dute erregistro guztien aldaketen erdiko balioa, baina horretarako, beharrezkoa da datu guztiak bilduta edukitzea. Horretarako, beste aukera bat izan daiteke aurreko denboraldietako aldaketen erdiko balioa erabiltzea: esaterako, $t-2$ eta $t-1$ artean; betiere, baldin eta aldaketa hori $t-1$ y t aldaketaren parekoa dela jotzen bada.

Eragina kalkulatzekoan, aurreko denboraldiko informazioa ere kontuan har daiteke:

$$F_{ij,t} = [\max(x_{ij,t}, \hat{x}_{ij,t-1})]^c$$

$0 \leq c \leq 1$ erabiliz. Formulazio hori dela eta, c erabiltzen da eragin-parametroaren garrantzia kontrolatzeko. Adibidez, azterlan batek (Latouche and Berthelot, 1992) 0,5eko c balioari zentzudun iritzi zion jasotako datuetan. Bestalde, gehieneko funtzioak ziurtatzen du $x_{ij,t}$ balioan errore bat eginez gero, azpizenbatespen bat izanda ere, gutxienez, aurreko denboraldiko bertsio araztuaren eragin bera $\hat{x}_{ij,t-1}$ sortuko duela.

“Score” funtzioa eraikitzeke beste estrategia batzuk

“Score” funtzioak kalkulatzeko azaldu ditugun moduak oinarrituta daude balio gordinaren eta itxarondako balioaren arteko desbiderapenean. Funtzio horiek dira estatuko estatistika-institutuetan gehien erabiltzen direnak. Hori horrela bada ere, beste estrategia-mota batzuk ere proposatu dira, nahiz eta oraindik “ohiko” estrategia bezain onartuak ez izan.

o Erroredun datuetarako eredu parametrikokoak

Aukera bat eredu parametrikoa zehaztea da, eta eredu horretan datuetan egon daitezkeen erroreak kontuan hartzea. Eredu horrek jotzen du erroredun datuak eta errorerik gabeko datuak iturri ezberdinetatik datozela. Estrategia hori zenbait adituk proposatu dute (Ghosh-Dastidar and Schafer, 2006), (Di Zio, Guarnera and Luzi, 2008) eta (Bellisai et al., 2009). Egile horiek ontzat hartu dute datu zuzenak banaketa normal batetik datozela, batez bestekoa μ dela, eta bariantza σ^2 dela; eta datu okerrak banaketa normal batetik datozela, batez besteko bera dutela, baina bariantza $c > 1$

faktore batek hanpatuta dagoela. Hipotesi horiek eredu normal kutsatua sortzen digute, honako dentsitate-funtzio hau duena:

$$f_x = \pi N(\mu, \sigma^2) + (1 - \pi) N(\mu, c\sigma^2)$$

errorerik gabeko datuen proportzioaren probabilitatea π da. Eredu hori erabiliz gero, baldintzazko probabilitatea, $\hat{\pi}_i$, zenbatetsi daiteke. Probabilitate horrek zehazten du erregistro bat erroretik gabekoa den, behatutako balioaren arabera: $\Pr(x_i = x_i^* | x_i)$. Behatutako datuen baldintzazko probabilitate horri ondorengo probabilitatea esaten zaio. Ebakipuntutik beherako balioak ezohikotzat joko dira eta arazteko bidaliko dira.

Aurreko eredu hori zabal daiteke, balio galduak onartzen badira. Horretarako, aldagaia logaritmoaren bidez eralda daiteke, simetrikoagoa izan dadin, eta normaltasun-hipotesia errealistagoa izan dadin. Era berean, μ batez bestekoaren balioa aldatzea onartzen duten aldagaikideak erabiliz zabal daiteke.

o **Balioztatze-*editei* lotutako estrategia**

Hau da proposatu den beste estrategia bat (Hedlin, 2003): erregistro batek balioztatze-*editak* zenbateraino urratu dituen zehaztea. Bestela esanda, zenbat *edit* ez duen betetzen, eta *editak* ez betetzeko aldea zenbatekoa den argitzea. Estrategia horrek oinarri hartzen du errore eragingarriek *edit* batzuk zapalduko dituztela; edo errorea bera handi samarra izango dela. Azterlanaren amaieran, (Hedlin, 2003), Hedlinek frogatu zuen “*score*” funtzioa erabiltzeak emaitza hobeak ematen zituela, balioztatze-*editei* lotutako estrategiarekin alderatuta.

o **Iragarpen-ereduaren estrategia**

Estrategia horrek proposatzen du eredu bat eraikitzea, zeinek azterketa-aldagaien eragingarriak diren errorearen agerpena eta tamaina erregistro bereko beste aldagai iragarle batzuekin erlazionatzen dituen. Estrategia horrek entrenamendu-datu batzuk behar ditu. Datu horiek jatorrizko datu gordinak eta datu araztuak eduki behar dituzte. Entrenamendu-datu horiek erabiliz gero, erregistro bakoitzaren arazketak guztizko zenbatespenean izan dezakeen eragina kalkula daiteke.

Erregistro bakoitzaren eragina lortutakoan, π “errorearen probabilitatea” lor dezakegu: adibidez, erregistro bakoitza sei kategoriadun aldagai batean sailkatzen badugu, Van Lancen-ek egin zuen bezala (Van Lancen, 2002). Lehenengo kategorian errorerik gabeko erregistroak sartuko dira, bigarrenetik bosgarrenera erroredun erregistroen %20 sartuko dira, eta azken kategorian errore eragingarrienak. Mota bakoitzari π probabilitatea esleitu behar zaio: 0; 0,2; 0,4; 0,6; 0,8; eta 1. Probabilitate hori iragartzeko, erregresio logistikoko eredua erabil daiteke. Bertan aldagai iragarleak sartuko dira, bai eta entrenamendu-datuak erabili ere. Ereduaren parametroak kalkulatu ondoren, uneko datuetan aplikatuko dira, eta erregistro bakoitzeko zenbatetsiko da errore eragingarria izateko probabilitatea.

Ez da frogatu estrategia hori “*score*” funtzioen kalkulua baino hobe denik (de Wall, Pannekoek and Scholtus, 2011).

“Score” funtzio globala

Erregistro oso bat hautatu eta araztu ahal izateko, beharrezkoa da zenbait “score” funtzioaren informazioa konbinatzen duen balio bat. Balio horri esaten zaio puntuazio globala edo “score” globala. Puntuazio horrek erakutsi behar du erregistro osoa arazteko garrantzia. Puntuazio lokalak konbinatu ahal izateko, garrantzitsua da “score” funtzio lokalak alderagarriak dien eskaletan neurtuta egotea. Horretarako, ohikoena da balio lokalak eskalatzea: zatiketa egin behar da balio horien eta haien guztizkoaren edo itxarondako guztizkoaren artean.

“Score” funtzio lokalak, aldezturik eskalatuak, konbinatzeko aukerarik ohikoena, honako hauek dira:

- “Score” funtzio lokalen batuketa (Latouche and Berthelot, 1992):

$$S_i = \sum_{j=1}^J s_{ij}$$

- “Score” funtzio lokalen maximoa (Latouche and Berthelot, 2000):

$$S_i = \max(s_{ij})$$

- Aurreko biak barne hartzen dituen proposamen bat (Hedlin, 2008):

$$S_i^{(\alpha)} = \left(\sum_{j=1}^J s_{ij} \right)^{1/\alpha}$$

Honako hauek dira osagaiak: $S_i^{(\alpha)}$ da α parametroaren arabera balio globala da; j -garren “score” funtzio lokalaren balioa s_{ij} da; eta J balio lokalen kopurua da.

Hau da funtzio lokalak konbinatzeko lehenengo moduren desabantaila: desbiderapen handia duten baina gutxi diren erregistroek baino lehentasun handiagoa izaten dute desbiderapen neurritsuko erregistroek, asko baldin badira. Horren aldean, honako abantaila hau dauka bigarren kasuak: desbiderapen handiak egonez gero, kasu horiek lehenetsiko dira. Hala eta guztiz ere, aukera horrek ezin izango ditu hauexek bereizi: puntuazio lokal bakar handia duten erregistroak eta puntuazio lokal handi ugari dituzten erregistroak. Azken aukera horretan, α balioak zehazten du balio lokalen eragina balio globalean. $\alpha = 1$ balio batek lehenengo aukera dakar berekin, hots, “score” funtzioen batuketa. α balio batek, ∞ baliotik gertukoa, bigarren aukera ondorioztatzen du, hots, “score” funtzio lokalen maximoa.

Badago beste aukera bat: aldagai bakoitzak balio globalean izango duen pisu espezifiko hautatzea, eman nahi zaion garrantziaren arabera. Adituek eslezi dezakete pisu hori, eta, adibidez, honako hauen artekoa izan: 0, 1, 10 eta 100.

Ataria zehaztea

“Score” funtzio baten azkeneko helburua zehaztea geroago araztu beharreko erregistroak hautatzea da. Arazketak datu guztiak bildu arte itxaron ahal badu, arazketa eten liteke intereseko parametroak jada funtsean aldatzen ez direnean. Oro har, arazteko modu horrek epe luzeak eskatzen ditu, datuen kopurua handia eta aldakorra baldin bada. Datuen arazketa bilketa-fasean hasi ahal izateko, beharrezkoa da erabakia hartzea, erregistro bakoitzaren “score”-an oinarrituta, beste erregistro batzuenarekin erkatu beharrik gabe. Horregatik, atari bat zehaztu behar da. Hortaz, erregistro baten “score” funtzio globalak atariko balioa gainditzen baldin badu, erregistro hori araztu beharko da.

Hauxe da atari hori zehazteko modu ohikoena: simulazio-azterlan bat egitea, zenbait atari-balioren efektua aztertzeko; bestela esanda, erregistro gehiago edo gutxiago arazteak intereseko parametroen gainean duen efektua ikertzea. Simulazio-azterlan horrek jatorrizko datu gordinak erabiltzen ditu, eta datu horiek eskuz araztu behar dira.

Simulazio-azterlana egiteko, erregistroak ordenatzen dira “score” funtzio globalean duten balioari jarraituta. Beraz, lehen p % erregistroak hautatu behar dira, eskuz araztu beharrekoak. Eta hori egiteko, erregistro horiek kendu, eta haien ordeaz, araztutako baseko datuak erabili beharko ditugu. Araztutako p % erregistroen azpitaldeari E_p esaten zaio. Urrats horiek p -balioen hein batean errepikatuko dira. Orduan, p % erregistro araztu dituen base berrian oinarrituta, intereseko parametroa zenbatetsiko dugu; eta gero, parametro hori eta erabat araztutako datu-basean zenbatetsitako parametroa erkatu beharko ditugu. Zenbatesleen arteko diferentzia erlatiboa sasi-alborapena deitzen da:

$$AB_j(p) = \frac{1}{\hat{X}_j} \left| \sum_{i \in E_p} w_i (x_{ij} - \hat{x}_{ij}) \right|$$

Balio horri esaten zaio sasi-alborapen absolutua: erroreak baino araztuko ez bagenitu, benetako alborapena izango litzateke, erregistro guztiak ez arazteagatik lorturikoa. Baina, araztutako datuak datu zuzenak direla ziurtatzerik ez dagoenez, alborapen hori egiazko baliora hurbildu baino ez da egiten; hortaz, sasi-alborapena da.

Erregistroen p % arazteagatik sorturiko sasi-alborapena jo daiteke, erregistroen gainerako $1 - p$ % araztuz gero, zenbateslearen zehaztapena zenbateraino handitu den neurtzeko baliabidetzat. Beraz, p balioen hein baterako sasi-alborapena kalkulatzan badugu, p balioen arabeko hobekuntza zehaztapenean nolakoa izan daitekeen jakingo dugu. Hortaz, p balioen puntu batean erabaki liteke erregistroen arazketa bertan behera uztea, jada intereseko parametroan ia hobekuntzarik sortzen ez duelako.

Azkenik, simulazio-azterketaren bidez, arazketa selektiboko prozesuaren eraginkortasuna egiazta daiteke. “Score” funtzio globalean balio handiak dituzten erregistroek benetako errore eragingarriak dituztenetz ziurtatzeko aukera ematen du,

baita alderantzizkoa ere, hau da, balio global txikiko erregistroek errore ez-eragingarriak dituztenetz.

4. Makrohautaketa

Datuen bilketa bukatuta edo ia bukatuta dagoenean, zer erregistro araztu hautatzea da makrohautaketaren funtsa.

Agregazio-metodoa

Datu guztiak bildu ondoren, intereseko aldagaien agregatu nagusiak kalkulatu behar dira. Agregatu horiek eta itxarondako balioak (adibidez, aurreko denboraldietako datuetan oinarrituta) oso bestelakoak baldin badira, berrikusi egin beharko dira.

Agregatuak eta itxarondako balioak oso bestelakoak izan daitezke, hainbat arrazoi direla eta:

- Baliteke datuek errore eragingarriak edukitzea.
- Baliteke diseinuan erabilitako pisuak zuzenak ez izatea.
- Baliteke ezusteko aldakuntzak, errealak direnak, gertatu izana.

Hauxe da makro mailako “score” funtzio baten adibidea:

$$S_j = X_j - \tilde{X}_j$$

Honako hauek dira osagaiak: x_j aldagairen zenbatesle agregatua X_j da, eta araztu gabeko datuetan oinarrituta dago; eta zenbatesle agregatuaren itxarondako balioa $\tilde{X}_j$ da.

Halaber, agregatuaren eta itxarondako balioaren arteko diferentzia erlatiboa kalkula dezakegu:

$$S_j = \frac{X_j - \tilde{X}_j}{\tilde{X}_j}$$

Kasu batzuetan, agregatuak berak bereizita erabiltzea baino eraginkorragoa izan ohi da agregatuen arteko ratioak erabiltzea.

$$S_j = \frac{X_j}{X_k} - \left(\frac{\tilde{X}_j}{\tilde{X}_k} \right)$$

Edo kopuru erlatiboak erabiltzea:

$$S_j = \frac{\frac{X_j}{X_k} - \left(\frac{\tilde{X}_j}{\tilde{X}_k} \right)}{\left(\frac{\tilde{X}_j}{\tilde{X}_k} \right)}$$

Kasu horren adibide izan daiteke fakturazioaren guztizkoaren eta kostuen guztizkoaren arteko ratioa; edo soldaten guztizkoaren eta industria bateko langile kopuru osoaren artekoa.

Agregatuaren bariantza kontrolatzeko modu bat zera da: agregatuetan edo haien ratioetan jasotako diferentziaren eta desbiderapen estandar erlatiboaren arteko zatiketa egitea.

$$S_j = \frac{X_j - \tilde{X}_j}{d.e.(X_j - \tilde{X}_j)} \text{ eta } S_j = \frac{\frac{X_j}{X_k} - \left(\frac{\tilde{X}_j}{\tilde{X}_k} \right)}{d.e.\left(\frac{X_j}{X_k} - \left(\frac{\tilde{X}_j}{\tilde{X}_k} \right) \right)}$$

Mikrohautaketan gertatzen zen bezala, agregatuen edo ratioen arteko diferentziak batuketa edo biderketa gisa adieraz daitezke.

Datuetan agregatu susmagarri bat hauteman ondoren, maila txikiagoko agregatuak ikertu beharko dira, adibidez, industria-mota zehatz baten agregatua. Prozesu horren amaieran, erregistroak berak gainbegiratu beharko dira, balizko errore eraginkorrak bilatzeko.

Segidan, mikrohautaketaren eta, datu-bilketa amaitutakoan, teknika horien aplikazioaren arteko diferentzia nagusiak aipatuko ditugu. Honako hauek dira:

1. Egungo datuak itxarondako balioak zehazteko informazio-iturri gisa erabil daitezke. Adibidez, egungo datuen talde homogeneo baten erdiko balioa erabil daiteke, aurreko denboraldi batekoaren orde. Egungo datu-basea oraindik araztu ez denez, garrantzitsua da erdiko balioak erabiltzea, balio ezohikoak daudenean zenbatespen sendoak baitira.
2. Zenbatesleak kalkulatzekoan, ez da beharrezkoa pisuen hurbilketa egitea v_i diseinuaren pisuak erabiliz, w_i azkeneko pisuak baititugu.
3. Ez da beharrezkoa *a priori* atari bat zehaztea, "score" funtzioen puntuazioek beraiek hurrenkera bat zehazten dutelako, eta hurrenkera horren arabera araztuko dira erregistroak. Halaber, azkeneko zenbateslea zenbat aldatzen den kontrola daiteke. Arazketa eten daiteke, zenbateslearen zehaztapenean sortzen duen hobekuntza hutsaltzat jotzen denean.

Banaketa-metodoa

Banaketa-metodoak identifikatu nahi du printzipioz zer balio ez diren erraz moldatzen behatutako banaketara. Hori egiteko, tresna grafikoak eta estatistikak neurtzeko tresnak erabiltzen dira. Ezohiko balioak, erregistro susmagarrienak, ziurtatu egin behar dira. Ezohiko balio bat okerra eta eraginkorra dela frogatzen baldin bada, zuzendu egin behar da.

Metodoa aldagai kuantitatiboei aplikatzen zaie, eta datuak normal samarrak direla hartzen du, edo, bestela, asimetrikoak direla. Eta hori betetzen ez denean, oro har, datuak eraldatzeko modu bat erabiltzen da.

Ezohiko balioak antzemateko neurri sendoa $x_{ij} - \text{mediana}(x_{ij})$ erdiko balioan oinarritzen da. Hori “score” funtzio baten antzekoa da, funtzioan itxarondako balioa erdiko balioa baita. Erdiko baliotik aldentzen diren desbiderapenak ziurtatu ahal izateko, diferentzia hori haren balio absolutuen erdiko balioaren arabera estandarizatu ohi da talde bakoitzeko erregistroetan.

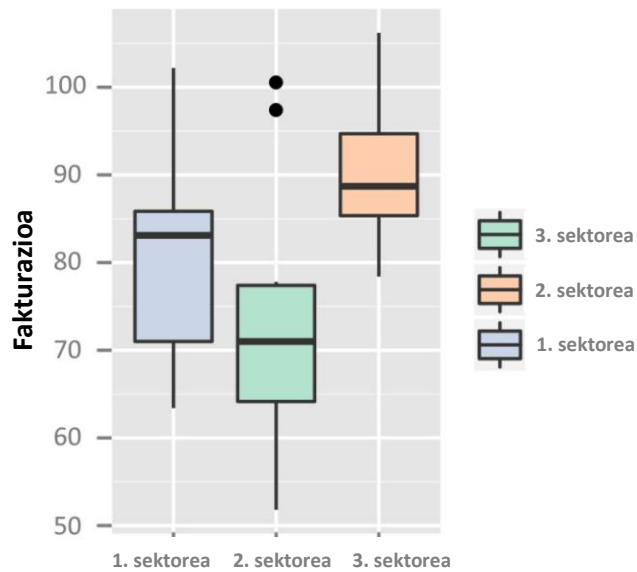
$$o_{ij,c} = |x_{ij} - \text{med}(x_{ij,c})| / (1,4826 \times \text{DAM}_c(x_{ij,c}))$$

Honako hauek dira osagaiak: c taldeko erregistroen erdiko balioa $\text{med}(x_{ij,c})$ da; eta erdiko balioaren desbiderapen absolutua $\text{DAM}_c(x_{ij,c})$ da erregistro horietan. Honela lortu da:

$$\text{DAM}_c(x_{ij,c}) = \text{med}(|x_{ij} - \text{med}(x_{ij,c})|)$$

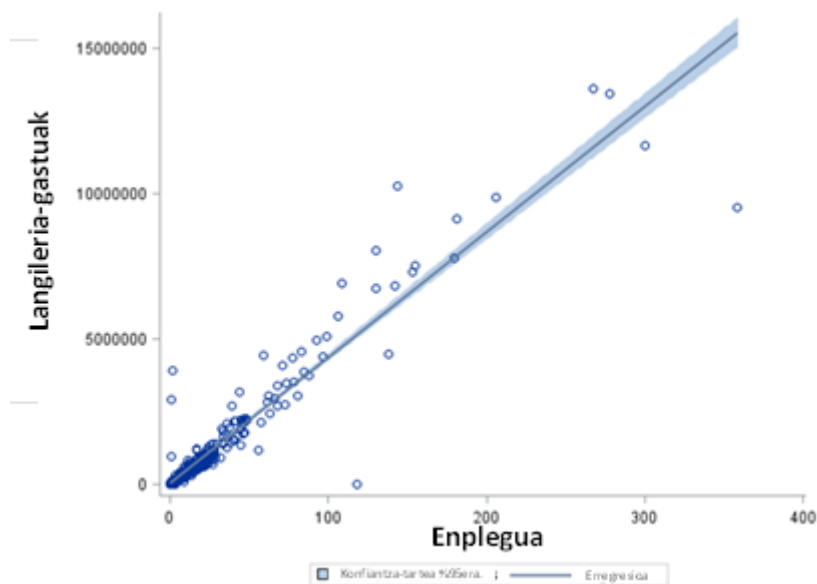
Banaketa normal batean, desbiderapen estandarra zehazteko zenbatesle sendoa $1,4826 \times \text{DAM}$ da. *Outlier*-ak antzemateko erabiltzen diren beste neurri sendo batzuk Windsor batez bestekoak edo batez besteko moztuak dira. Horrez gain, sakabanatze-neurri ez-sendoak ere erabil daitezke, hala nola bariantza edo desbiderapen estandarra.

Era berean, ohikoa da kutxa-diagramak edo *boxplot*-ak erabiltzea erdiko baliotik aldentzen diren desbiderapenak irudikatzeko. Grafiko horiek haxe erakusten dute: batetik, erregistroen %50 duen kutxa bat; eta beste batetik, lerro batzuk, zeinek lehenengo eta hirugarren kuartilei dagokien kuartilarteko tartea 1,5 bider mugatzen duten. Lerro horietatik kanpoko balioak ezohikotzat jotzen dira.



Grafiko honetan irudikatu da hiru sektoretako establezimenduen fakturazioaren banaketa, mila eurotan. Ikus bezala, 2. sektorean bi ezohiko balio daude.

Ezohiko balioak antzemateko beste teknika grafiko bat puntu-diagrama (*scatterplot*) da. Kutxa-diagramak ez bezala, puntu-diagrama bi aldagai jarraitu erkatzen direnean erabili ohi da.



Grafiko honetan irudikatu da enpleguaren eta langileria-kostuaren arteko erlazioa “makinak eta tresnak” sektorean.

5. Simulazioa

Kapitulu honetan “*score*” funtzio batzuen portaera aztertuko dugu; horretarako, zenbait errore-mota sortuko ditugu, datu-base simulatu baten bidez.

Datu-base simulatua

Merkataritza Erregistroko datuak simulatzen dituen datu-base bat osatu da. Datu-base hori sortzerakoan, hainbat ezaugarri kontuan hartu ditugu, ahalik eta errealena izan zedin.

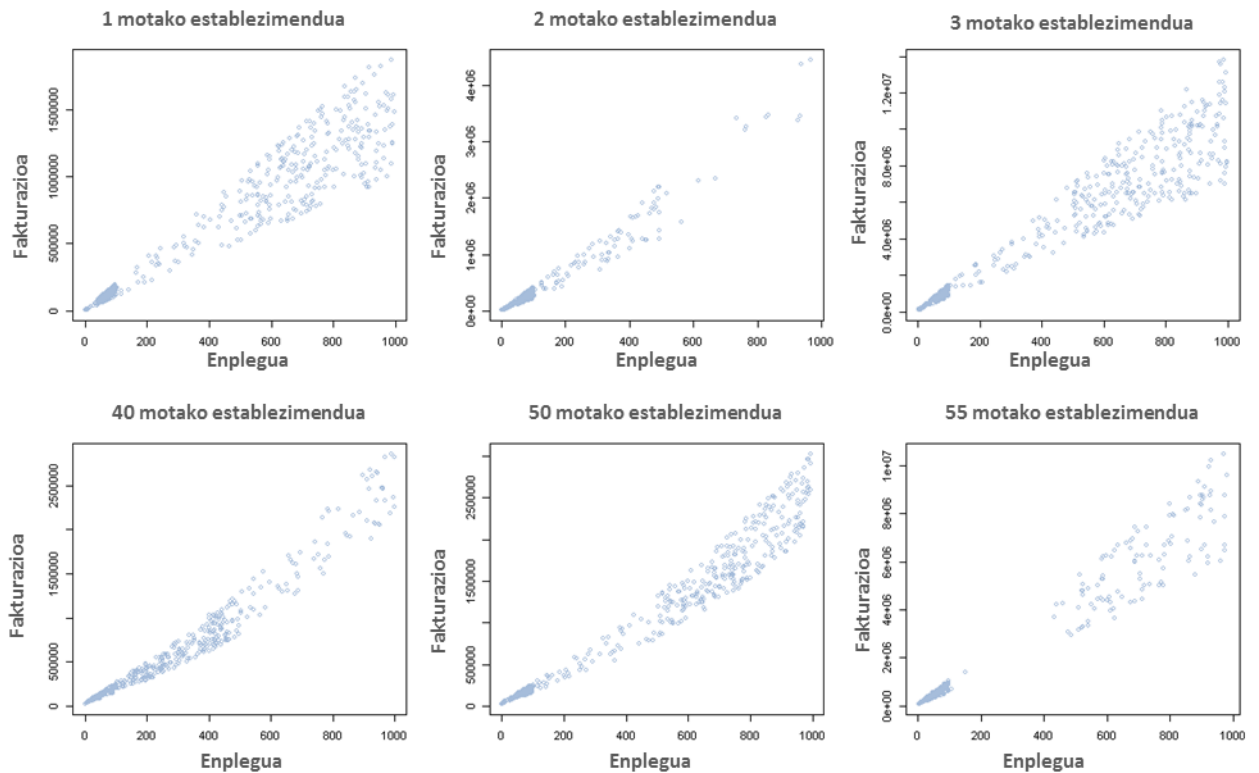
Lehenik, enpresei buruzko datuak sailkatu dira enpresen jardueraren arabera, 2009ko EJSN sailkapenari jarraituta, bai eta haien tamainaren arabera ere, langileen kopurua zein den.

Enpresa bakoitza lurralde historiko bati esleitu zaio: Bizkaian kokatzeko probabilitate hurbildua %0,50 da; Gipuzkoan %0,33; eta Araban %0,17. Simulatu den datu-basean jasotako enpresa-kopurua 36.719 da.

Enpresa bakoitzaren fakturazioa sortzeko, jarduera eta langileen kopurua aintzat hartu dira. Zenbat eta langile gehiago, orduan eta fakturazio handiagoa. Horretarako, funtzio linealak eta kuadratikoa erabili dira. Horrez gain, nolabaiteko aldakortasuna ere gehitu zaie. Establezimenduko langileen kopurua igo ahala, fakturazioaren aldakortasuna ere handituko da.

“*Score*” funtzio batzuk kalkulatzeko aurreko denboraldiko balioa zein zen jakin behar denez, aurreko balioak ere sortu ditugu fakturazio- eta langileen kopuru-aldagaiei dagokienez. Era berean, fakturazio-aldagaia ere sortu da. Kasu honetan ere aldakortasuna gehitu da, beraz, aurreko denboraldiko balioak ausazko aldakortasun handiagoa dauka, egungo balioaren aldean.

Grafiko honek erakusten du fakturazioaren eta enpleguaren arteko erlazioa zenbait enpresatan, sei jarduera-mota barne hartuta.



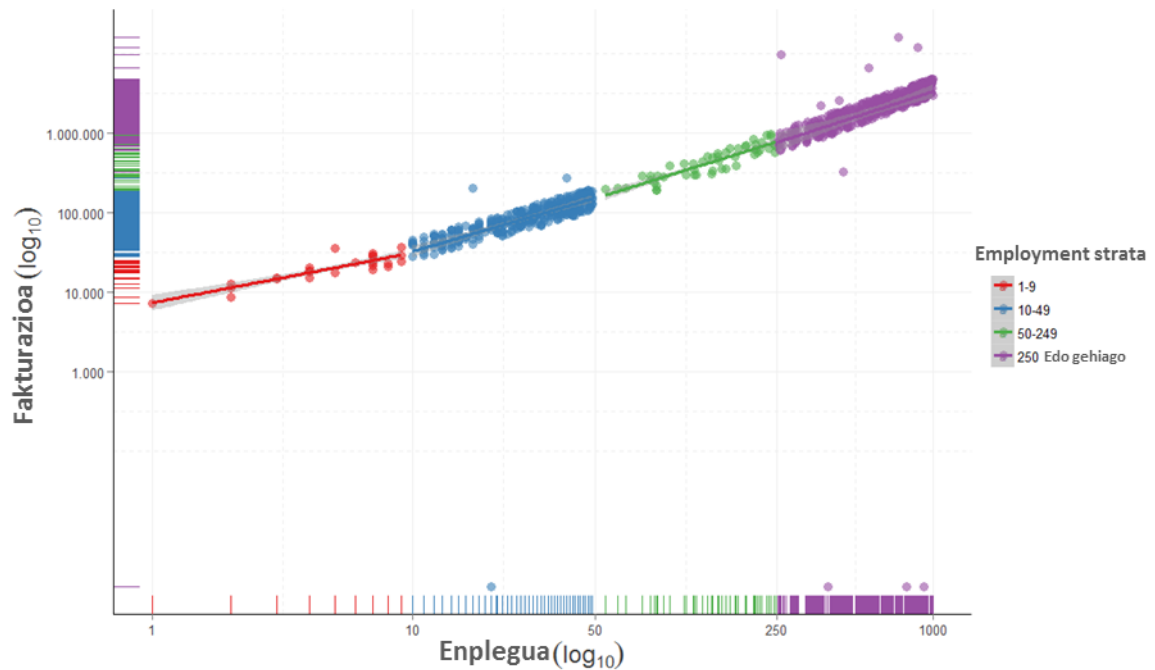
Errore-motak sortzea

Jatorrizko datu-basea sortu ondoren, hurrengo urratsean fakturazio-aldagaiari erroreak gehitu genizkion. "Score" funtzioek erroreak antzematerakoan nola funtzionatzen duten jakingo badugu, ahalik eta errealistenak izan behar dute errore horiek.

Lehenik eta behin, datu-baseko 36.719 enpresetatik 500 enpresa ausaz hautatu genituen. Ondoren, ausazko errore bat gehitu genien: banaketa normala zuen, batez bestekoa 0 izan zen, eta desbiderapen estandarrak jatorrizko fakturazioa bikoizten zuen. Eraitza negatiboa izanez gero, fakturazioa zero balioarekin parekatuko zen. Horrek sortzen dituen erroreek fakturazioa lau bider edo gehiago biderka dezakete, edo fakturazioa 0 bihurtu dezakete.

Gainera, unitate-erroreak ere gehitu dira. Ausaz, 50 enpresa hautatu ziren, eta haien fakturazioa 1.000 zenbakiz biderkatu zen. Bestalde, beste 50 enpresaren fakturazioa 1.000 zenbakiz zatitu zen.

Honako grafiko honek erakusten du jarduera-mota bateko enpresa guztien fakturazioaren eta enpleguaren arteko erlazioa. Eskala logaritmikoa erabili da enplegu-estratu bakoitza behar bezala bereizteko, eta erregresio linealeko lerroa enplegu-tarteen arabera adierazita dago. Portaera “anomaloak” antzeman daitezke.



“Score” funtzioak

Segidan jaso dira simulatutako datu-basearen bidez kalkulatu genituen “score” funtzioak. Enpresa bakoitzari funtzio bakoitzaren arabera dagokion puntuazioa kalkulatu dugu.

1. egoera: fakturazioa eta itxarondako balioa talde homogeneo batean

Baturakorra $s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \frac{ x_{ij} - \tilde{x}_{ij} }{\tilde{x}_{ij}}$	Biderkatzailea $s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \left(\max \left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}} \right) - 1 \right)$
---	---

Hauexek dira osagaiak:

x_{ij} → Fakturazioa

$\tilde{x}_{ij}$ → Fakturazioa talde homogeneo batean (jarduera eta langileen kopuru berak)

2. egoera: langileen kopuruaren eta itxarondako balioaren arteko ratioa, talde homogeneo batean

Batukorra	Biderkatzailea
$s_{ij} = \frac{w_i \left(\frac{\tilde{x}_{ij}}{\tilde{y}_{ij}} \right)}{\left(\frac{\tilde{X}_j}{\tilde{Y}_j} \right)} \times \frac{\left \frac{\tilde{x}_{ij}}{\tilde{y}_{ij}} - \left(\frac{\tilde{x}_{ij}}{\tilde{y}_{ij}} \right) \right }{\left(\frac{\tilde{x}_{ij}}{\tilde{y}_{ij}} \right)}$	$s_{ij} = \frac{w_i \left(\frac{\tilde{x}_{ij}}{\tilde{y}_{ij}} \right)}{\left(\frac{\tilde{X}_j}{\tilde{Y}_j} \right)} \times \left(\max \left(\frac{\frac{\tilde{x}_{ij}}{\tilde{y}_{ij}}}{\left(\frac{\tilde{x}_{ij}}{\tilde{y}_{ij}} \right)}, \frac{\left(\frac{\tilde{x}_{ij}}{\tilde{y}_{ij}} \right)}{\frac{\tilde{x}_{ij}}{\tilde{y}_{ij}}} \right) - 1 \right)$

Hauexek dira osagaiak:

$\frac{x_{ij}}{y_{ij}} \rightarrow$ Fakturazioa zati langileen kopurua

$\left(\frac{\tilde{x}_{ij}}{\tilde{y}_{ij}} \right) \rightarrow$ Fakturazioa zati langileen kopurua talde homogeneo batean (jarduera eta langileen kopuru berak)

3. egoera: aurreko denboraldiko balioaren eta itxarondako balioaren arteko ratioa, talde homogeneo batean

Batukorra	Biderkatzailea
$s_{ij} = \frac{w_i \left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right)}{\left(\frac{\tilde{X}_{j,t}}{\hat{X}_{j,t-1}} \right)} \times \frac{\left \left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right) - \left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right) \right }{\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right)}$	$s_{ij} = \frac{w_i \left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right)}{\left(\frac{\tilde{X}_{j,t}}{\hat{X}_{j,t-1}} \right)} \times \left(\max \left(\frac{\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right)}{\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right)}, \frac{\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right)}{\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right)} \right) - 1 \right)$

Hauexek dira osagaiak:

$\frac{x_{ij,t}}{\hat{x}_{ij,t-1}}$ → Fakturazioa t denboraldian zati fakturazio araztua $t-1$ denboraldian

$\left(\frac{\tilde{x}_{ij,t}}{\hat{x}_{ij,t-1}} \right)$ → Fakturazioa t denboraldian zati fakturazio araztua $t-1$ denboraldian
(jarduera eta langileen kopuru berak)

4. egoera: eragina, fakturazioa erabilia eta arriskua langileen kopuruaren ratioarekin

Batukorra

$$s_{ij} = \frac{w_i \tilde{x}_{ij,t-1}}{\tilde{X}_{j,t-1}} \times \frac{\left| \frac{x_{ij,t}}{y_{ij,t}} - \left(\frac{\tilde{x}_{ij,t-1}}{\tilde{y}_{ij,t-1}} \right) \right|}{\left(\frac{\tilde{x}_{ij,t-1}}{\tilde{y}_{ij,t-1}} \right)}$$

Biderkatzailea

$$s_{ij} = \frac{w_i \tilde{x}_{ij,t-1}}{\tilde{X}_{j,t-1}} \times \left(\max \left(\frac{\frac{x_{ij,t}}{y_{ij,t}}}{\left(\frac{\tilde{x}_{ij,t-1}}{\tilde{y}_{ij,t-1}} \right)}, \frac{\left(\frac{\tilde{x}_{ij,t-1}}{\tilde{y}_{ij,t-1}} \right)}{\frac{x_{ij,t}}{y_{ij,t}}} \right) - 1 \right)$$

Hauexek dira osagaiak:

$x_{ij,t}$ → Fakturazioa t denboraldian

$\tilde{x}_{ij,t-1}$ → Fakturazio araztua $t-1$ denboraldian

$\frac{x_{ij,t}}{y_{ij,t}}$ → Fakturazioa zati langileen kopurua t denboraldian

$\left(\frac{\tilde{x}_{ij,t-1}}{\tilde{y}_{ij,t-1}} \right)$ → Fakturazio araztua zati langileen kopurua $t-1$ denboraldian

5. egoera: erregresio sendotik eratorritako fakturazioa eta zenbatespena

Batukorra

$$s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \frac{|x_{ij} - \tilde{x}_{ij}|}{\tilde{x}_{ij}}$$

Biderkatzailea

$$s_{ij} = \frac{w_i \tilde{x}_{ij}}{\tilde{X}_j} \times \left(\max \left(\frac{x_{ij}}{\tilde{x}_{ij}}, \frac{\tilde{x}_{ij}}{x_{ij}} \right) - 1 \right)$$

Hauexek dira osagaiak:

x_{ij} → Fakturazioa

$\tilde{x}_{ij}$ → Fakturazioaren gaineko zenbatespena, enpresako langileen kopurua zein den. Estratu bakoitzean erabili dira erregresio sendoko metodoak.

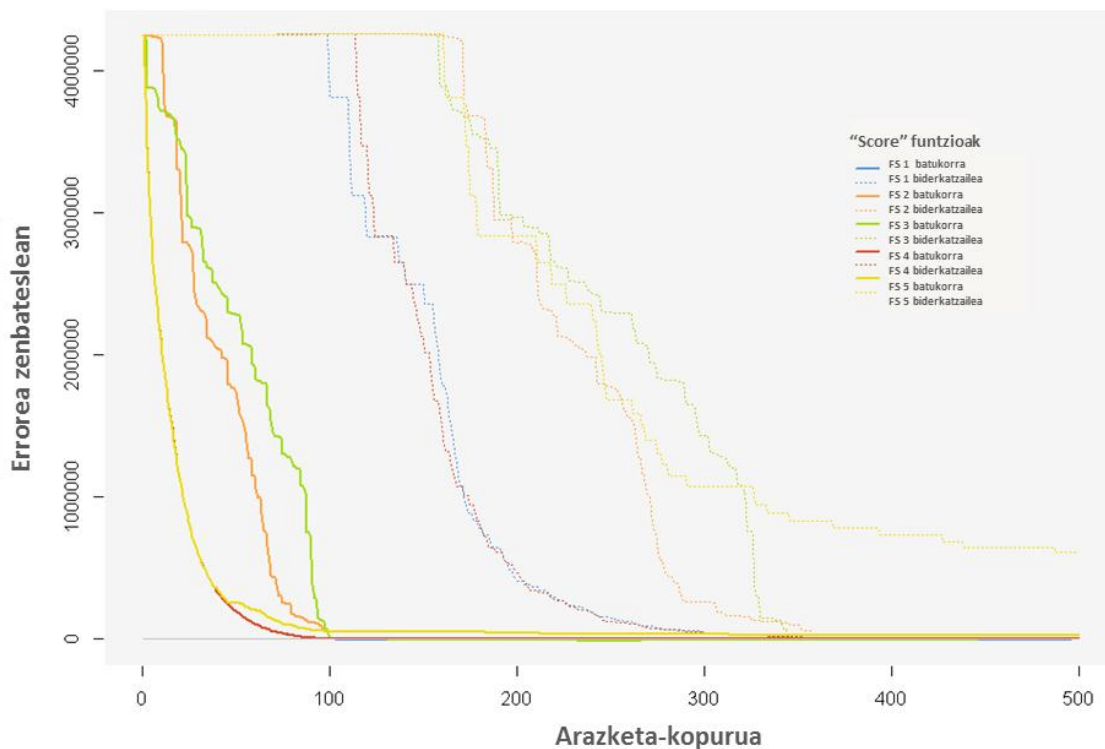
Simulazioaren emaitzak

“Score” funtzioaren bidez enpresa bakoitzari dagokion puntuazioa kalkulatu ondoren, puntuazioak ordenatu dira. Puntuazio handieneko enpresak hautatu, eta lehenik araztuko dira.

Beheko grafikoak erakusten du nola bilakatu den zenbateslearen errorea –egoera honetan fakturazioaren batez bestekoa–, enpresak araztu ahala. Errore hori errorerik gabeko fakturazioaren batez bestekoaren eta erroredun fakturazioaren batez bestekoaren arteko diferentzia gisa kalkulatu da.

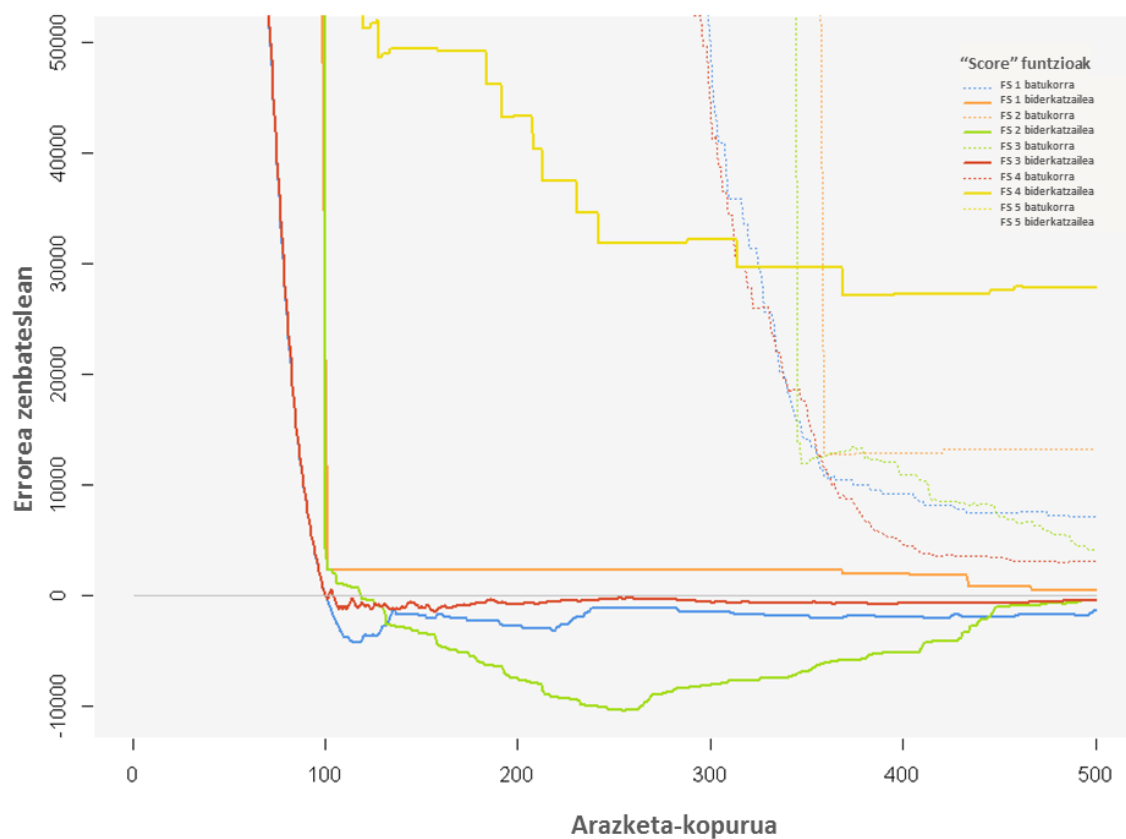
Grafikoek erakusten duten bezala, arriskuan eskala biderkatzailea erabiltzen duten “score” funtzioek eraginkortasun txikiena dute, errorea zenbateslean txikiagotzeko tarte handiagoa behar dutelako.

“Score” funtzio batukorren artean, badirudi 1 eta 4 onenak direla. Funtzio horietan “eraginkortasun” terminoak aintzat hartzen du enpresa bakoitzean itxarondako fakturazioak zenbatesle orokorrari egiten dion ekarpena.



Honako grafiko honetan *zoom*-a egin dugu zenbatesleko errorearen ardatzaren gainean. Horrela, fakturazioaren errore nagusiak ezabatu ondoren, “score” funtzioek nola funtzionatzen duten aztertu ahal izango dugu.

Grafikoak erakusten du 4. “score” funtzioa lehena dela errore nulura hurbiltzen, eta ia denbora osoan balore horretatik gertu egoten dela. Hori dela eta, gure simulazioan funtzionamendurik onena izan duen “score” funtzioa hauxe da: haren “eraginkortasun” terminoak enpresa bakoitzaren fakturazioak zenbatesle orokorrari egiten dion ekarpena kontuan hartzeaz gain, arrisku-terminoa fakturazioaren eta enplegatu-kopuruaren arteko ratioa aintzat hartuta kalkulatzen duena.



6. SAS Makroa

Segidan, SAS makroak aurkeztuko ditugu. Horrek bide emango digu datu-base batean arazketa selektiboa egiteko.

Informazio gehiago eskuratu nahi izanez gero, honako eskuliburu hau kontsulta daiteke: *Manual de usuario de la aplicación SAS:Depuración Selectiva*.

FUNCION_SCORE makroa

FUNCION_SCORE SAS makroaren bidez, aldagai baterako hiru “score” funtzio kalkula daitezke erregistro batean.

- I. motako “score” funtzioa: erreferentziatzen hartzen ditu aurreko denboraldiko balioa eta erregistroaren eragina zehaztu dugun estratuan.
- II. motako “score” funtzioa: erreferentziatzen hartzen ditu zehaztu dugun erdiko balioa eta erregistroaren eragina estratu horretan.
- III. motako “score” funtzioa: erregresio sendoaren bitartez zenbatetsi den balioa zehaztu dugun estratuan zein den eta erregistroaren eragina estratu horretan erreferentziatzen hartzen ditu.
- IV. motako “score” funtzioa: erdiko balioa erreferentziatzen hartzen du; eta zatitzaile hartzen ditu zehaztu dugun estratuaren kuartil arteko zabalera eta erregistroaren eragina estratu horretan.

Sarrerako datuak

SAS datu-multzoak gutxienez hauexek eduki behar ditu:

- t unean araztu nahi dugun aldagaia;
- $t-1$, araztu nahi dugun aurreko denboraldiko aldagaia.
- Aukerakoa: arazketa-aldagaiari lotzeko aldagai bat, erregresioa egin ahal izateko.
- Aukerakoa: estratuak identifikatzeko aldagaiak. Estratu horietan erdiko balioa kalkulatu dugu, edo erregistro bakoitzaren eragina “score” funtzio bakoitzean.

Makroaren sintaxia

Hona hemen beharrezko argumentuen deskribapen laburra:

- *dataset*= SAS datu-multzoa, non arazketa selektiboa egingo baita.
- *var*= t unean araztu nahi dugun aldagaia.

- `var_ant`= $t-1$ aurreko denboraldiko aldagaia, araztu nahi duguna.
- `var_reg`= arazketa-aldagaiari lotzeko aldagai bat, erregresioa egin ahal izateko.
- `tipolestrato`= estratu horretan erregistroaren eragina kalkulatu da I. motako "score" funtziorako.
- `tipollestrato`= estratu horretan erregistroaren eragina kalkulatu da II. motako "score" funtziorako.
- `tipolllestrato`= estratu horretan erregistroaren eragina kalkulatu da III. motako "score" funtziorako.
- `tipolV.estrato`= estratu horretan erregistroaren eragina kalkulatu da IV. motako "score" funtziorako.
- `varpositiva`= erregresioa zenbatesteko, aldagaiak balio positiborik baino hartuko ez duen zehaztu behar dugu. "T" balioa lehenetsita dago, eta positiboa dela adierazten du. Kontrakoa "F" izango litzateke.
- `n_estrato`= erregistroen gutxienerako kopurua estratu bakoitzean, erregresioa sendoa zehaztu ahal izateko.
- `fscore_1` = I. motako "score" funtzioa kalkulatzeko aukera ematen du. "T"= bai; eta "F" =ez.
- `fscore_2` = II. motako "score" funtzioa kalkulatzeko aukera ematen du. "T"= bai; eta "F" =ez.
- `fscore_3` = III. motako "score" funtzioa kalkulatzeko aukera ematen du. "T"= bai; eta "F" =ez.
- `fscore_4` = IV. motako "score" funtzioa kalkulatzeko aukera ematen du. "T"= bai; eta "F" =ez.

FS_GLOBAL makroa

FS_GLOBAL SAS makroaren bidez, "score" funtzio globala, *fs_global*, kalkula daiteke. Funtzio horrek "score" funtzio lokalak, aurretik kalkulatuak, konbinatzen ditu.

Sarrerako datuak

SAS datu-multzo bat behar dugu: bertan *FUNCION_SCORE* makroaren bitartez, gutxienez, aldagai baterako "score" funtzio lokalak kalkulatu behar izan dira.

Makroaren sintaxia

Hona hemen beharrezko argumentuen deskribapen laburra:

- `dataset`= SAS datu-multzoa, non arazketa selektiboa egingo baita.
- `var`= t unean araztu nahi dugun aldagaia.

- vartext= aldagaiaren literalak.
- max= hiru "score" funtzioaren maximoa kalkulatzeko. "T" balioa lehenetsita dago. Hiru "score" funtzioaren batuketa kalkulatzeko, max="F" ezarri behar dugu, besterik gabe.
- w1= I. motako "score" funtzioaren pisua. 1 lehenetsita dago.
- w2= II. motako "score" funtzioaren pisua. 1 lehenetsita dago.
- w3= III. motako "score" funtzioaren pisua. 1 lehenetsita dago.
- w4= IV. motako "score" funtzioaren pisua. 1 lehenetsita dago.
- pisuak= behaketa bakoitzaren pisua populazioan. 1 lehenetsita dago.

7. Ezarpen praktikoa zerbitzu-estatistika deituriko jardueran

Euskal Estatistika Erakundearen zerbitzu-estatistika deritzon jarduera berria egiteko, batetik, establezimenduei bereziki zuzendutako galdera-sorta batetik ateratako datuak erabili dira; eta, bestetik, hiru administrazio-erregistroetatik etorritako datuak. Honako hauek dira aipatutako erregistroak: Merkataritza Erregistroa, Kooperatiben Erregistroa eta Elkarteen eta Fundazioen Erregistroa.

Erregistro-informazioa erabili ahal izateak zenbatespenak hobetu ditu, informazio-kopuru handiagoa eskuratu baita. Aldi berean, landu beharreko datuen kopurua hain handia izateak hainbat zailtasun-mota eragin ditu. Merkataritza Erregistroa datu gehien eman dituen iturria da; zehatzago esanda, 2012ko Zerbitzu Estatistika egiteko, 16.251 merkataritza-enpresatatik etorritako datuak erabili dira.

Horregatik, beharrezkoa da arazketa-prozedurak eraginkorrak eta fidagarriak izatea: adibidez, datu-baseen arazketa selektiboa, hain zuzen ere.

Arazketa selektiboaren ezarpen praktikoa

Establezimenduei bereziki zuzendutako galdera-sortatik etorritako informazioa eta administrazio-erregistroetatik lorturiko informazioa konbinatzen dituen datu-basean erabili da arazketa selektiboa egiteko makroa. Hauexek dira iturri gisa erabilitako erregistroak: Merkataritza Erregistroa, Kooperatiben Erregistroa eta Elkarteen eta Fundazioen Erregistroa.

Bestalde, honako aldagai hauek erabili ditugu: negozioen kopuruaren zenbateko garbia; faktoreen kostuari erantsitako balioa eta langileria-kostua. Horiek hartu dira aldagai eskuragarri nagusiak. Gainera, landun langileria-aldagaiarekin lotura handiena dutenak dira.

○ **Negozioen kopuruaren zenbateko garbiaren aldagaia araztea**

Makroa erabiltzerakoan, honako aukera hauek hautatu ziren. Lehenik eta behin, aldagai eta egoera hauek nolakoak ziren, II. eta III. “score” funtzioak egokienak zirela hartu genuen. Funtzio horiek hurrenez hurren bilatzen dituzte estratuaren erdiko baliotik edo erregresio sendoan zenbatetsitako baliotik aldentzen diren erregistroak. Doikuntza-aldagaia enplegua izan da.

Establezimendu bakoitzari buruzko aurreko denboraldiko daturik ez geneukanez, I. motako “score” funtzioa ez genuen kalkulatu, aurreko denboraldiko balioarekiko desbiderapena kontuan hartzen baitu. IV. “score” funtzioa ez kalkulatzeari erabaki genuen. Funtzio horrek ematen dio garrantzi handiena estratu bakoitzaren aldagarritasunari.

“Score” funtzioak honako estratu honetan kalkulatzeari erabaki genuen: EJSN kodea, bi digitukoa, eta enplegu-estrata konbinatzen dituenean. Erregresio sendoa zenbatetsi ahal izateko, erregistro bakoitzean gutxienez 20 establezimendu hautatu ziren; eta “score” funtzioak global batean konbinatzeko, haien arteko maximoa erabili zen.

Adibide moduan, beheko taulan jaso dira arazketa-makroaren emaitzetatik eratorritako lehenengo 10 establezimenduak.

7.1. taula. Negozioen kopuruaren zenbateko garbiaren aldagaia araztea

Establezimendua	Arrazoa	Enplegua	EJSN09	Enplegu-estrata	Negozioen kopuruaren zenbateko garbia	Erdiko balioa	Zenbatespena	sf globala
a1	2.- estratu-balorea	3	5229	1	3.470.393	25.344		1,530
a2	3.- erregresio-balorea	1	7022	2	20.881.251	105.883	79.446	0,486
a3	3.- erregresio-balorea	2	5221	1	1.244.845	25.344		0,479
a4	3.- erregresio-balorea	1	7112	1	423.372	44.925	45.666	0,178
a5	2.- estratu-balorea	1	6910	1	308.531	41.729	40.776	0,128
a6	3.- erregresio-balorea	1	7111	1	284.882	44.925	45.666	0,111
a7	2.- estratu-balorea	1	6910	1	288.304	41.729	40.776	0,110
a8	3.- erregresio-balorea	1	6621	1	170.406	32.962		0,104
a9	3.- erregresio-balorea	2	7022	2	11.482.794	105.883	119.914	0,097
a10	3.- erregresio-balorea	2	7022	2	11.443.789	105.883	119.914	0,096

7.1 taulak erakusten duen bezala, establezimendu horiek biziki aldentzen dira haien estratuaren erdiko baliotik edo estratu horretan itxaron daitekeen baliotik, langile-kopuru horren arabera. Establezimendu horietan oso handia da negozioen kopuruaren zenbateko garbia, sektore bereko edo langile-kopuru bereko establezimenduekin erkatuta. Langileen kopuruaren arabera doikuntza egin arren, itxarondako balioa nabarmen gaitzen du.

Estratuko establezimenduen kopurua 20tik beherakoa baldin bazen, estratu horietan ez zen baliorik zenbatetsi negozioen kopuruaren zenbateko garbirako.

- **Negozioen kopuruaren zenbateko garbiaren aldagaia araztea pertsona bakoitzeko**

Atal honetan negozioen kopuruaren zenbateko garbiaren ratioa araztuko dugu, establezimenduko langileen kopurua aintzat hartuta. Horretarako, estratuaren erdiko balioa soilik hartuko dugu aintzat, eta establezimendu guztiei pisu bera emango diegu.

7.2. taula. Negozioen kopuruaren zenbateko garbiaren aldagaia araztea pertsona bakoitzeko eta establezimendu bakoitzerako.

Establezimendua	Arrazoa	Enplegua	EJSN09	Enplegu- estratua	Negozioen kopuruaren zenbateko garbiaren ratioa pertsona bakoitzeko	Erdiko balioa	sf globala
b1	2.- estratu-balorea	1	7022	2	20.881.251	65.193	0,213
b2	2.- estratu-balorea	3	5229	1	1.156.798	25.344	0,062
b3	2.- estratu-balorea	18	9001	4	807.432	45.749	0,031
b4	2.- estratu-balorea	2	5221	1	622.423	25.344	0,017
b5	2.- estratu-balorea	2	7022	2	5.741.397	65.193	0,016
b6	2.- estratu-balorea	2	7022	2	5.721.895	65.193	0,016
b7	2.- estratu-balorea	1	7112	2	5.311.000	70.454	0,014
b8	2.- estratu-balorea	108	9312	7	438.467	40.003	0,014
b9	2.- estratu-balorea	1	7219	2	1.305.609	65.612	0,010
b10	2.- estratu-balorea	1	6820	2	4.266.072	45.036	0,010

Negozioen kopuruaren zenbateko garbiaren ratioa eta estratuaren portaera oso bestelakoak direla ondorioztatu da; hortaz, berrikusi behar da.

- **Faktore-kostuari erantsitako balioaren aldagaia araztea**

Aldagai hori arazteko, negozioen kopuruaren zenbateko garbia arazteko hautatu ziren aukerak berriro hartu ziren. Lehenik eta behin, establezimendu bakoitzari buruzko aurreko denboraldiko daturik ez geneukanez, I. motako “score” funtzioa ez genuen kalkulatu, aurreko denboraldiko balioarekiko desbiderapena kontuan hartzen baitu. Halaber, berriro erabaki genuen IV. “score” funtzioa ez kalkulatzeko. Funtzio horrek ematen dio garrantzi handiena estratu bakoitzaren aldagarritasunari.

Era berean, balio erantsia kalkulatzeko, II. eta III. “score” funtzioak egokienak zirela hartu genuen. Funtzio horiek hurrenez hurren bilatzen dituzte estratuaren erdiko

baliotik edo erregresio sendoan zenbatetsitako baliotik aldentzen diren erregistroak. Doikuntza-aldagaia enplegua izan da.

“Score” funtzioak honako estratu honetan kalkulatzeko erabaki genuen: EJSN kodea, bi digitukoa, eta enplegu-estratua konbinatzen dituenean. Erregresio sendoa zenbatetsi ahal izateko, erregistro bakoitzean gutxienez 20 establezimendu hautatu ziren; eta “score” funtzioak global batean konbinatzeko, haien arteko maximoa erabili zen.

Beheko taulak adierazten ditu arazketa-makroaren emaitzetatik eratorritako lehenengo 10 establezimenduak.

7.3. taula. Faktore-kostuari erantsitako balioa araztea.

Establezimendua	Arazoia	Enplegua	EJSN09	Enplegu-estratua	Faktoreen kostuari erantsitako balioa	Erdiko balioa	Zenbatespena	sf globala
c1	3.- erregresio-balorea	1	6820	1	537.319	23.443	26.655	3,116
c2	2.- estratu-balorea	1	7112	1	925.311	27.002	25.698	2,466
c3	3.- erregresio-balorea	2	5221	1	718.272	17.668		1,428
c4	3.- erregresio-balorea	2	7022	2	11.021.163	63.658	76.697	0,271
c5	3.- erregresio-balorea	1	4725	1	142.580	8.187	7.633	0,255
c6	3.- erregresio-balorea	1	6910	1	262.088	30.000	29.363	0,173
c7	3.- erregresio-balorea	1	9001	1	285.340	26.355		0,172
c8	3.- erregresio-balorea	1	6910	1	251.182	30.000	29.363	0,158
c9	2.- estratu-balorea	1	7220	1	60.703	20.234		0,142
c10	3.- erregresio-balorea	1	5320	1	44.390	30.717		0,117

○ Pertsona bakoitzeko faktore-kostuari erantsitako balioa araztea

Atal honetan balio erantsia araztuko dugu establezimenduko langileen kopuruaren arabera. Horretarako, estratuaren erdiko balioa soilik hartuko dugu aintzat, eta establezimendu guztiei pisu bera emango diegu.

7.4. taula. Pertsona bakoitzeko faktore-kostuari erantsitako balioa araztea.

Establezimendua	Arazoia	Enplegua	EJSN09	Enplegu-estratua	Pertsona bakoitzeko faktore-kostuari erantsitako balioaren ratioa	Erdiko balioa	sf globala
d1	2.- estratu-balorea	2	7022	2	5.510.582	39.980	0,066

d2	2.- estratu-balorea	2	5221	1	359.136	16.517	0,043
d3	2.- estratu-balorea	1	7112	1	925.311	26.220	0,041
d4	2.- estratu-balorea	5	4742	3	515.179	9.097	0,034
d5	2.- estratu-balorea	1	6820	1	537.319	22.294	0,030
d6	2.- estratu-balorea	2	7022	2	3.419.244	39.980	0,025
d7	2.- estratu-balorea	108	9312	7	407.720	32.938	0,021
d8	2.- estratu-balorea	7	4764	3	382.244	9.097	0,019
d9	2.- estratu-balorea	50	5210	6	842.255	54.571	0,016
d10	2.- estratu-balorea	10	6820	4	580.367	69.618	0,016

Pertsona bakoitzeko faktore-kostuari erantsitako balioaren ratioa eta estratuaren portaera oso bestelakoak direla ondorioztatu da; hortaz, berrikusi behar da, balioa zuzena den edo araztu behar den jakiteko.

- **Langileria-kostuak pertsona bakoitzeko araztea**

Egoera honetan langileria-kostuaren eta establezimendu bakoitzeko enpleguaren arteko ratioa zuzenean erabiltzea erabaki da.

7.5. taula. Langileen kostuak pertsona bakoitzeko araztea.

Establezimendua	Arrazoa	Enplegua	EJSN09	Enplegu-estratua	Langileen kostuen ratioa, pertsona bakoitzeko	Erdiko balioa	sf globala
e1	2.- estratu-balorea	108	9312	7	340.707	30.095	0,078
e2	2.- estratu-balorea	23	9312	5	340.707	27.364	0,032
e3	2.- estratu-balorea	65	9312	6	144.770	22.734	0,025
e4	2.- estratu-balorea	32	9312	5	267.197	27.364	0,019
e5	2.- estratu-balorea	246	8220	7	80.607	19.478	0,017
e6	2.- estratu-balorea	30	9312	5	218.031	27.364	0,013
e7	2.- estratu-balorea	106	9312	7	144.770	30.095	0,012
e8	2.- estratu-balorea	6	9102	3	79.146	31.681	0,012
e9	2.- estratu-balorea	7	4764	3	340.707	27.338	0,011
e10	2.- estratu-balorea	108	9312	7	340.707	30.095	0,078

Taula horretan ikus daiteke establezimendu batzuetan langileria-kostuaren ratioak, enpleguaren arabera, nabarmen gaitzen duela haren estratuaren erdiko balioa.

Arazketa selektiboaren ezarpenaren laburpena

Erakundearen garatu den arazketa selektiboko makroa, SAS-en programatuta, baliagarria izan da hainbat iturritatik etorritako ekonomia-datuak arazteko; betiere, datu horiek

balioztatu, integratu eta Eustat-en ekonomia-jardueren direktorioarekin erkatu ondoren.

Bestalde, honako hauek erabili dira aldagai nagusi gisa: negozioen kopuruaren zenbateko garbia; faktoreen kostuari erantsitako balioa eta langileria-kostua. Horiek hartu dira aldagai eskuragarri nagusitzat. Gainera, landun langileen aldagaiarekin lotura handiena dutenak dira.

Makroak ematen duen analisiaren bidez, muturreko balioak edo balio eraginkorrak dituzten establezimenduak bereizi ahal izan dira gorapen-estratuetan (jarduera eta enplegu-estratua); eta estratu horietan sortzen duten eragina ere kontuan hartu da. Horri esker, informazioa era eraginkorrean eta fidagarrian zenbatetsi ahal izan da, landu beharreko datuen kopurua oso handia izan arren.

8. Ondorioak

Azkenik, atal honetan jaso ditugu datu-baseen arazketa selektiboari buruzko lan honen ondorio nagusiak, bai eta laburpen bat ere. Lan hori garatu da estatistika- eta matematika-metodologietan prestatzeko eta ikertzeko beka jaso bitartean.

Arazketa selektiboaren gaineko laburpena eta ondorioak

Estatistika-erakundeek funtsezko iristen diote datuak arazteko metodo eraginkorrak izateari. Izan ere, datuen kalitatea hobetzeko prozesuan, datuak eskuz edo interaktiboki araztea da, hain zuzen, denbora eta diru gehien eskatzen duen alderdia.

Araztu beharreko erregistroen kopurua askoz ere txikiagoa izan daitekeela frogatu da. Izan ere, eskuzko arazketak ia ez du eraginik hainbat erregistrotan, behintzat, intereseko parametro nagusien zenbatesleei dagokienez.

Horrelako testuinguru batean, hautaketa-estrategia bat behar da erregistroak bi zatitan bereizteko: zati bat kritikoa izango da, eta balizko errore eraginkorrak bilduko ditu; eta besteak, araztuz gero, argitaratu beharreko emaitzetan aldaketa handirik sortuko ez luketen erregistroak bilduko ditu.

Arazketa selektiboaren arabera, argitaratu beharreko emaitzetan eragin handia izango duten erregistroak baino ez dira araztu eta zuzendu behar. Hortaz, estrategia horri jarraituta, kostuak eta entregatze-epea murriztuko dira.

Arazketa selektiboaren tresna nagusia “*score*” funtzioa da. Funtzio horrek puntuazio jakin bat esleitzen dio erregistro bakoitzari, aztertu den aldagai bakoitzeko. Puntuazio horrek adierazten du erregistro hori araztuz gero, zer eragin izango duen zenbatetsi beharreko parametroan. Lehenik, puntuazio handiko erregistroak araztu beharko dira.

Gainera, “*score*” funtzio-mota batzuk kalkula daitezke, erreferentzia-balioa zein den, edo aldagai batzuk aldi berean arazten ari garenean. Horregatik, erregistro oso bat hautatu eta araztu ahal izateko, beharrezkoa da zenbait “*score*” funtzioen informazioa konbinatzen duen balio bat. Balio horri esaten zaio puntuazio globala edo “*score*” globala. Puntuazio horrek erakutsi behar du erregistro osoa arazteko garrantzia.

Lan honetan erakutsi dugu “*score*” funtzioa tresna balioduna eta eraginkorra dela ezohiko erregistro eraginkorrak bereizteko. Lehenik eta behin, teorikoagoa zen esparru batean, datu-base simulatua erabili genuen. Horren bidez frogatu genuen “*score*” funtzio-mota batzuek nola funtzionatzen duten, araztu beharreko erregistroak hautatu behar dituztenean.

Ondoren, teknika horiek egiazko esparru batean ezarri genituen; zehazki, Euskal Estatistika Erakundearen “Zerbitzu-estatistika” deituriko jarduera berrian. Makroa, SAS-en programatuta, baliagarria izan da hainbat iturritatik etorritako ekonomia-datuak arazteko; betiere, datu horiek balioztatu, integratu eta Eustat-en ekonomia-jardueren direktorioarekin erkatu ondoren.

Nabarmentzekoa da bai testu honetan azaldutako teknikak, bai programatu diren SAS makroak, araztu beharreko edozein datu-basetan erraz ezarri eta aldatu ahal direla. SAS makroen parametroak alda daitezke, eta aldagai bakarra ezarrita edo zenbait aldagai batera ezarrita erabil daitezke. Lau “*score*” funtzio-motaren artean aukera daiteke, bai eta konbinatu ere, bakoitzari pisu zehatz bat esleituta. “*Score*” funtzio bakoitza hainbat estratutan kalkula daiteke. Gainera, erregistroek pisu bera eduki dezakete, edo pisu ezberdinekoak izan, zer interesatzen zaigun.

SAS makroak Eustaten webgunean eskuragarri egongo balira, erakundeek, institutuek edo ikertzaileek erabil ahal izango lituzkete, nahi izanez gero.

Azkenik, koaderno tekniko hau osatzeko, oinarri hartu da Holandako Estatistika Institutu Nazionaleko metodologia aplikatua. Metodologia hori zenbait koadernotan argitaratu da (Hoogland, van der Loo, Pannekoek and Scholtus, 2011; eta de Wall, 2008). Horrez gain, EDIMBUS europar proiektuan ere oinarrituta dago (Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys 2007).

Bibliografia

BELLISAI, D.; DI ZIO, M.; GUARNERA, U. AND LUZZI, O. (2009)

A selective editing approach based on contamination models: An application to an ISTAT business survey. Working Paper No. 27, UN/ECE Work Session on Statistical Data Editing, Neuchatel.

DE WAAL, T. (2008)

An overview of statistical data editing. Statistics Netherlands, The Hague/Heerlen.

DE WAAL, T., PANNEKOEK, J. AND SCHOLTUS, S. (2011)

Handbook of statistical data editing and imputation. Wiley.

DI ZIO, M.; GUARNERA, U. AND LUZZI, O. (2008)

Contamination Models for the Detection of Outliers and Influential Errors in Continuous Multivariate Data. Working Paper No. 22, UN/ECE Work Session on Statistical Data Editing, Vienna

EUREDIT PROJECT (2004a)

Towards Effective Statistical Editing and Imputation Strategies. Findings of the Euredit Project, Volume 1. Hemen eskuragarri: <http://www.cs.york.ac.uk/euredit/results/results.html>

EDIMBUS (2007)

Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys. Manual prepared by ISTAT, Statistics Netherlands and SFSO.

GHOSH-DASTIDAR, B. AND SCHAFER, J. L. (2006)

Outlier Detection and Editing Procedures for Continuous Multivariate Data. Journal of Official Statistics 22, 487-506 orr.

GRANQUIST, L (1995)

Improving the Traditional Editing Process. Business Survey Methods. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott, eds. John Wiley & Sons, New York, 385–401.

GRANQUIST, L AND KOVAR (1997)

Editing of Survey Data: How Much Is Enough? Survey Measurement and Process Quality. L.E. Lyberg, P. Biemer, M. Collins, E.D. De Leeuw, C. Dippo, , N. Schwartz, and D. Trewin, eds. John Wiley & Sons, New York, pp. 415–435.

HEDLIN, D (2003)

Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. Journal of Official Statistics 19, 177-199

HEDLIN, D (2008)

Local and Global Score Functions in Selective Editing. Working Paper No. 31, UN/ECE Work Session on Statistical Data Editing, Vienna.

HIDIROGLOU, M. A. AND BERTHELOT, J. M. (1997)

Statistica Editing and Imputation for Periodic Business Surveys. Survey Methodology 12, 73-78 orr.

HOOGLAND, J (2002)

Selective editing by means of Plausibility Indicators. UNECE Work Session on Statistical Data Editing, Helsinki, working paper no. 33.

HOOGLAND, J; VAN DER LOO, M; PANNEKOEK, J. AND SCHOLTUS, S. (2011)

Data editing. Detection and correction of errors. Statistics Netherlands, The Hague/Heerlen.

LATOUCHE, M. AND BERTHELOT, J. M. (1992)

Use of a Score Function to prioritise and Limit Recontacts in Editing Business Surveys Data editing. Journal of Official Statistics 8, 389-400 orr.

LAWARENCE, D. AND MCKENZIE, R. (2000)

The General Application of Significance Editing. Journal of Official Statistics 16, 243-253 orr.

VAN LANGEN, S (2002)

Selective Editing by Using Logistic Regression. Report, Statistics Netherlands, Voorburg.

Organismo Autónomo del



www.eustat.eus