Ikerketa beka
JOSE IGNACIO RUIZ OLABUENAGA
Beca de investigación

Komunikabideen audientzia ikerkuntzarako metodologia berriak esploratzen: datuen fusioa oinarri duen azterketa aplikatua

Explorando nuevas metodologías de investigación de audiencias: un análisis aplicado basado en la fusión de datos

euskal soziologia eta zientzia politikoaren elkartea
asociación vasca de sociología y ciencia política

EUSKO JAURLARITZA GOBIERNO VASCO
LEHENDAKARITZA PRESIDENCIA

Lan honen bibliografia-erregistroa Eusko Jaurlaritzako Liburutegi Nagusiaren katalogoan aurki daiteke:
Un registro bibliográfico de esta obra puede consultarse en el catálogo de la Biblioteca General del Gobierno Vasco:
https://www.katalogoak.euskadi.eus/katalogobateratua

Jatorrizko obra:	Komunikabideen audientzia ikerkuntzarako metodologia berriak esploratzen: datuen fusioa oinarri duen azterketa aplikatua
Obra original:	Explorando nuevas metodologías de investigación de audiencias: un análisis aplicado basado en la fusión de datos.
Argitaraldia:	1.a 2023ko urria
Edicion:	1ª, Octubre 2023
Ale-kpourua:	310 ale
Tirada:	310 ejemplares
©	Euskal Autonomia Erkidegoko Administrazioa Lehendakaritza Administración de la Comunidad Autónoma del País Vasco Departamento de Departamento de Presidencia
Internet:	www.euskadi.eus/argitalpenak
Argitaratzailea	Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia
Edita:	Servicio Central de Publicaciones del Gobierno Vasco Donostia/San Sebastián kalea 1. 01010 Vitoria-Gasteiz
Egileak: Autoras/es:	Naroa Burreso Pardo Josu Amezaga Albizu Libe Mimenza Castillo Edorta Arana Arrieta Irantzu Barrio Beraza
Azala marraztea, maketazioa eta inprimaketa: Diseño de portada, maquetación e impresión:	Irudi, S.L.
ISBN:	978-84-457-3719-4
Lege-gordailua: Depósito Legal:	LG G 709-2023

II. Ikerketa beka
JOSE IGNACIO RUIZ OLABUENAGA
II. Beca de investigación

Komunikabideen audientzia ikerkuntzarako metodologia berriak esploratzen: datuen fusioa oinarri duen azterketa aplikatua

Explorando nuevas metodologías de investigación de audiencias: un análisis aplicado basado en la fusión de datos

Naroa Burreso Pardo (Grupo de Investigación NOR, EHU/UPV)
Josu Amezaga Albizu (Grupo de Investigación NOR, EHU/UPV)
Libe Mimenza Castillo (Grupo de Investigación NOR, EHU/UPV)
Edorta Arana Arrieta (Grupo de Investigación NOR, EHU/UPV)
Irantzu Barrio Beraza (Grupo de Investigación MATHMODE, EHU/UPV)

euskal soziologia eta zientzia politikoaren elkartea

asociación vasca de sociología y ciencia política

EUSKO JAURLARITZA GOBIERNO VASCO
LEHENDAKARITZA
Koordinazio eta Gizarte Komunikaziorako Idazkaritza Nagusia
Prospekzio Soziologikoen Kabinetea
PRESIDENCIA
Secretaría General de Coordinación y Comunicación Social
Gabinete de Prospección Sociológica

Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia
Servicio Central de Publicaciones del Gobierno Vasco

Vitoria-Gasteiz, 2023

Edukien aurkibidea

Índice de contenidos

Sarrera

Hemen aurkezten dena Eusko Jaurlaritzako Prospekzio Soziologikoen Kabineteak eta Soziologia eta Zientzia Politikoko Euskal Elkarteak deitutako *Jose Ignacio Ruiz Olabuenaga II. Ikerketa Beka* deialdiari esker garatutako proiektuaren emaitza nagusia da.

Bekaren helburua, deialdiaren arabera, Euskal Herriko errealitate soziopolitikoa hobeto ezagutzen laguntzea zen, besteak beste ikerketa sozialeko metodologia berriak proposatuz eta gizarte-joera berrien azterketan Big Data erabiliz.

Helburuak irakurritakoan, berehala ondorioztatu genuen deialdia egokia zela Euskal Hedabideen Behategiaren inguruan garatu nahi genuen ikerketa ildo bat aurkezteko[1]. Izan ere, azken urteetako analisiek argiro erakutsi digute joera berri eta garrantzitsuak ari direla gertatzen gizarteko komunikazio praktiketan, eta horrek ondorio sakonak izan ditzakeela gizartean; gizartean oro har, eta baita euskal gizartean ere. Azken horri dagokionez zehazki, eta Behategiaren lan eremutik begiratuta, euskarazko komunikazio praktiketan. Argi

1 Behategia Hekimenen eta lau unibertsitateren (EHU, DU, MU eta UEU) arteko hitzarmen baten emaitza da. Egoitza EHUn du, eta Hekimen eta EHUren arteko lankidetza hitzarmenen bidez finantzatzen da,

Introducción

Lo que a continuación se presenta es el principal fruto del trabajo de investigación realizado gracias a la *2ª Beca de Investigación José Ignacio Ruiz Olabuenaga*, convocada por el Gabinete de Prospección Sociológica del Gobierno Vasco y la Asociación Vasca de Sociología y Ciencia Política.

El objetivo de la beca, según la convocatoria, era contribuir a conocer mejor la realidad sociopolítica del País Vasco, proponiendo nuevas metodologías de investigación social y utilizando el Big Data en el análisis de las nuevas tendencias sociales.

A la vista de los objetivos, vimos enseguida la pertinencia de la convocatoria en relación con una línea de investigación que pretendíamos desarrollar en torno al Observatorio de los Medios de Comunicación en euskera, Behategia[1]. Los análisis realizados en los últimos años nos han demostrado que se están generando nuevas y significativas tendencias en las prácticas de comunicación social, las cuales pueden tener repercusiones profundas en la sociedad en general y en la sociedad vasca en particular. En este último caso, y desde el ámbito de trabajo del

1 Behategia es un marco de colabroación entre Hekimen y cuatro universidades (UPV/EHU, DU, MU y UEU). Tiene su sede en la UPV/EHU y se financia mediante convenios de colaboración entre Hekimen y la UPV/EHU.

genuen horrenbestez ikerketa metodologia berriak behar genituela, eta datu ugaritasunaren garaian ezinbestekoa genuela diziplinartekotasuna garatzea, analisi soziologikoari analisi matematikoa gehituta.

Hori guztia buruan genuela argitaratu zen ikerketa bekaren deialdia, eta erabaki erraza izan zen proiektua aurkeztea. Gure proiektuak (D^i - Datu Integralak) aukeratua izateko zoria izan zuen, eta horrek ahalbidetu zigun aurrera egitea.

Lehen unetik argi genuen gure lanaren emaitza ezin zela euskarazko komunikazio praktiken azterketara mugatu, eta euskal soziologiari ere ekarpena egin nahi genion. Hartara, hemen aurkezten den lana ikerketa soziologikoan aurrerapausoa emateko baliagarria delakoan argitaratzen dugu. Izan ere, bizitzen ari garen gizarte aldaketen bizkortasunak batetik, eskura dauden datuen ugaritasunak bestetik, eta metodologietan garatzen ari diren bideek azkenik, horretarako aukerak ematen dituzte.

Lantaldea arlo ezberdinetako pertsonek osatu dugu. Edorta Arana Arrieta, Libe Mimenza Castillo eta ni neu EHUko NOR Ikerketa Taldeko kideak gara, soziologia eta komunikazio zientzietako formazioa dugu, eta komunikazioa, kultura eta nortasunak aztertzen ditugu ikus-

Observatorio, en las prácticas de comunicación en euskera. Teníamos claro entonces que precisamos de nuevas metodologías de investigación y que, en la época de abundancia de datos, se hace imprescindible el recurso a la interdisciplinariedad, incorporando al análisis sociológico el análisis matemático.

Eso teníamos en mente cuando se publicó la convocatoria de la beca de investigación, por lo que la decisión de presentar el proyecto no fue difícil. Nuestro proyecto (D^i - Datos Integrales) tuvo la suerte de ser seleccionado y ello nos permitió avanzar por esa vía.

Teníamos claro desde el primer momento que el resultado de nuestro trabajo no podía limitarse al análisis de las prácticas comunicativas en euskera, y quisimos hacer una mayor aportación a la sociología vasca. De esta forma, presentamos aquí este trabajo desde la convicción de que puede servir para avanzar en la investigación sociológica. De hecho, la rapidez de los cambios sociales que se están produciendo, por un lado, la abundancia de los datos disponibles, por otro, y las nuevas metodologías en desarrollo, ofrecen posibilidades para ello.

El equipo de trabajo lo hemos formado personas de diferentes áreas. Edorta Arana Arrieta, Libe Mimenza Castillo y yo mismo formamos parte del Grupo de Investigación NOR de la UPV/EHU, te-

pegi soziologikotik. Naroa Burreso Pardo matematikatik dator, eta proiektu honi esker (bai eta Lanbideren laguntza bati esker) NORkide bihurtu da. Irantzu Barrio Beraza MATHMODE Ikerketa taldekoa da, EHUkoa baita ere. Horrela aldi berean bi diziplinatan mugitu gara urte betean zehar, talde aberasgarria osatuta.

Nori berea da gizalegea, eta Eusko Jaurlaritzako Prospekzio Soziologikoen Kabineteak eta Soziologia eta Zientzia Politikoko Euskal Elkartearentzat dira lehen esker onak. Haiek proiektuan jarritako konfiantzarik gabe nekez iritsiko ginen lerro hauetara. Laguntza ekonomikoaz gain, lanerako ildo batzuk eta kontrastea eskaini dizkigute. CIES enpresari ere eskerrak ematea dagokigu; urte askotako lankidetzak sortu duen konfiantza giroan erabat prest agertu baita beren datuak proiektu honetarako emateko, musutruk. Berdin egin du AIMC elkarteak ere; oraingo honetan haien inkestetako mikrodatuak erabili ez arren, eskura dugu, haien eskuzabaltasunari zor, inoiz erabiltzeko aukera. Haien laguntza teknikoa, bestalde, ezinbestekoa izan da proiektuan aurrera egiteko, hasierako gure intuizioetatik emaitzen baliozkotzeraino. Hor izan dira ere EUSTATeko langileak hasiera-hasieratik amaiera-amaieraraino laguntza ematen, behar izan dugun bakoitzean. Esan dezakegu beraz, proiektuaren beste emaitza

nemos formación en Sociología y Ciencias de la Comunicación, y trabajamos en el ámbito de la comunicación cultura e identidades desde una perspectiva sociológica. Naroa Burreso Pardo proviene de las matemáticas y gracias a este proyecto (así como a una ayuda de Lanbide) se ha convertido en miembro de NOR. Irantzu Barrio Beraza, matemática, es miembro del grupo de investigación MATHMODE, también de la UPV. De esta forma nos hemos movido a lo largo de un año entre dos disciplinas, formando un equipo enriquecedor.

Nori berea da gizalegea dice el refrán en euskera: a cada cual lo que en justicia le corresponde, por lo que nuestro primer agradecimiento es para el Gabinete de Prospección Sociológica del Gobierno Vasco y para la Sociedad Vasca de Sociología y Ciencia Política. Sin su confianza en el proyecto difícilmente habríamos llegado a estas líneas. Además de la ayuda económica, nos han ofrecido unas líneas de trabajo y un contraste del mismo. Debemos también mostrar nuestro agradecimiento a CIES. Dentro el clima de confianza generado tras muchos años de colaboración, nos han mostrado su total disposición a aportar sus datos para este proyecto. La asociación AIMC también lo ha hecho así; a pesar de no utilizar los microdatos de sus encuestas en esta ocasión, aun contamos con la posibilidad de utilizarlos gracias a su generosidad. Su apoyo técnico

bat —batere txikia ez, gainera— lankidetza sare indartsu bezain atsegina ehuntzea izan dela.

Amaitzeko, ezin dut ukatu, maila pertsonalean, hunkigarria ere izan dela proiektu hau aurrera eramateko parada. Irakasle fin izan nuen José Ignacio Ruiz Olabuenaga, Soziologia ikasten ari nintzelarik. Eta seguruenik bere ikasle izandako gehientsuenak bezalaxe, estimu handia nion. Gizon handia zen, hala bere lan arloan nola pertsona gisa ere, eta hark ikasleoi transmititu zizkigun jakintza, zorroztasuna zein maitasuna, dena batera eta irripartsu, nola edo hala itzultzeko aukera izan dut lan honen bitartez. Pena dut, handia, emaitza hari aurrez-aurre aurkeztu ezina. Horren ordez, inguruko soziologoei eskaini nahi diet. Euskal soziologiari hainbeste eman zion maisuari ere ilusioa egingo zion, urte askoren ondoren bada ere, hau haren lanaren itzala baita.

Josu Amezaga Albizu

ha sido fundamental para avanzar en el proyecto, desde nuestras intuiciones iniciales hasta la validación de los resultados. Por otro lado el personal de EUSTAT también nos ha ayudado cada vez que lo hemos necesitado, desde el principio hasta el final. Podemos decir pues que otro de los resultados del proyecto, nada baladí, ha sido la posibilidad de tejer una red de cooperación tan estimulante como amable.

Para finalizar, y a nivel personal, no puedo negar que las emociones también han participado en este proyecto. Tuve el placer de tener a José Ignacio Ruiz Olabuenaga como profesor de Sociología y, seguramente igual que sucedió con la mayoría de sus antiguos estudiantes, se ganó mi gran aprecio. Era un gran hombre, tanto en su faceta profesional como humana, a quien a través de este trabajo he tenido la oportunidad de agradecer de una u otra manera el saber, el rigor y el cariño que nos transmitió a quienes con él aprendimos. Me hubiese gustado poder presentárselo en persona. En su defecto querría dedicárselo a las sociólogas y sociólogos de nuestro entorno. Creo que es lo que le hubiese gustado a alguien que tanto le diera a la sociología vasca; de alguna manera es la extensión de su aportación a la investigación, al cabo de muchos años.

Josu Amezaga Albizu

Lehen Atala: arazoaren planteamendua

1. Egoeraren deskribapena

Ez da oso berria esatea gizarte-eraldaketa sakonak bizitzen ari garela, ez eta aldaketa horietako asko biztanleriaren komunikazio-praktiketako aldaketekin lotuta daudela ere, horiek kultura-dinamika disruptoreak sortzen baitituzte. Analisi-mailan behera egiten badugu, ordea, ez dago hain zehaztuta nola gertatzen ari diren aldaketa horiek komunikazio-praktiketan; are gutxiago, zer ondorio zehatz izan ditzaketen ezagutzen ditugun gizarteetan.

Bide horretan ezagutza argiztatzeko asmoz, komunikabide berriek lehendik zeudenak ordezkatuko ote zituzten ala eguneroko dieta mediatikoan erantsiko ziren hausnartzen hasi zen 80ko hamarkadan. Galdegai horrek bide eman zion Ordezkapenaren Teoriari (Kaye & Johnson, 2003). Baina gaia ez zen guztiz originala, komunikabide berri bat agertzen zen bakoitzean (irratia liburuaren aurrean, telebista irratiaren eta zinemaren aurrean, edo argazkia pinturaren aurrean) antzeko galderak planteatzen baitziren.

Beste behin, komunikabideen birdefinizioa bizi dugu egun: orain

Capítulo primero: planteamiento de la cuestión

1. Descripción de la situación

No es muy novedoso decir que estamos viviendo transformaciones sociales profundas, ni que gran parte de esas transformaciones tienen relación con los cambios en las prácticas comunicativas de la población, que generan a su vez dinámicas culturales disruptivas. Si descendemos en el nivel de análisis, sin embargo, no está tan detallado cómo se están produciendo esos cambios en las prácticas comunicativas; y menos aún qué efectos concretos pueden tener en las sociedades que conocemos.

En este sentido, ya en los años 80 se suscitaba la cuestión sobre si los entonces nuevos medios de comunicación sustituirían a los existentes, o se añadirían a estos en la dieta mediática diaria. Este debate, que dio pie a la Teoría de la Sustitución (Kaye & Johnson, 2003), no era del todo novedoso, puesto que planteaba preguntas que ya habían surgido cada vez que aparecía un nuevo medio (la radio frente al libro, la televisión frente a la radio y el cine, o incluso la fotografía frente a la pintura).

Hoy en día disponemos de evidencias que muestran claramen-

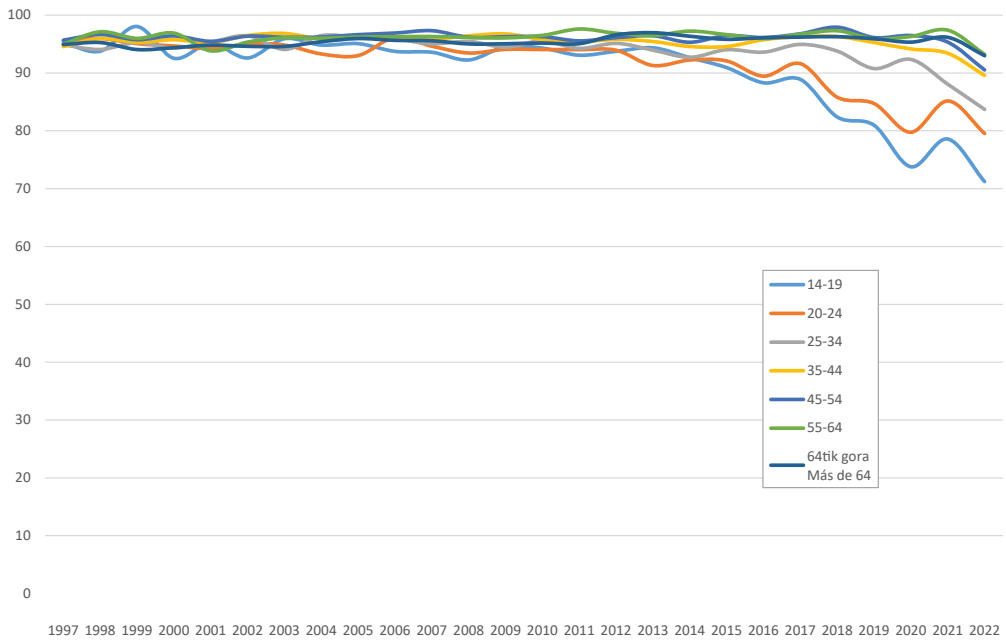
dela gutxira arte tradizionaltzat jo izan diren komunikabideetatik sareko komunikazio-jardueretara trantsizio bat gertatzen ari dela erakusten duten ebidentzia nahikoa daukagu.

Eskura dauden datuek agerian jartzen dutenez, komunikabide tradizionalak alboratzen ari dira, batez ere gazteen artean. Prentsa —paperean, PDF formatuan eta edizio digitalean—, irrati lineala eta telebista lineala kontsideratzen ditugu baliabide tradizionaltzat.

te que efectivamente se está produciendo una sustitución o transición desde los medios hasta hace poco tradicionales hacia las prácticas comunicativas en red.

Los datos disponibles muestran un abandono real, encabezado por los sectores más jóvenes de la población, de los medios tradicionales. Entendemos aquí como medios tradicionales la prensa (en papel, en formato PDF y en edición digital), la radio lineal y la televisión lineal.

1 grafikoa: prentsa (paperean, PDF formatuan eta edizio digitalean), irrati lineala edo telebista lineala egunero kontsumitzen duten biztanleen ehunekoa, adinaren arabera (Araba, Bizkaia, Gipuzkoa eta Nafarroa, CIES) / Gráfico 1: porcentaje de población que consume diariamente prensa (en papel, PDF o edición digital), radio lineal o televisión lineal, por grupo de edad (Araba, Bizkaia, Gipuzkoa y Nafarroa, CIES)



1. grafikoan, euskarri horietako bat gutxienez egunero kontsumitzen duten biztanleen ehunekoa ageri da, adin-taldeka. Geuk landutako datuak dira, CIES-ek Araba, Bizkaia, Gipuzkoa eta Nafarroarako egindako Komunikabideen Audientzia Azterlanetik abiatuta (CIES, 2022); hala ere, ez dugu arrazoirik ikusten gure inguruko beste gizarteekin alderatuta gehiegi aldatuko direnik pentsatzeko.

El Gráfico 1 muestra el porcentaje de población, por grupos de edad, que consume diariamente al menos uno de esos soportes. Son datos de elaboración propia a partir del Estudio de Audiencia de Medios de Comunicación de CIES para Araba, Bizkaia, Gipuzkoa y Nafarroa (CIES, 2022), si bien no vemos razones para pensar que varíen en exceso con respecto a otras sociedades de nuestro entorno.

Grafikoak argi eta garbi agerrarazten du azken hamar urteetan izan den eraldaketa, horrelako datuak ditugunetik (1997) gaur arte (2022). Ordura arte, gutxienez euskarri tradizional bateko ohiko kontsumitzaileen ehunekoa % 95 ingurukoa edo handiagoa zen adin-talde guztietan. Egoera horretatik abiatuta, oraindik gelditu ez den aldaketa bat hasi zen duela hamarkada bat: pertsona gazteenak (normalean inkesten bidez behatzen diren gazteenak, 14tik 19ra urte bitarteko taldeak ordezkatzen ditu) ildo nagusitik askatzen hasi ziren, beheranzko joerarekin; eta gazteen beste taldeak (20-24 eta 25-34, baita 35-44 eta gehiago ere) atzetik ekin zioten portaera aldaketari. 2022ra iritsi arte, non 14-19 taldean hedabide tradizionalen baten eguneroko kontsumoa % 71 baita. Bestela esanda, ia hiru gaztetik batek ez du kontsumitzen, egunero, ez prentsa, ez irrati lineala, ez telebista lineala.

Vemos nítidamente la transformación que opera desde hace diez años. Hasta entonces, desde que disponemos de datos de este tipo (1997-2022), el porcentaje de personas consumidoras habituales de al menos un soporte tradicional rondaba o superaba el 95 % en todos los grupos de edad. Sin embargo, hace una década comienza un cambio que no parece haberse detenido aun: las personas más jóvenes (de entre aquellas a las que habitualmente se accede a través de encuestas, el grupo de 14 a 19 años) comienzan a despegarse de la línea general, con una tendencia decreciente que es seguida a cierta distancia por los otros grupos jóvenes (20-24 y 25-34, incluso 35-44 y más), hasta llegar a 2022 con un porcentaje de consumo cotidiano de algún medio tradicional del 71 % en el grupo 14-19. Dicho de otro modo, casi una de cada tres personas jóvenes ya no consume, de manera diaria, ni prensa, ni radio lineal, ni televisión lineal.

Grafikoak erakusten duen irudia zirpilduz doan soka batena da, eta metafora saihestezina bihurtzen da hainbat hamarkadatan komunikabideek —beste eragile batzuekin batera— txirikordatutako gizartearen aldaketa deskribatzeko, non komunikabideek indarra galtzen baitute kohesio elementu gisa.

Deigarria da, zalantzarik gabe, joera-aldaketa hori telefono adimendunen erabileraren hedapenarekin bat etortzea. CIESen datuekin jarraituz, 2011n (datua biltzen den lehen urtean) 14 urtetik gorako biztanleen % 37k telefono adimenduna zuen, eta ehuneko hori % 55ekoa zen 14-19 urte bitartekoen artean; bi urte geroago, 2013an, datuak % 62 eta % 89 ziren, hurrenez hurren.

Komunikabide tradizionalak sareko praktikekin ordezkatzeak sozializazio-prozesuetan ondorio sakonak ote dituen galdetu behar diogu gure buruari, landu nahi dugun esparruari dagokionez behintzat. Sozializazio hori zentzu soziologiko klasikoan ulertzen dugu hemen, zeinaren arabera sozializazio primarioa (identitate indartsuenak sortzen dituena) lau eragile nagusiren bidez egiten baita: familia, eskola, parekoen taldea eta komunikabideak. Eragile horiek jasan edo sortzen dituzten aldaketek influentziaren bat izango dute sozializazio-prozesuetan; eta, zalantzarik gabe, gaur egun aldaketa sakonak gertatzen ari dira.

La imagen que muestra el gráfico es el de una cuerda que se va deshilachando, convirtiendo en inevitable la metáfora para describir una sociedad trenzada durante décadas por -entre otros agentes- los medios de comunicación, en la que éstos estarían perdiendo fuerza como elemento cohesionador.

Sin duda llama la atención la coincidencia del cambio de tendencia con la extensión del uso de los teléfonos inteligentes. Siguiendo con los datos de CIES, en 2011 (primer año en que se recaba el dato) el 37 % de la población mayor de 14 años disponía de un teléfono inteligente, siendo este porcentaje del 55 % para el grupo 14-19 años; dos años más tarde, en 2013, los datos alcanzan el 62 % y el 89 % respectivamente.

Cabe preguntarse si esta sustitución de los medios tradicionales por las prácticas en red tiene realmente implicaciones profundas en los procesos de socialización, al menos en relación al ámbito que pretendemos abordar. Y entendemos aquí socialización en un sentido sociológico clásico, según el cual la socialización primaria (aquella que crea identidades más fuertes) se realiza a través de cuatro grandes agentes: familia, escuela, grupo de pares y medios de comunicación. Los cambios que operan pues en estos agentes han de tener impacto en los procesos de socialización, y no hay duda de que hoy en día aquellos están sufriendo profundas transformaciones.

Familiaren kasuan, forma berrien aniztasun eta onarpen sozial handiagoaz (pertsona baten ardurapeko familiak, familia berregituratuak, eta abar) eta emakumeek gizartean (eta, beraz, baita familian ere) duten rolaren eraldaketaz gain, fenomeno demografikoek datu bat uzten digute kontuan hartzeko: ama atzerritardun jaiotzak % 22 izan ziren 2021ean Araban, Bizkaian, Gipuzkoan eta Nafarroan; horrek berekin dakartzan inplikazioekin, familiaren bidez sozializatzeari begira, hots, gizarte jakin batean integratzea.

Bestalde, Interneten agerpena eta sarearen bitartekotza duten harreman primarioen, sekundarioen, tertziarioen eta kuaternarioen ugaritasuna (Calhoun, 1991) eskolaren bidezko sozializazioan zein parekoen taldearen bitartez gertatzen den sozializazioan islatzen dira: espazio horiek gero eta gehiago kolonizatzen ditu teknologiak, horrek erreprodukzio sozialean izan ditzakeen ondorio guztiekin. Algoritmoak leku pribilegiatua hartzen ari dira eskola-harremanen eta parekoen arteko harremanen esparruan. Familiaren espazio sozializatzailea ere inbaditzen ari dira, gazteak familiaren elkarreraginetik ateratzen baitituzte, makinekiko eta makinen bidezko harremana areagotzeko.

Testuinguru horretan, ohiko komunikabide tradizionalen kontsumotik sareko praktiketarako

En el caso de la familia, aparte de una mayor variedad y aceptación social de nuevas formas (familias a cargo de una persona, familias reestructuradas, etcétera) y de la transformación del rol de las mujeres en la sociedad (y por tanto también en la familia), los fenómenos demográficos nos dejan un dato a considerar: el 22 % de los nacimientos en 2021 en Araba, Bizkaia, Gipuzkoa y Nafarroa lo fueron de madre extranjera, con las implicaciones que ello conlleva de cara a la socialización (entendida como integración en una sociedad concreta) a través de la familia.

Por otro lado, la aparición de internet y la profusión de relaciones primarias, secundarias, terciarias y cuaternarias (Calhoun, 1991) mediadas por la red se refleja tanto en la socialización a través de la escuela como en aquella que tiene lugar a través del grupo de pares: esos espacios están, cada vez más, colonizados por la tecnología, con todas las implicaciones que ello puede tener para la reproducción social. Los algoritmos están ocupando un lugar privilegiado en el ámbito de las relaciones escolares y de las relaciones entre pares. Están también invadiendo el espacio socializador de la familia, al retraer a las personas jóvenes de la interacción familiar para aumentar la relación con y a través de las máquinas.

En este contexto, nos parece sumamente pertinente analizar

jauzia nola gertatzen ari den aztertzea guztiz egokia iruditzen zaigu. Baliteke euskarri batzuek beste batzuk ordezkatzea, baina gizarte-erreprodukzioan ondorio berak izatea; ebidentziak, ordea, kontrakoa iradokitzen du. Sareko edukien globaltasunak, beste hizkuntzetako produktuetara eta, beraz, beste komunikazio-espazio batzuetara (batez ere ingelesera) gero eta errazago iristeak (bai herritarren gaitasun handiagoagatik, bai baliabide tekniko handiagoengatik), edo produktu-eskaintza eskerga baten berehalako eskuragarritasunak zein eskura dagoen bibliografiak, sarean kontsumitzen edota ekoizten diren edukiak eta euskarri tradizionaletan kontsumitzen direnak oso desberdinak direla pentsarazten dute. Beraz, ez dirudi lan honetan inplizitua dagoen hipotesia zentzugabea denik, alegia, euskarri batzuetatik besteetarako trantsizioak eragin garrantzitsuak dituela gaur egun ezagutzen ditugun gizarteen sozializaziorako eta birsorkuntzarako.

Asko idatzi da soziologiak komunikabideak lehen mailako sozializazio-agentetzat hartu zituenetik. Horretan lagundu dute, besteak beste, epe luzeko efektuei buruzko teoriek, hala nola G. Gerbnerren Kultiboaren Teoriak (Gerbner, 1986). Teoria horretan komunikabideei emandako garrantzia hain-

cómo se está produciendo el tránsito desde el consumo de medios de comunicación tradicionales a las prácticas en red. Bien podría ser que unos soportes sustituyan a otros, pero continúen con los mismos efectos sobre la reproducción social. Sin embargo la evidencia parece sugerir lo contrario. El hecho de la globalidad de los contenidos en red, del acceso cada vez más fácil a productos en otras lenguas y por tanto a otros espacios de comunicación, especialmente el inglés (tanto por mayor capacitación de la población como por mayores recursos técnicos), o de la disponibilidad instantánea de una inmensa oferta de productos, así como la bibliografía existente, hacen pensar que los contenidos que se consumen o producen en red son muy diferentes a aquellos que se consumen en soportes tradicionales. Por lo que no parece descabellada la hipótesis, implícita en este trabajo, de que la transición de unos soportes a otros tiene impactos importantes para la socialización y para la reproducción de las sociedades que conocemos actualmente.

Mucho se ha escrito desde la consideración que hace la Sociología de los medios como agentes de socialización primaria. A esta consideración han contribuido, entre otros, teorías de los efectos a largo plazo, como la Teoría del Cultivo, de G. Gerbner (Gerbner, 1986). La importancia dada a los medios de comunicación en estas

bestekoa izango litzateke ezen, Roda Fernandezen arabera komunikabideen sozializazio lanak «lagundu egiten baitu Comte, Tönnies eta Durkheimenek zuten gizarte-desintegrazioaren mamua uxatzen, berau gizarte modernoaren mehatxu nagusitzat jo baitzuten» (Roda Fernández, 1989).

Gerbnerrek balioetan sozializatzeaz aztertzen duenaz gain, eta bereziki gizarte nazionalen birsorkuntzan medio tradizionalek izan duten tokia kontuan hartuta, esan dezagun nahiko finkatuta dagoela komunikabideek nazioen eraikuntzan duten garrantzia. K. Deutschen lan aitzindarietatik (Deutsch, 1953) B. Andersonen lanetara (Anderson, 1983) , asko dira tesi honen alde egin duten autore ezagunak. Denboran hurbilago, M. Billigen nazionalismo hutsalaz ere hitz egin genezake (Billig, 1995).

Lerro paraleloan, J. Habermasek esfera publikoaren euskarri nagusietako bat finkatzen du komunikabideetan (Habermas, 1962). Geroagoko ekarpen interesgarri batean (Curran, 1991), komunikabideen bidezko aisiaren esparrua gehitu zitzaion eremu publikoaren kontzeptu horri —hasieran informazioaren ikuspuntutik pentsatua izan baitzen—.

Era berean, Espainiako Estatuan ekarpen esanguratsuak egin dira hedabideek gizarte

teorías sería tal que, según Roda Fernández su acción «contribuye a exorcizar el fantasma de la desintegración social que Comte, Tönnies y Durkheim habían considerado la principal amenaza de la sociedad moderna» (Roda Fernández, 1989).

Más allá de la socialización en valores de la que habla Gerbner y centrándonos en concreto en el lugar que los medios tradicionales han ocupado en la reproducción de las sociedades nacionales, digamos que está bastante establecida la importancia de los medios en la construcción de las naciones. Desde los trabajos pioneros de K. Deutsch (Deutsch, 1953) hasta los de B. Anderson (Anderson, 1983), son numerosos los autores reconocidos que defienden esta tesis. Más cercano en el tiempo podríamos hablar también de nacionalismo banal de M. Billig (Billig, 1995).

En una línea paralela, J. Habermas ve los medios como uno de los principales soportes de la esfera pública (Habermas, 1962). En una posterior y certera anotación, J. Curran (Curran, 1991) añadía el ámbito del ocio a través de los medios a ese concepto de esfera pública, inicialmente más pensado desde el punto de vista de la información.

También desde el Estado español se han hecho aportaciones significativas para entender la relación de los medios con la

nazionalen eraikuntzarekin duten harremana ulertzeko, hala nola komunikazio espazio nazionalaren —katalanaren— nozioa (Gifreu, 1989).

Komunikabideen kontsumoaren eta praktika linguistikoen arteko erlazioa ez dago hain finkatua ordea, soziolinguistikaren ikuspuntutik behintzat. Arlo horretako erreferente nagusietako bat, J. Fishmanek, ez zuen hain nabarmentzat jotzen duela hiru hamarkada (Fishman, 1991). Hala ere, azken urteotan hainbat lan argitaratu dira, bi praktiken artean lotura dagoela pentsarazten dutenak —harreman hori hainbat mekanismoren bidez gauzatuko litzateke (Cormack, 1998; Jones, 2013; Moring et al., 2011)—. Badira bestalde zenbait azterlan sareko praktiken gorakada hizkuntza-praktikekin lotzen dituztenak, bereziki hiztun-kopuru mugatua duten hizkuntzetan. Islandian, adibidez, harreman horren susmoak hizkuntza-politika protekzionistak eragiten ditu (Hilmarsson-dunn, 2006). Edonola ere, ez dirudi erabateko adostasunik dagoenik inpaktuaren magnitudeari buruz (Sigurjónsdóttir & Nowenstein, 2021).

Aldaketa horiek ziztu bizian gertatzen ari dira. Agian horregatik, ez dugu bibliografia askorik aurkitzen trantsizioak gizarte-erreprodukzioan dituen ondorioen gain. Bai, ordea, trantsizio horren alderdi zehatzean zentratua (kasu-azterketak, ere-

construcción de las sociedades nacionales, como fue la noción del “espacio nacional –catalán- de comunicación” (Gifreu, 1989).

No está tan afianzada la relación entre consumo de medios y prácticas lingüísticas, al menos desde la sociolingüística. Uno de los principales referentes de esta rama, J. Fishman, no la consideraba hace tres décadas tan evidente (Fishman, 1991). Sin embargo en los últimos años se han publicado numerosos trabajos que hacen pensar que efectivamente existe una relación entre ambas prácticas, materializada a través de diferentes mecanismos (Cormack, 1998; Jones, 2013; Moring et al., 2011). De manera específica, existen ya algunos estudios que relacionan el auge de las prácticas en red con las prácticas lingüísticas, especialmente en lenguas con limitado número de hablantes. Es el caso por ejemplo del islandés, donde la sospecha de esa relación está llevando a políticas lingüísticas proteccionistas (Hilmarsson-dunn, 2006). Aun así, no parece haber un consenso total sobre la magnitud del impacto (Sigurjónsdóttir & Nowenstein, 2021).

Estos cambios están sucediendo de manera vertiginosa. Tal vez por esa razón, no encontramos una bibliografía abundante sobre las consecuencias de la transición en la reproducción social. Sí la hay sobre aspectos concretos de esa transición (estudios de ca-

mu jakinetako azterketak, hala nola hezkuntza, genero-auzia, eta, azken aldian, desinformazioari buruzkoak, eta abar), baina ez hainbeste lurraldeari lotutako komunitate-zentzuko gizarte-erreprodukzioari, hizkuntza-erabilerari eta kultura-birsorkuntzaren beste alderdi batzuei buruz.

2. Egoera aztertzeko eskura ditugun tresnak

Populazioaren komunikazio ohiturak aztertzeko tresna ugari daude gaur egun. Garatuenak, eta datuetan emankorrenak, publizitatearen merkatuari begira egindako audientzia ikerkuntzak eskaintzen dituenak dira. Komunikabideak publizitateagatiko diru sarrerak behar dituzten neurrian, oso aspaldian arduratu ziren haien audientziak ezagutzeaz, horretarako tresna berezituak garatuz. Historia apur bat eginez, Paul F. Lazarsfeld (1901-1976) gogora dezagun. Soziologia enpiriko kuantitatiboaren aitzindari nagusitzat hartua da bera, matematikan doktorea zen. 1931n Austrian irrati entzuleen ezaugarriak eta haien lehentasunak ezagutzeko egin zuen ikerketa, *RAVAG Ikerketa* izenez ezaguna (*Die Hörerbefragung der Ravag*, 1932koa), lehen inkesta zientifikotzat hartua da (Pavlik, 2017). Hortik aurrera, irratiaren eta komunikabideen audientzien ikerketari eman zion bere lanaren zati handia matematikasos, análisis en ámbitos concretos como la educación, la cuestión de género y, últimamente, sobre la desinformación, etcétera), pero no tanto sobre cuestiones de reproducción social en sentido de comunidad ligada al territorio, sobre usos lingüísticos, y sobre otros aspectos de la reproducción cultural.

2. Herramientas disponibles para el análisis de la situación

Existen en la actualidad abundantes herramientas para el estudio de los hábitos de comunicación de la población. Las más desarrolladas, y las más fructíferas en datos, son las ofrecidas por la investigación de audiencias orientada al mercado publicitario. Los medios de comunicación, en la medida en que necesitan ingresos por publicidad, se preocuparon desde hace bastante tiempo de conocer sus audiencias, desarrollando para ello herramientas específicas. Haciendo un poco de historia recordemos a Paul F. Lazarsfeld (1901-1976). Considerado el principal precursor de la sociología empírica cuantitativa, era doctor en matemáticas. Su estudio realizado en 1931 en Austria para conocer las características y preferencias de las personas radioyentes, conocido como *Estudio RAVAG* (*Die Hörerbefragung der Ravag*, 1932), es considerado como la primera encuesta científica (Pavlik, 2017). A partir de ahí, la socióloga matemática austriaca y americana (incluida la *Radio*

ri-soziologo austriar-amerikarrak (besteak beste *Radio Research Project* barnean). Garai hartan, ordura arte medio idatziek monopolioan hartutako espazioa okupatzen hasiak ziren ikus-entzunezkoak (irratia eta zinema lehenik, telebista geroago), eta euskarri berri haien erabilera komertzialez harago, 1930-1940 hamarkadetan izandako gertaera politikoekin zein masa mugimendu izugarriekin izan zezaketen harremana aztertu beharra zegoen. Testuinguru hartan, Lazarsfeld eta beren lankideek medioen inguruan egindako lanek elikatu zituzten, *Mass Communication Research* ez eze, soziologia osoaren metodoak (Jerabek, 2001).Geroztik oparoa izan da audientzien ikerketa, publizitatearen merkatuak bultzatuta (Quintas-Froufe et al., 2021).

2.1. Audientzia ikerketak Europan

Esan gabe doa audientzien neurketa garrantzitsua dela gero eta negozio bolumen handiagoa duen publizitatearen merkaturako. Merkatu honek 127.000 milioi euro mugitu zituen 2017an Europako Batasunean (European Commission, 2023). Horrek erabakien erdigunean kokatzen ditu audientzien ikerketak: 14.500 milioi euro

Research Project) dedicó gran parte de su trabajo a la investigación de las audiencias de la radio y los medios de comunicación en general. En aquella época, los medios audiovisuales (la radio y el cine primero, la televisión más tarde) empezaron a ocupar un espacio hasta entonces monopolizado por los medios escritos, y más allá de los usos comerciales de aquellos nuevos soportes, surgió el interés por analizar su relación con los acontecimientos políticos ocurridos en los años 1930-1940, así como con los fenómenos de masas de la época. En este contexto, los trabajos realizados por Lazarsfeld y sus compañeros en torno a los medios alimentaron los métodos de toda la sociología, no solamente la de la conocida como *Mass Communication Research* (Jerabek, 2001). La investigación de las audiencias así impulsada por el mercado de la publicidad ha resultado, desde entonces, ciertamente fructífera en términos de investigación social (Quintas-Froufe et al., 2021).

2.1. La investigación de audiencias en Europa

Huelga decir que la medición de audiencias es crucial en un mercado de la publicidad con un volumen de negocio creciente. Este mercado movilizó 127.000 millones de euros en 2017 en la Unión Europea (European Commission, 2023). Esto sitúa los estudios de audiencias en el centro de muchas decisiones sobre publicidad:

gastatu ziren, urte berean, merkatuen ikerkuntzan kontinente zaharrean.

Ikerketa horien emaitzek truke balio handia dute; hortik datorkio, neurketak ematen duen datuari, sektorean jarri ohi zaion terminoa: *audience currency*. Merkatuko eragileek, hots, publizitate saltzaileek (komunikabideek eta hauei gaina hartzen ari zaizkien bestelako plataforma eta euskarriek), zein erosleek (publizitate agentziek eta iragarleek) trukerako onartu behar dituzte ikerkuntzak emandako datuak. Dirua bezala, trukea gerta dadin ezinbestekoa da batzuek eta besteek *currency* edo monetari balioa aitortzea.

2.1.1. Metodologiak eta teknikak

Europako 22 herrialdetako neurtzaileak EMRO (*European Media Research Organisation*) elkartean biltzen dira. Urtero argitaratzen duten Emro Audience Survey Inventory (EASI) datu-baseak erakusten duenez, antzekotasunak eta ezberdintasunak agertzen dira herrialdeen artean (EMRO European Media Research Organisation, 2022). Metodologia eta teknika ugari egon arren, ezaugarri nagusien laburpen bat egin daiteke sistema ezberdinak honako lau ardatzen inguruan sailkatuta:

por esa razón, se gastaron ese mismo año 14.500 millones de euros en investigación de mercados en el Viejo Continente.

Los resultados de estos estudios tienen un gran valor de intercambio; de ahí el término que se le suele poner en el sector al dato obtenido: *audience currency*. Tanto los agentes del mercado, es decir, tanto quienes venden soportes para la publicidad (medios de comunicación y soportes o plataformas que les están sustituyendo en ese ámbito) como quienes los compran (agencias de publicidad y anunciantes) deben acordar el valor de los datos aportados por la investigación. Al igual que con la moneda, para que el intercambio se produzca es imprescindible que unos y otros reconozcan su valor de cambio.

2.1.1. Metodologías y técnicas

Los medidores homologados de audiencias de 22 países europeos se agrupan en la EMRO (European Media Research Organisation). Esta organización publica anualmente la base de datos Emro Audience Survey Inventory (EASI), la cual permite identificar similitudes y diferencias entre mediciones en diferentes países (EMRO European Media Research Organisation, 2022). Así, a pesar de la variedad de metodologías y técnicas, se puede realizar un resumen de las principales características de la investigación clasificando los

- Aztertzen diren komunikabide, euskarri edo komunikazio praktiken arabera: monomedia vs. multimedia.

- Jasotako informazioaren adierazgarritasunaren arabera: laginketa bidezko datuak vs. errolda-datuak.

- Jasotako informazioaren izaeraren arabera: aitortua vs. behatua.

- Jasotako informazioen jatorriaren araberakoa: *user-centric* vs. *site-centric*.

Hurrengo lerroetan ezaugarri horien bereizgarri nagusiak aurkeztuko dira, eta ardatz horiei beste bat gehituko zaie: konbinazioa edo fusioa.

Monomedia vs. multimedia

Azterketa batzuek euskarri, komunikabide edo komunikazio praktika bakarra dute jomugan; adibidez telebista linealaren kontsumoa soilik aztertzen dutenak, edo prentsa idatziarenak, edota sareko ikus-entzunezkoen kontsumoan oinarritzen direnak. Beste batzuek komunikabide, euskarri edo praktika bat baino gehiago aztertzen dituzte aldi berean; horiei multimedia edo *cross-media* deitu ohi zaie, eta tresna bakarrean (adibidez inkesta bakarrean) hainbat komunikabide zein praktikari buruzko informazioa jasotzen da.

diferentes sistemas en torno a cuatro ejes:

- Según los medios de comunicación, soportes o prácticas comunicativas analizadas: monomedia vs. multimedia.

- Según la representatividad de la información recogida: datos muestrales vs. datos censales.

- Según la naturaleza de la información recabada: dato declarado vs. dato observado.

- Según el origen de las informaciones recabada: *user-centric* vs. *site-centric*.

A continuación se presentan los aspectos principales de estos cuatro ejes, a los que se añada otro, la combinación o fusión.

Monomedia vs. multimedia.

Algunos estudios tienen como objeto un único soporte, medio de comunicación o práctica comunicativa. Por ejemplo, aquellos que analizan únicamente el consumo de televisión lineal, prensa escrita, o el consumo de audiovisuales en la red. Otros, por el contrario, estudian simultáneamente más de un medio de comunicación, soporte o práctica. A estos se les suele llamar multimedia o *crossmedia*, y en una sola herramienta (por ejemplo una encuesta) se recoge información sobre diversos medios de comunicación y prácticas.

Badirudi azken urteetan multimedia ikerketak indarra hartzen ari direla euskarrien araberako neurketen aldean, komunikabideen beraien arteko mugak lausotu ahala (Publicom AG, 2017).

Oso ohikoa da bestalde bi eratako iturriak konbinatzea, eta ikusiko dugun bezala konbinazio hori egiteko, sarritan, fusio teknikak erabiltzen dira. Alegia, bateratuta oinarrizko multimedia inkesta batetik (non euskarri ezberdinei buruzko datuak jasotzen diren) eta monomedia inkesta bestetik (euskarri bakarrean fokatuz). Bi inkesten datuak fusionatzeak panorama orokorra eskaintzen du (multimediako datuei esker), baina euskarri zehatz bateko ezagutzan xehetasun gehiago lortuta (monomediako datuei esker).

Laginketa bidezko datuak vs. errolda-datuak

Lazarsfelden lehen lanetatik hona inkesta izan da, zalantzarik gabe, audientzia ikerkuntzako tresnarik erabiliena, baina ez bakarra. Populazio handiei buruzko araketa behar denetan laginak erabili ohi dira helburu horrekin. Gorago aipatu dugun EASI txostenean ikus daitekeenez, oso lagin handiak erabili ohi dira inkesta eta panel ezberdinetan. Erresuma Batuan, adibidez, 6.224 laguneko lagina

La investigación multimedia parece estar ganando peso en los últimos años respecto a las mediciones por soportes individuales, a medida que se van difuminando las hasta hace poco claramente definidas líneas que distinguían un soporte de comunicación de otro (Publicom AG, 2017).

Suele ser habitual, por otra parte, combinar los dos tipos de fuentes y, como veremos, esta combinación se realiza a menudo mediante técnicas de fusión. Es decir, una encuesta multimedia básica (en la que se recogen datos sobre diferentes soportes) y una encuesta monomedia (enfocando un único soporte). Así, la fusión de los datos de las dos encuestas ofrece un panorama general (gracias a los datos multimedia) pero con un mayor detalle en el conocimiento de un soporte concreto (gracias a los datos monomedia).

Datos muestrales vs. datos censales.

Desde los primeros trabajos de Lazarsfeld ha sido la encuesta sin duda la herramienta de investigación de audiencia más utilizada; aunque no la única. Se suelen utilizar muestras con este fin, en la medida en que se pretende una exploración de grandes poblaciones. Como se puede observar en el citado informe EASI, a menudo son utilizadas muestras muy grandes en diferentes encuestas y paneles. En el Reino Unido, por ejemplo, tienen una muestra de

dute multimedia inkestan, baina 100.000koa irratikoan. Alemanian 305.890 multimedian, eta 65.000 irratian.

Baina badira panel mugatuagoak ere, beren ezaugarriengatik zenbaki handietara eraman ezin daitezkeenak.

Laginez gain, errolda-datuak ere erabili ohi dira audientzien ikerkuntzan; era honetako teknika tradizionalen adibide bat prentsaren tiradaren edo zinema aretoetako sarrera-txartelen salmentaren erregistroak lirateke. Datu horien adierazgarritasuna % 100 da jakina, tartean laginik ez dagoelako; baina ematen duten datua ematen dute, bere mugekin: tiradaren kasuan, zenbat ale banatu dituen komunikabide batek (horiek zenbat eta nolako pertsonek edo zein sekzio eta zenbat denbora irakurri diren jakiterik ez dagoelarik); sarrera salmentaren kasuan, berriz, zenbat ikuslek ikusi duten aretoan proiektatutakoa (baina ez nolakoa edo zer gustatu zaien), eta abar.

Datu aitortuak vs. datu behatuak

Inkesten bidez, normalean, informatzaileak aitortzen duen datua jasotzen da: hark esaten digu halako komunikabidea erabili duen, noiz eta zenbat, zer interpretatu duen edo zer gustatu zaion, eta beste. Jakina, erantzun horren azpian faktore baldintzatzaile ugari egon dai-

6.224 personas en la encuesta multimedia, pero 100.000 en la de radio. En Alemania 305.890 en multimedia, y 65.000 en radio.

Se recurre también a paneles más limitados que por sus características no se pueden llevar a grandes contingentes poblacionales.

Además de las muestras, la otra gran fuente de la investigación de audiencias son los datos censales. Un ejemplo tradicional lo constituyen los registros de la tirada de la prensa o del taquillaje en salas de cine. La representatividad de estos datos es del 100%, al no estar basados en muestras, pero normalmente aportan un dato muy limitado: en el caso de la tirada, cuántos ejemplares ha distribuido un medio de comunicación (sin que se pueda saber cuántas y qué tipo de personas lo han leído, ni qué secciones, ni cuánto tiempo, etcétera); o cuántas personas han visto lo proyectado en la sala (pero no si eran mujeres u hombres, qué les ha gustado), etc.

Dato declarado vs. dato observado.

A través de las encuestas, habitualmente, se recoge el dato proporcionado por la persona informante: es ella quien nos dice si ha utilizado tal medio de comunicación, cuándo y cuánto, qué ha interpretado o qué le ha gustado, etcétera. Por supuesto, tras esta respuesta pueden existir

teke: informatzailearen oroimena, emango den erantzunari buruzko aurreiritzi soziala, eta abar luzea.

Datuak jasotzeko honako ohiko teknikak erabiltzen dira:

- Aurrez aurreko elkarrizketak: PAPI (*Paper and Pencil Interviewing*) eta CAPI (*Computer-Assisted Personal Interviewing*) —azken hori tableta erabiliz—.
- Telefono bidezkoak: CATI (*Computer-Assisted Telephone Interviewing*).
- Web bidezkoak: CAWI (*Computer Aided Web Interviewing*). Mota honetakoak gero eta gehiago erabiltzen dira, merkeagoak direlako; baina zalantzak daude sortzen duten desbideraketaren inguruan —teknologia gehiago erabiltzen duten pertsonen aldeko desbideraketa alegia—.

Beste kasu batzuetan, aldiz, behatutako datuekin egiten da lan; hau da, behatu, frogatu eta kontabilizatu daitezkeen datuekin: zerbait gertatu izanaren ziurtasunarekin egiten da lan (tipologia honetakoa dira esaterako salmenta bat, klik bat, eta abar).

User-centric vs. site-centric

Komunikabideen kontsumoari buruzko datuak jasotzeko iturri posible bat kontsumitzaile

múltiples factores que la condicionan: su recuerdo, el prejuicio social sobre la respuesta que se va a dar, y un largo etcétera.

Para la recogida de datos mediante encuestas, por otro lado, se utilizan diferentes técnicas:

- Entrevistas cara a cara: PAPI (*Paper and Pencil Interviewing*) y CAPI (*Computer-Assisted Personal Interviewing*), esta última utilizando una tablet.
- Telefónicas: CATI (*Computer-Assisted Telephone Interviewing*).
- Vía web: CAWI (*Computer Aided* Web Interviewing). Cada vez más utilizada por resultar más económica, aunque pueden plantearse dudas sobre la desviación que pudiera generarse (debido al sesgo a favor de las personas con mayores habilidades tecnológicas).

En otros casos, sin embargo, se trabaja con datos reales, es decir, con los datos observables, demostrables y contabilizables: se tiene la certeza de que ha sucedido algo (como son una venta, un clic, etc.).

User-centric vs. site-centric.

En las técnicas *user-centric*, la fuente para la recogida de datos sobre el consumo de medios de

potentzialak dira (*user-centric*): haiei galdetzen edo behatzen zaie halako kontsumoa edo halako praktika egin duten.

Beste iturri posible bat komunikabidea bera da (*site-centric*): hark daki zenbat ale zabaldu dituen, zenbat klik egin dioten, eta abar.

Tradizionalki biak erabili izan dira audientzien ikerkuntzan. Populazioari eginiko inkestak adibidez, *user-centric* gisa sailka ditzakegu; eta tiradaren kontrola, *site-centric*. Hala ere, bi termino hauek nagusiki sareko portaerei lotzen zaizkie gehiago. Lehen kasuan, ohikoa izaten da erabiltzaile panel bati software pieza bat instalatzea beren gailuetan (telefono adimentsua, tableta, ordenagailua...), eta hark jasotzen du sareko portaera. Bigarren kasuan, webgune edo sare bakoitzak analitika digitalerako tresnak ditu, jasotzen dituen klik guztiak erregistratu eta aztertzeko.

Berriro ere, metodologia bakoitzak bere aukerak eta bere mugak ditu.

Konbinazioak eta fusioak

Tradizionalki metodo konbinatuak erabili izan dira audientzien ikerketan lehenxeago adierazi

comunicación son las personas consumidoras potenciales; es a ellas a quienes se consulta u observa para concluir si han realizado este consumo o tal práctica.

En las técnicas *site-centric*, por el contrario, la mirada se dirige hacia el propio medio de comunicación o soporte, que es quien conoce el número de ejemplares que ha difundido, el número de clicks recibidos, etc.

Tradicionalmente ambas han sido utilizadas en la investigación de audiencias. Las encuestas a la población, por ejemplo, pueden considerarse *user-centric*; mientras que el control de la tirada es *site-centric*. Sin embargo, estos dos términos se asocian más a los actuales comportamientos en la red. En las primeras, es habitual que instale una pieza de software en los dispositivos (*smartphone*, ordenador, tablet...) de un panel de personas usuarias, el cual registra el comportamiento en la red. En los segundos, cada sitio web o red social cuenta con herramientas de analítica digital que registran y analizan todos los contactos que recibe.

Una vez más, aquí también cada metodología tiene sus posibilidades y sus limitaciones.

Combinaciones y fusiones.

Tradicionalmente ha sido habitual el uso combinado de diferentes métodos en la investiga-

den gisan: monomedia inkestak multimedia inkestekin osatuta, datu aitortuak datu behatuekin konbinatuta, eta abar. Ardatz horietako kategorien gurutzaketa ezberdinekin sistema ezberdinak sortzen dira.

Bestalde, eraldaketa handi bat suertatzen ari da audientzien neurketan azken hamarkadetan; komunikazio praktiken eraldakuntza sakonaren ondorioz bezain beste, eskura dauden datuen ugaritasunaren zein datuak jaso eta tratatzeko teknologia eta tekniken ondorioz. Horrek bultzatu du, geroago ikusiko dugun moduan, iturri ezberdinetatik datozen eta metodologia ezberdinekin jaso diren datuen fusioa (*statistical matching*) egiteko.

2.1.2. Antolaketa, gobernantza eta kostuak

Europan zein mundu osoan, audientziak neurtzen dituzten sistemak hiru motakoak izaten dira, jabetzari dagokionez:

- **JIC** (*Join Industry Commitee*): merkatuak —komunikabideen, publiz itate agentzien eta iragarleen sindikazioak— kontrolatzen eta finantzatzen du datuen ekoizpena. Europan eredu hedatuena da. Gure adibide hurbilena AIMC-EGM eta CESP lirateke.

ción de audiencias: encuestas monomedia complementadas con encuestas multimedia, datos declarados combinados con datos observados, etc. Los diferentes cruces de categorías de estos ejes se generan a su vez diferentes sistemas de medición.

Por otra parte, en las últimas décadas se está produciendo un gran cambio en la medición de audiencias, debido por un lado a la abundancia de datos disponibles y al desarrollo de las tecnologías y técnicas de recogida-tratamiento de datos; y por otro a la profunda y rápida transformación de las prácticas de comunicación. Esto ha promovido, como veremos más adelante, la fusión (o *statistical matching*) de datos procedentes de diferentes fuentes, los cuales a su vez han sido recogidos mediante diferentes técnicas.

2.1.2. Organización, gobernanza y costes

Tanto en Europa como en el resto del mundo, los sistemas que miden las audiencias se clasifican en tres tipos en términos de gobernanza:

- **JIC** (*Join Industry Commitee*): la generación de datos está controlada y financiada por el mercado (sindicación de medios de comunicación, agencias de publicidad y anunciantes). Es el modelo más extendido en Europa. Nuestro ejemplo más cercano serían AIMC-EGM y CESP.

- **RA** (*Research Agency*): agentzia batek —enpresa batek— datuak eskuratu, ekoitzi eta saltzen ditu. Adibideak: Kantar, Comscore edo CIES.

- **MOC** (*Media Owner Contract*): sektore batek —komunikabide talde batek edo komunikabide bakar batek— kontrolatzen eta finantzatzen du datuen ekoizpena. Adibidez OJD (prentsa eta aldizkari enpresek eratutako sozietate autonomoa).

- **Erakunde publikoen ikerketa ofizialak:** populazioaren osaera, portaerak eta kontsumoak kulturalak eta medioatikoak ikertzen dituzten erakunde publikoak dira. Adibidez: EUSTAT, NASTAT, INSS, INE edota Europa mailako EUROSTAT.

Ikus daitekeenez, merkatuak arautzen du —formula ezberdinekin— audientzien neurketa. Herrialde gutxi batzuetan ordea erregulatzaile publikoek badute ardurarik audientzien neurketan (Italian kasu); beste batzuetan emandako emakiden edo finantzazio publikoaren truke komunikabideek beren funtzio soziala ondo betetzen duten aztertzen da (Erresuma Batuan adibidez).

Ia herrialde guztietan audientzia neurtzaile bakarra dago, euskarri bakoitzean (telebista, irratia, prentsa, sarea).

Lagin eta metodologia horiekin egindako lanak, bistan da, ga-

- **RA** (*Research Agency*): una agencia (una empresa) obtiene, produce y vende los datos. Ejemplos: Kantar, Comscore o CIES.

- **MOC** (*Media Owner Contract*): la producción de datos está controlada y financiada por un sector (un grupo de medios de comunicación o un único medio). Por ejemplo, OJD (sociedad autónoma constituida por empresas de prensa y revistas).

- **Estudios oficiales de instituciones públicas:** son los organismos públicos quienes investigan la composición de la población, así como los comportamientos y consumos culturales y mediáticos. Por ejemplo: EUSTAT, NASTAT, INSS, INE o, a nivel europeo, EUROSTAT.

Puede deducirse pues que es el mercado quien regula, con distintas fórmulas, la medición de audiencias. En unos pocos países, sin embargo, los reguladores públicos tienen alguna responsabilidad en la medición de audiencias (como es el caso de Italia); en otros desde lo público se analiza si los medios de comunicación cumplen bien su función social, a cambio de concesiones o financiación pública (como en el Reino Unido).

Por otra parte, lo habitual es que haya un único medidor de audiencia en cada país y para cada soporte (televisión, radio, prensa, red).

restiak dira. Neurketa agentziek haien aurrekontuak argitaratzeko uzkur jokatzen badute ere, zenbait datu argitaratuta daude (Publicom AG, 2017): Austriako 4 milioi eurotatik Frantziako 10 milioi eurotara mugitzen dira aurrekontuak, beti ere prentsaren audientzien kasuan (telebistaren audientzien neurketa garestiagoa izan ohi da).

2.2. Publizitate merkatua helburu duten sistemen mugak

Sistema hauek indartsuak dira izan ere. Mugak dituzte ordea publizitatearen merkaturako txanpon gisa erabiltzetik gizartearen komunikazio ohiturak neurtzeko tresna gisa erabiltzera igaro nahi dugunean.

Alde batetik, publizitatearen merkatuaren beharretara egokituta daude. Aztertutako gaiak eta egindako analisiak baldintzatzen ditu ezaugarri horrek. Ikerketa aplikatua da zentzu horretan, eta akademiaren interes zientifikoaren ikuspegitik, hainbatetan, emaitza mugatukoa. Aplikazioa, bestalde, publizitatearen arloari begirakoa da; komunikazio politika publiko batetik edo are hizkuntza politika batetik beharko litzatekeen ikerkuntzatik at egiten dena.

Este tipo de investigación basada en grandes muestras es, evidentemente, económicamente muy costoso. Si bien las agencias de medición se muestran reticentes a publicar sus presupuestos anuales, contamos con unos pocos datos publicados (Publicom AG, 2017). Entre ellos, los presupuestos se mueven de 4 millones de euros en Austria a 10 millones de euros en Francia, siempre en el caso de las audiencias de la prensa (la medición de audiencias de la televisión suele ser más cara).

2.2. Limitaciones de los sistemas orientados al mercado publicitario

Estos sistemas son robustos, sin duda alguna. Sin embargo, tienen limitaciones cuando queremos pasar de utilizarlos como moneda para el mercado de la publicidad a utilizarlos como herramienta para medir los hábitos de comunicación de la sociedad.

Por un lado están adaptados a las necesidades del mercado de la publicidad. Esto condiciona los temas y los análisis realizados. Se trata, en este sentido, de investigación aplicada, y desde el punto de vista del interés científico de la academia puede ofrecer con frecuencia resultados limitados. Su aplicación está orientada al ámbito de la publicidad, ámbito diferente al de las políticas públicas de comunicación o de las políticas lingüísticas.

Metodologiaren ikuspegitik, bestalde, inkesta bidez lortutako datu aitortuek ez dute, sarritan, zuzenean behatutako errolda-datuek lortzen duten zehaztasuna ematen. Azken horiek, ordea, ez dute inkestek ematen duten aberastasunik (adibidez datu soziodemografiko ugari, medio eta plataforma ezberdinetako erabilera, eta abar). Tekniken arteko ezberdintasunak gainditzeko metodologia berriak garatzen ari dira, aurrerago ikusiko dugunez; eta bide horretatik joango da gure lana ere.

Publizitateari begirako audientzia ikerkuntzaz gain, jakina, beste ikerketa ildo asko burutzen dira akademiatik zein beste arlo batzuetatik. baina kontuan hartu behar da ordea finantzazioa arazo handia dela oinarrizko ikerkuntzan. CIESek Araba, Bizkaia, Gipuzkoa eta Nafarroan urtero egiten duen 8.600 laguneko inkesta ez ohikoa litzateke beste ikergune askotan, eta gauza bera gertatuko litzateke AIMCk Espainian urtero egiten duen 279.000 laguneko galdeketarekin (Asociación para la Investigación de Medios de Comunicación AIMC, 2022)[2]. Komunikazio azturak bestalde oso arin aldatzen ari dira, 1. grafikoak erakutsi duen moduan, eta etengabeko jarraipena be-

2 Ikus xehetasunak 4.4. azpiatalean. Bestalde gogora dezagun Espainiako *Centro de Investigaciones Sociológicas* (CIS) ikerguneak urtero egiten duen Barometroak 4.000 laguneko laginak erabiltzen dituela.

Por otra parte, desde el punto de vista metodológico, los datos declarados obtenidos a través de la encuesta no proporcionan a menudo la precisión que obtienen los datos censales observados directamente. Estos últimos, sin embargo, carecen de la riqueza que aportan las encuestas (por ejemplo, numerosos datos sociodemográficos, información sobre el uso de diferentes medios y plataformas, etc.). Para superar estas diferencias entre técnicas se están desarrollando nuevas metodologías, como veremos más adelante. Y por ahí transitará también nuestro trabajo.

Además de la investigación de audiencia publicitaria, por supuesto, otras muchas líneas de investigación se llevan a cabo tanto desde la academia como desde otros ámbitos. Pero hay que tener en cuenta que la financiación es un gran problema para la investigación básica. La encuesta anual de 8.600 personas que realiza CIES en Araba, Bizkaia, Gipuzkoa y Nafarroa sería considerada inusual en muchos centros de investigación, y lo mismo ocurriría con la encuesta anual de 279.000 personas que realiza AIMC en España (Asociación para la Investigación de Medios de Comunicación AIMC, 2022)[2]. Por otra parte, los hábitos de comunicación están cambiando muy rápidamente, como hemos mos-

2 Ver detalles en el subapartado 4.4. Por otro lado, cabe recordar que el Barómetro anual del Centro de Investigaciones Sociológicas (CIS) utiliza muestras de 4.000 personas.

harko litzateke suertatzen ari diren aldaketen ezagutza zehatza eskuratzeko. Tamaina horietako inkestak nahikoa maiztasunez egitea —adibidez urtero— oso urrun dago edozein ikerguneren ahalmenetatik.

Audientzia ikerkuntzak bestalde, inbertsio handiak eta tresna indartsuak erabili arren, zailtasunak izaten ditu komunitate txikietako komunikazio praktikak neurtzeko. Euskal Herriari dagokionez, aipatu ditugun sistema batzuek oso gutxi begiratzen diote, adibidez, euskarazko komunikazioari; lagin handiak edo tresna teknologiko sendoak erabili arren analisia ezin baita jaitsi zenbaitetan kasu horien mailara, eta beste batzuetan ez da ikuspuntu hori txertatzen analisian.

2.2.1. Hizkuntza gutxituetako komunikazio praktikak ezagutzeko mugak

Esan bezala, publizitatearen negozioak gidatzen du komunikabideen audientzien ikerkuntza gehiena, aspalditik. Merkatuko eragileak dira, egitura ezberdinen bidez, komunikazioaren kontsumoaren eta erabileraren ikerkuntza gehien egiten dutenak.

Komunikazioa, publizitatea txertatzeko baino gehiago beste helburu sozialetarako baliatu nahi duten erakundeek bi aukerako litzateke suertatzen ari diren aldaketen ezagutza zehatza eskuratzeko. Tamaina horietako inkestak nahikoa maiztasunez egitea —adibidez urtero— oso urrun dago edozein ikerguneren ahalmenetatik.

Audientzia ikerkuntzak bestalde, inbertsio handiak eta tresna indartsuak erabili arren, zailtasunak izaten ditu komunitate txikietako komunikazio praktikak neurtzeko. Euskal Herriari dagokionez, aipatu ditugun sistema batzuek oso gutxi begiratzen diote, adibidez, euskarazko komunikazioari; lagin handiak edo tresna teknologiko sendoak erabili arren analisia ezin baita jaitsi zenbaitetan kasu horien mailara, eta beste batzuetan ez da ikuspuntu hori txertatzen analisian.

2.2.1. Limitaciones en el estudio de las prácticas comunicativas en lenguas minorizadas

Como ya se ha indicado, el negocio de la publicidad es el que más tiempo lleva liderando la investigación de las audiencias de los medios de comunicación. Los agentes del mercado son, a través de las diferentes estructuras existentes, los que realizan una mayor investigación sobre el consumo y uso de la comunicación.

Los organismos que pretendan utilizar la investigación sobre comunicación para otros fines socia-

ra dituzte eragin-esparru duten errealitate hori ezagutzeko: jadanik dauden informazioa eta ezagutza baliatu; edo ikerkuntza propioa egin. Kontuan hartu behar da edozein ezagutza adarretan gertatzen den bezala, ikerketaren atzetik dauden helburuek moldatu egiten dituztela emaitzak. Adibidez, publizitatearen merkaturako komunikabide bakar baten intzidentzia datu garrantzitsua da (eskuratzen duen arretaren terminoetan, hartzaile eta denbora kopurutan neurtua); baina hizkuntza plangintzaren ikuspegitik hori bezain interesgarria izan liteke komunikabideek, sistema gisa hartuta, hizkuntzari egiten dioten ekarpenaren berri izatea.

Ildo beretik gauza jakina da komunikabide handiek, audientziak neurtzeko inkestak urteko zein sasoitan egiten diren jakitun, sasoi horietan esfortzu handiagoa egin ohi dutela zabalkundea handitzeko (edota, erakunde neurtzaileetan eskuhartzerik badute, neurketak haien zabalkunde handienetako garaietan egin daitezen baldintzatzeko). Gauza jakina da baita ere audientzia erosle potentzial gisa ikusten duenak arreta berezia jarriko diola erosketa ahalmena duenari, eta ez hain handia erosketa ahalmen txikia duenari.

Arazo hori aski ezaguna da hizkuntza gutxitua duten

les más allá de la publicidad tienen dos posibilidades de conocer esa realidad: bien recurrir a la información y el conocimiento existentes, o bien recurrir a la investigación propia. Hay que tener en cuenta que, al igual que cualquier conocimiento, el objetivo perseguido por la investigación va a modelar los resultados. Por ejemplo, la penetración de un único medio de comunicación para el mercado publicitario es un dato importante (medido en términos de atracción que genera, en términos de número de *receptores* y de tiempo dedicado), pero desde el punto de vista de la planificación lingüística puede ser más interesante conocer la contribución que el sistema de los medios de comunicación hace en su conjunto a la lengua, como sistema y no individualmente.

En la misma línea es sabido que los grandes medios de comunicación, sabedores de los momentos del año en la que se realizan las encuestas de medición de audiencias, suelen realizar en esos momentos un mayor esfuerzo por aumentar su difusión (o, si pueden, intentan intervenir en las instituciones medidoras para condicionar que las mediciones se realicen en las épocas de mayor difusión). También es sabido que quien vea en la audiencia un potencial comprador prestará especial atención a aquellos sectores con mayor poder adquisitivo frente al resto.

Estos problemas son bien conocidos en las comunidades de

Mendebaldeko Europako komunitateetan. Euskal Herritik begiratuta, Katalunia, Gales eta Galizia izan daitezke erreferente nagusiak, hizkuntza gutxituetako komunikabideei dagokienez; izan ere, horiei buruz egindako ikerketa konparatiboan (Zabaleta et al., 2018), euskarazko 123 komunikabide identifikatu ziren —gurea izanik, katalanaren atzetik (751), bigarrena komunikabide kopuruari dagokienean; hirugarrena galesera zen, 84 komunikabiderekin; eta laugarrena galegoa, 60rekin; irlandera eta bretoiera zetozen jarraian 12na komunikabiderekin, eta beste hainbat hortik beherako zenbakiekin—.

Lau komunitate handienetatik bakarrean egon da saio esanguratsurik bertako hizkuntzako audientziak sakon aztertzeko: Katalunian, 2007tik 2012ra abian egon zen *Baròmetre de la Comunicació i la Cultura* lanarekin (Sabaté, 2011, 2020). Labur esanda, lagin handiko inkesta batean oinarritutako urteroko txosten sorta bat zen, baina arrazoi ezberdinak direla medio itxi egin zen: bereziki aurrekontu handia batetik, eta lurralde berean bi datu —beraz bi *currency*— egoteak sortzen zuen ziurgabetasuna bestetik (Amezaga Albizu, 2022).

Jarraian, Euskal Herrian komunikazio praktikak ikertzeko eskura

Europa Occidental con lenguas minorizadas. Visto desde el País Vasco, los principales referentes en cuanto a medios de comunicación en lenguas minoritarias son Cataluña, Gales y Galicia. En el estudio comparativo realizado sobre los medios en lenguas europeas minorizadas, Zabaleta y otros autores (Zabaleta et al., 2018), identificaron 123 medios de comunicación en euskera —siendo esta lengua, por detrás del catalán (751), la segunda en número de medios de comunicación. En tercer puesto se encuentra el galés, con 84 medios de comunicación. Y en el cuarto el gallego, con 60. Les siguen el irlandés y el bretón con 12 medios de comunicación, contando el resto de lenguas con cifras aún más bajas—.

Sólo en una de estas cuatro comunidades lingüísticas ha habido intentos significativos para analizar en profundidad las audiencias en lengua propia: en Cataluña, con el *Baròmetre de la Comunicació i la Cultura,* que se publicó desde 2007 hasta 2012 (Sabaté, 2011, 2020). Se trataba de una serie de informes anuales basados en una encuesta muestral amplia, que se truncó por diferentes razones: sobre todo por un elevado presupuesto y por la incertidumbre que generaba la existencia de dos datos – por tanto dos *currency*- en un mismo territorio (Amezaga Albizu, 2022).

A continuación se presentan las principales herramientas dispo-

dauden tresna nagusiak aurkezten dira.

2.3. Audientzia ikerketak Hegoaldeko Euskal Herrian, gaur

Duela ia mende bat bezala, audientzia ikerkuntza zientifikoa abiatu zenean alegia, orain ere aldaketa garai betean dago komunikazioaren eremua. Euskarri eta ohitura berriak komunikabide tradizionalek betetzen zituzten espazioak okupatzen ari dira (baita komunikabideetatik at zeudenak ere, hala nola familiaren, eskolaren eta lagunartearen espazioak). Aldaera horren ondorio sozial, kultural zein politikoak ez dira lehen baino txikiagoak, inondik inora ere; eta prozesu horiek ezagutzeko beharra begi bistakoa da.

Bestalde, inoiz baino informazio gehiago dugu gizarte fenomenoak aztertzeko —komunikazio jarduerak barne— eta horrek aukera berriak irekitzen dizkio begirada soziologikoari.

Euskal Herrira etorrita, eta komunikazioaren arlora mugatuta, informazio iturri anitz daude. Hurrengo lerroetan horietako garrantzitsuenak laburbilduko ditugu.

nibles hoy en el País Vasco para la investigación de las prácticas comunicativas.

2.3. Estudios de audiencia en Araba, Bizkaia, Gipuzkoa y Nafarroa, hoy

Al igual que hace casi un siglo, cuando se inició la investigación científica de las audiencias, el campo de la comunicación social se encuentra en plena época de cambio. Los nuevos soportes y prácticas comunicativas están ocupando espacios que anteriormente ocupados por los medios de comunicación tradicionales (así como otros espacios más allá de los medios, como la familia, la escuela y el grupo de pares). Sus consecuencias sociales, culturales y políticas no son en absoluto menores que antes. Y la necesidad de conocer estos procesos parece obvia.

Por otro lado, tenemos más información que nunca para analizar los fenómenos sociales, incluidas las actividades de comunicación. Y eso abre nuevas posibilidades a la mirada sociológica.

En País Vasco, y ceñidos al ámbito de la comunicación, existen m últiples fuentes de información. A continuación se resumen los aspectos más relevantes de las mismas.

2.3.1. CIES (Euskal Autonomia Erkidegoko eta Nafarroako komunikabideen audientziari buruzko azterlana)

Jose Ignacio Ruiz Olabuenagaren bi ikasle ohik, Carlos Zufiak eta Fernando Lacabek, abian jarri zuten Deustuko Soziologia Fakultatean ikasitakoa, eta CIES enpresa sortu zuten 1981ean Iruñean. 1984an lehen audientzia ikerketa burutu zuten, Araba, Bizkaia, Gipuzkoa eta Nafarroako 4.887 lagunek osatutako lagina baliatuta. Geroztik urtero-urtero egin dute datu-bilketa[3], lagina handituz urteak igaro ahala (gaur egun 8.600 lagunekoa da) eta bilakaera historikoa aztertzea ahalbidetzen duen egitura mantenduz. Ezaugarri horiek Euskal Herriari buruzko datu-base garrantzitsuenetako bat bihurtzen dute[4].

Inkestak, hasieratik, datu soziodemografikoak, demolinguistikoak (euskararen ezagutzari dagozkionak), ekipamenduzkoak eta komunikabideen kontsumozkoak biltzen ditu, nazioartean audientzien neurtzaileek erabiltzen dituzten estandar berberetan. Publizitatearen

3 Bi salbuespenekin: 1986an ez zen inkestarik egin; eta 2020an, COVID-19aren pandemia medio, lagina erdira jaitsi zen (4.300), udaberriko uhina ezin izan zelako egin.

4 Euskal Hedabideen Behategiak (*behategia.eus*) datu historikoen bilduma egin du, 1984tik aurrerako serie historiko osoak bilduz. Datuok une honetan Euskararen Adierazle Sisteman kargatzen ari dira, eta modu irekian kontsulta daitezke (*https://labur.eus/HbRPO*).

2.3.1. CIES (Estudio de Audiencia de Medios de Comunicación en la C.A. de Euskadi y Nafarroa)

Dos antiguos alumnos de José Ignacio Ruiz Olabuenaga, Carlos Zufia y Fernando Lacabe, pusieron en marcha lo aprendido en la Facultad de Sociología de Deusto y montaron hace casi cuatro décadas la empresa CIES en Pamplona. En 1984 se realizó el primer estudio de audiencia con una muestra de 4.887 personas de Araba, Bizkaia, Gipuzkoa y Nafarroa. Desde entonces, este estudio se ha realizado anualmente[3], ampliando la muestra (actualmente es de 8.600 personas) y manteniendo una estructura que permite analizar la evolución histórica. Estas características las convierten en una de las bases de datos más importantes del País Vasco[4].

La encuesta recoge desde sus comienzos datos sociodemográficos, demolingüísticos (referidos al conocimiento del euskera), así como de equipamiento y de consumo de medios de comunicación, en los mismos estándares que los medidores de audiencias

3 Con dos excepciones: en 1986 no se realizó la encuesta. Y en 2020, debido a la pandemia del COVID-19, la muestra fue la mitad de la habitual (4.300) debido a la imposibilidad de realizar la ola de primavera.

4 El Observatorio de los Medios de Comunicación en Euskera (*behategia.eus*) ha realizado una recopilación de series históricas enteras desde su existencia. Estos datos se están cargando actualmente en el Sistema de Indicadores del Euskera y se pueden consultar en abierto (*https://labur.eus/HbRPO*).

merkatuan —iragarleek, publizitate agentziek eta komunikabideek— onartutako datua eskaintzen dute, Europan ezohikoa den egoera sortuz. Izan ere, Europan neurtzaile bakarra egon ohi da estatu bakoitzean (medio guztietarako bakarra edo medio bakoitzerako bakarra), datu homogeneoa mantentze aldera. Espainiako Estatuan alabaina estatu mailako agentziez gain (AIMC komunikabide guztietarako, eta Kantar Media telebistarako) CIES dago, Euskal Herriko hegoalderako, eta aurrekoen homologazio maila berbera lortu du. Honek egoera pribilegiatuan jartzen gaitu, gure errealitate soziala ezagutzeari dagokionez.

Inkesta izanik CIESen funtsa, jasotako datuak aitortuak dira, eta lagin bidez ateratakoak. Multimedia motako inkesta da, eta hortaz, euskarri ezberdinen kontsumoa neurtzen du.

Galderak erantzungo dituen lagina aukeratzeko laginketa aleatorio estratifikatu bat erabiltzen da —honako estratifikazio-aldagaiak kontuan hartuta: sexua, adina (bost urteko taldeetan), probintzia, habitata eta eskualdea—.

CIES inkestak 200 galdera inguru ditu, tematikoki 7 talde nagusitan banatuta:

1. Soziodemografikoak.
2. Prentsa.

a nivel internacional. Ofrece un dato reconocido en el mercado de la publicidad (anunciantes, agencias de publicidad y medios de comunicación), creando una situación no habitual en Europa. De hecho, en Europa lo habitual es que haya un solo medidor en cada país (único para todos los medios o único para cada medio) lo que permite la existencia de un dato homogéneo y aceptado. En el Estado español, sin embargo, además de las agencias estatales (AIMC para todos los medios de comunicación y Kantar Media para la televisión), se encuentra CIES, para las provincias peninsulares del País Vasco, que ha alcanzado el mismo nivel de homologación que los anteriores. Esto nos sitúa en una posición privilegiada, de referencia al conocimiento de nuestra realidad social.

Tratándose de una encuesta, los datos obtenidos son declarados y muestrales. Es una encuesta multimedia (que mide el consumo de los diferentes soportes).

Para elegir dicha muestra, se utiliza un muestreo aleatorio estratificado en el que las variables de estratificación son el sexo, la edad (en grupos quinquenales), la provincia, el hábitat y la comarca.

La encuesta de CIES consta de alrededor de 200 preguntas divididas en 7 grupos principales:

1. Sociodemográficas.
2. Prensa.

a. Paperezko prentsa.
b. Egunkari digitalak.
3. Irratia.
4. Telebista.
a. Orotariko telebista.
b. Tokiko telebista.
5. Aldizkako argitalpenak.
6. Internet: konexio-bitartekoak, web-orriak eta Interneteko zerbitzuak.
7. Ekipamendua.

2.3.2. AIMC (Estudio General de Medios EGM)

Espainiako audientzien neurtzaile erreferentzial bilakatu da eta horrela aitortzen diote eragile desberdinek[5] (telebistaren kasuan Kantar Mediarekin batera, nahiz eta alderdi ezberdinak neurtzen dituzten) eta 1968an sortu zen. CIESen txostenaren antzeko egitura du, euskarri batzuetan (aldizkariak adibidez) xehetasun gehiago emanez eta lehenak ez dituen beste atal batzuk (kaleko publizitatea esaterako) kontuan hartuz.

CIESen multimedia inkestaren aldean, EGM inkesta ezberdinek osatua da: bata multimedia eta besteak euskarrikakoak (prentsa, irratia, aldizkariak eta telebista). Denak konbinatuz, 8.435 laguneko lagina osatzen du Araba, Bizkaia, Gipuzkoa eta Nafarroan.

5 Erreferentzialtasuna ez dagokio izendapen administratibo bati, merkatuaren aitortzari baizik.

a. Prensa en papel.
b. Diarios digitales.
3. Radio.
4. Televisión.
a. Televisión generalista.
b. Televisión local.
5. Publicaciones periódicas.
6. Internet: medios de conexión, páginas web y servicios de Internet.
7. Equipamiento.

2.3.2. AIMC (Estudio General de Medios EGM)

La Asociación de Investigación de Medios de Comunicación AIMC es el medidor referencial de las audiencias en España[5] (en el caso de la televisión, junto con Kantar Media, aunque miden aspectos diferentes) y se creó en 1968. Su estudio EGM, ya cásico, cuenta con una estructura similar al informe del CIES, con mayor detalle en algunos soportes (revistas, por ejemplo) y teniendo en cuenta otros apartados que no tiene aquel (como la publicidad en la calle).

A diferencia de la encuesta multimedia de CIES, el EGM está compuesto por diferentes encuestas, una multimedia y otras monomedia (prensa, radio, revistas y televisión). En combinación, la muestra alcanza un tamaño de 8.435 personas en Araba, Bizkaia, Gipuzkoa y Nafarroa.

5 La referencialidad no se refiere a una denominación administrativa, sino al reconocimiento por parte del mercado.

Aurrekoa bezala, jasotako datuak aitortuak dira, eta lagin bidez ateratakoak. Multimedia inkesta da (Arana Arrieta, 2011).

Aipatzekoa da AIMCk datu fusioen arloan egiten duen lana, bereziki Comscore enpresarekin batera —aurrerago emango dira xehetasunak—.

2.3.3. Applika+ proiektua (IKUSIKER panela)

EHUko *Applika+* proiektua 2017an sortu zen, helburu nagusia euskal gazteen ikus-entzunezkoen kontsumoak zein *online* jarduera aztertzea zuelarik. EHUrekin batera, orain arteko bi edizioetan eragile ezberdinek kofinantzatu dute: EiTB, Tabakalera, Hekimen, eta Kulturaren Euskal Behatokia. 2019an IKUSIKER panela jarri zen abian, unibertsitateko ikasleen kontsumoak neurtzeko asmoz. Laginketa hedatu eta gaur egun ia 3.000 laguneko panela osatzen da, 11 urtetik 23 urtera bitarteko gazteekin. DBHn, Batxilergoan eta unibertsitatean ikasten ari diren gazteak dira, Araba, Bizkaia, Gipuzkoa eta Nafarroan. Unibertsitarien kasuan, Whatsapp bidez aldiro-aldiro (urtean zortzi aldiz) inkesta labur baterako esteka igortzen zaie, eta berehalako erantzunak jaso eta erregistratzen dira, erantzuntasa altuekin. Bigarren Hezkuntzako ikasleen kasuan, inkestaldia urtean birritan egiten da. Azken hauen kasuan, 2022-2023

Al igual que el anterior, se basa pues en datos declarados, obtenidos mediante encuesta, y multimedia (Arana Arrieta, 2011).

Destaca la labor de AIMC en el campo de la fusión de datos, especialmente con la empresa Comscore, la cual se detallará más adelante.

2.3.3. Proyecto Applika+ (panel IKUSIKER)

El proyecto *Applika+* de la UPV/EHU se lanzó en 2017 con el objetivo principal de analizar el consumo audiovisual y la actividad online de las y los jóvenes vascos. Junto con la UPV/EHU, en las dos ediciones anteriores participan diversos agentes: EiTB, Tabakalera, Hekimen, y el Observatorio Vasco de la Cultura. En 2019 se puso en marcha el panel IKUSIKER para medir los consumos de las y los estudiantes universitarios. Posteriormente fue ampliado a sectores más jóvenes, y en la actualidad se completa un panel de casi 3.000 personas de 11 a 23 años. Se trata de jóvenes que cursan estudios de ESO, Bachillerato y Universidad en Araba, Bizkaia, Gipuzkoa y Nafarroa. En el caso de las personas universitarias, se les envía regularmente (ocho veces al año) y a través de Whatsapp un enlace a una breve encuesta y se recogen y registran respuestas inmediatas, con un alto nivel de retorno. En el caso del alumnado de Secundaria, la encuesta se realiza dos veces al año. También

ikasturtetik aurrera, Iparraldeko ikasketa-zetroetako gazteak kontuan hartzen ditu bere azterketan.

Lagina aukeratzeko, laginketa aleatorio estratifikatu bat erabiltzen da, non estratifikazio-aldagaiak sexua, adina (bi urteko taldeetan) eta probintzia diren. Inkestak egin ondoren, panelkide bakoitzak duen kode pertsonal baten bidez lotzen dira, galdera guztien informazioa biltzen duen datu-base bat lortzeko. Jasotako informazioak gazteen ikus-entzunezkoen kontsumoen berri ematen du, besteak beste kontsumo oroitua baliatuz: zein eduki (fikzioa, informazioa, eta abar), zein euskarri (*streaming*, VOD, telebista lineala, eta abar), zein hizkuntza (audioan zein azpidatzietan), noiz eta non, norekin eta bestelakoak erregistratuz. Horrez gain, gai monografikoak ere lantzen dira inkestetan lantzean behin —sare sozialak (Instagram edo Twitch kasu), COVID-19aren inpaktua, informazioaren garrantzia gazteen artean, eta beste—. Portaera datuak aldagai soziodemografiko, demolinguistiko eta ekipamenduzkoekin alderatzen dira.

Inkesta bakoitzaren ondoren txosten bat argitaratzen da proiektuaren webgunean webgunean eskuragarri dagoelarik (UPV/EHU et al., 2023).

Aurrekoak bezala, jasotako datuak aitortuak dira, eta panel

en este último caso, a partir del curso 2022-2023, la encuesta se ha extendido a centros del País Vasco continental.

Para elegir la muestra, se utiliza un muestreo aleatorio estratificado en el que las variables de estratificación son el sexo, la edad (en grupos de dos años) y la provincia. Una vez realizadas las encuestas, éstas se enlazan mediante un código personal del que dispone cada panelista, con el fin de conseguir una base de datos con la información de todas las preguntas para todos los registros. La información recopilada proporciona información sobre los consumos audiovisuales de la población joven a través, entre otros, del consumo recordado: contenidos (ficción, información, etc.), soporte (*streaming*, VOD, televisión lineal...), idioma (tanto en audio como en subtítulos), además del cuándo y dónde, con quién, etc. Además, las encuestas abordan esporádicamente temas monográficos —redes sociales (Instagram, Twitch...), impacto del COVID-19, importancia de la información para la gente joven, etc. —. Los datos de prácticas comunicativas se comparan con las variables sociodemográficas, demolingüísticas y de equipamiento.

Tras cada encuesta se publica un informe disponible en la página web del proyecto (UPV/EHU et al., 2023).

Al igual que los anteriores, los datos recogidos son declarados y

bidez ateratakoak. Halaber, honakoa ere multimedia panela da.

2.3.4. Hekimen (Hekimen Analytics)

Herri ekimeneko euskarazko komunikabideen elkartea da Hekimen. 2016an, Euskal Hedabideen Behategia proiektuaren baitan, Hekimen Analytics tresna abian jarri zuen. Tresna honek 51 euskarazko komunikabide digitalen trafikoa neurtzen du (webguneak zein Facebook eta Twitter kontuak), Google Analytics-en datuak baliatuta. Tresnaren ezaugarri nagusia komunikabideen artean datuak partekatzea da. Izan ere, analitika digitalean komunikabide bakoitzak bere datuak jaso ohi ditu, baina ez dira partekatzen, lehia arrazoiengatik. Hekimen Analytics-ek ordea sektorea du gogoan, ez komunikabide bakoitza, lankidetza lehia baino onuragarriagoa delakoan.

2013tik aurrerako datuak jasotzen dira, Google Analytics-ek eskaintzen dituen metrika nagusiekin (*The Ultimate Google Analytics Glossary*, 2022).

Une honetan Behategiak BEHA Analytics proiektua du abian, zeinaren bitartez *Google Analytics 4* tresna berria inplementatuko baita euskarazko komunikabide digital gehienetan, ia 70 medio barneratuz; tartean EITB erakunde publikoa. Tresna berri honek aldaketa

extraídos mediante paneles. Es un panel multimedia.

2.3.4. Hekimen (Hekimen Analytics)

Hekimen es la asociación de medios de comunicación de iniciativa popular en euskera. En 2016, y dentro del marco del Observatorio de Medios en Euskera Behategia, puso en marcha la herramienta Hekimen Analytics. Esta herramienta mide el tráfico de 51 medios digitales en euskera (webs y cuentas de Facebook y Twitter) a través de los datos de Google Analytics. Su principal característica es el uso compartido de datos entre los medios de comunicación. Hay que tener en cuenta que en la analítica digital cada medio recibe sus propios datos, pero no suele ser habitual que estos sean compartidos, debido a razones de competencia. Hekimen Analytics, sin embargo, está construido desde el punto de vista del sector, no de cada medio de comunicación, desde la premisa de que la cooperación es más beneficiosa que la competencia.

Hekimen Analytics recopila datos a partir de 2013, con las principales métricas que ofrece Google Analytics (*The Ultimate Google Analytics Glossary*, 2022).

Actualmente Behategia tiene en marcha el proyecto BEHA Analytics, por el que se implementa la nueva herramienta *Google Analytics 4* en la mayoría de los

handia dakar orain arteko neurketa digitalen aldean, besteak beste cookieak desagertuko direla aurreratzen delarik. Horrekin batera gutxieneko datu soziodemografikoak eskainiko ditu (sexua eta adina), analisiari ate berriak irekita.

Orain arte deskribatutako sistemekiko ezberdina da: jasotzen diren datuak errolda bidezkoak dira (trafiko osoa erregistratzen da, ez lagin bat); behatuak (ez erabiltzaileek aitortuak); monomedia dela esan genezake (trafiko digitala soilik neurtzen du); eta *site-centric* (aurrekoak ez bezala, *user-centric* izan baitira horiek).

2.3.5. Kantar Mediaren panela

1993tik aurrera hasi ziren telebista audientziak audimetriaren bidez neurtzen Espainiako Estatuan. Funtsean, etxebizitzen panel batean jarritako audimetroek egiten duten erregistroan oinarritzen da audimetria. Etxeko lagun bakoitzak bere botoia du audimetroan, eta telebista aurrean jartzen delarik botoia klikatzen du. Horren bidez, audimetroak segunduero jasotzen du telebista gailua erreproduzitzen ari den edukiaren berri, une horretan telebista ikusten ari direnen informazioarekin batera. Horrela osatzen da telebisten

medios digitales en euskera, alcanzando un número cercano a los 70 e incluyendo al medio público EITB. Esta nueva herramienta supone un cambio sustancial respecto a las mediciones digitales realizadas hasta la fecha, adelantándose, entre otras cosas, la desaparición de las cookies. Entre los datos que permite recopilar esta nueva herramienta se encuentran datos sociodemográficos básicos (sexo y edad), lo que abre nuevas perspectivas para el análisis.

El tipo de datos que se recoge difiere de los expuestos hasta ahora: son censales (se registra todo el tráfico, no una muestra); observados (no declarados); podríamos calificarlos de monomedia (solo se registra el tráfico digital); y *site-centric* (al contrario de los anteriores, *user-centric*).

2.3.5. Panel de Kantar Media

Las audiencias televisivas en el Estado español comenzaron a medirse mediante técnicas de audimetría a partir del año 1993. Básicamente, la audimetría se basa en el registro que realizan los audímetros instalados en un panel de viviendas. Cada una de las personas presentes en el hogar es identificada mediante un botón en el audímetro y pulsa el botón cuando está consumiendo televisión. Así, el audímetro registra cada segundo el contenido que se está reproduciendo, junto con la información sobre las personas que se encuentran viendo la tele-

audientzia neurketa erreferentziala. Gaur egun Espainiako Estatuan lan hau egiten duen enpresa Kantar Media da.

Euskal Autonomia Erkidegoan 430 etxebizitzetan dago audimetroa, horien bitartez 994 pertsonaren jarduerak neurtzen direlarik; lagin honek EAE unitate analisi gisa hartzea ahalbidetzen du. Nafarroan, aldiz, 90 bat neurgailu daude, orotara 200 bat pertsonaren datuak biltzen; datu-maila baxua dela eta herrialde hau ez da analisi unitatetzat hartzen.

Jasotzen diren datuak mistoak direla esan daiteke: aitortuak, batetik, ikusleak telebistaren aurrean jarri dela aitortu behar duelako; eta errolda bidezkoak, bestetik, behin presentzia adierazita audimetroak bere jokaera neurtzen baitu. Monomedia panela da, telebistara mugatua.

2.3.6. Bestelako sistemak

Aipatutako iturriez gain, beste zenbait eduki behar dira gogoan: OJD eta OJD Interactiva, prentsaren hedapena neurtzen dutenak —ez dira kontuan hartu euskarazko komunikabiderik apenas neurtzen dutelako—; Comscore eta GfK enpresek trafiko digitalaz egiten dituzten neurketak, user-centric panelak zein site-centric neurketak —

visión en ese momento. De esta manera se consigue lo que se ha convertido en la medición de referencia de la audiencia de la televisión. Kantar Media es la empresa que actualmente realiza este trabajo en el Estado español.

En la Comunidad Autónoma del País Vasco se utilizan 430 audímetros en otras tantas viviendas, lo que permite configurar un panel de 994 personas. Gracias a esta muestra la CAPV es tratada como una unidad analítica. Por el contrario, Nafarroa cuenta con unos 90 audímetros, con un alcance de cerca de 200 personas; pero este territorio no es considerado una unidad de análisis.

Podemos decir que los datos que se recogen son mixtos, ya que por una parte la persona informante ha de declarar que está viendo la televisión; pero por otra parte su comportamiento es observado a partir de ese punto. Se trata asimismo de un panel monomedia y actualmente limitado a la televisión.

2.3.6. Otros sistemas

Además de las fuentes citadas, hay que considerar otras como la Oficina de Justificación de la Difusión OJD y OJD Interactiva, que miden la difusión de la prensa basándose en la tirada y en la distribución. Sin embargo apenas miden medios de comunicación en euskera. Están también las mediciones que Comscore y GfK realizan en el tráfico digital (pa-

oraingoz Euskal Herria unitate gisa hartzen ez dutenak—; Frantziako Estatuko neurketak (Médiamétrie); eta abar.

Publizitatearen alorreko neurketak ere sar litezke multzo honetan (ARCE, Infoadex eta IAB esaterako).

Azkenik, era ezberdinetako estatistika ofizialak —alde batetık geografikoak (EUSTAT, NASTAT, INE, INSEE, Soziometroa) eta bestetik tematikoak (Kulturaren Euskal Behatokia kasu)—. aipa ditzakegu, etorkizunean datu hornitzaile gisa har litezkeenak.

neles user-centric y mediciones site-centric) que de momento no consideran al País Vasco como unidad. Finalmente, no debemos olvidar las mediciones realizadas en el Estado francés (principalmente Médiamétrie).

También podrían incluirse aquí las mediciones de inserciones publicitarias (ARCE, Infoadex e IAB).

Por último, podemos citar las estadísticas y encuestas oficiales de diversa índole —algunas de base geográfica (EUSTAT, NASTAT, INE, INSEE, Sociómetro), y otras de base temática (Observatorio Vasco de la Cultura) que podrían considerarse en el futuro como fuentes proveedoras de datos que incluyen alguna información sobre prácticas comunicativas.

2.3.7. Laburpena

Jarraian datorren 1. taulak modu sinoptikoan laburbiltzen ditu gaur egun euskarazko komunikazio praktikei buruzko datu bilketa handienak. Tresna bakoitzaren ezaugarri nagusiez gain, eskaintzen duen informazioa ere zehazten da: datu soziodemografikoak eta demolinguistikoak batetik; eta komunikazio euskarri zein komunikazio praktika zehatzak bestetik.

Ikusten denez, ez dago ikuspegi orohartzailea ahalbidetzen duen datu bilketarako tresnarik. Bakoitzak bere ekarpena egiten du,

2.3.7. Resumen

La tabla 1 resume de forma sinóptica las principales recopilaciones actuales de datos sobre las prácticas de comunicación en euskera. Además de las características principales de cada herramienta, se detalla la información que ofrecen, por un lado, sobre datos sociodemográficos y demolingüísticos; y por otro, sobre soportes comunicativos específicos y sobre prácticas comunicativas concretas.

Como se observa, no existen instrumentos de recogida de datos que permitan una visión con-

eta informazio ugari eskaintzen du dagokion eremuan eta bere metodologia baliatuz arabera. Baina ezinezkoa da iturri bakarra hartuta edo sistema bakarrean geratuta jokoan dauden aldagai guztien arteko harremanak aztertzea, gaur-gaurkoz.

Hain zuzen ere informazioaren zatikatze hori gainditzea bilatzen du *Jose Ignacio Ruiz Olabuenaga II. Ikerketa Beka* deialdiari esker garatu den proiektu honek, aldi berean komunikabideen audientzia ikerkuntzarako metodologia berriak esploratzeko azterketa aplikatu bat proposatuz. Epe luzerako helburua, ahalik eta datu-base gehien fusionatzea izango litzateke, binaka edo multzo handiagoetan posible balitz. Horrek informazio aberatsagoa emango liguke; oraingoan ordea, bi datu-baseren fusioarekin hasiko gara: CIES eta IKUSIKER.

junta. Cada uno realiza su propia aportación y genera mucha información en su ámbito y con su metodología. Pero no es posible analizar las relaciones entre todas las variables en juego.

Este proyecto, desarrollado gracias a la *2ª Beca de Investigación José Ignacio Ruiz Olabuenaga* busca precisamente superar esta fragmentación de la información, al tiempo que se propone un análisis aplicado para explorar nuevas metodologías de investigación de audiencias. El objetivo a largo plazo sería fusionar el mayor número posible de bases de datos, si fuese posible por parejas o en grupos más numerosos. Esto nos daría una información más útil; en esta ocasión nos limitaremos a la fusión de dos fuentes: CIES e IKUSIKER.

1. taula: Araba, Bizkaia, Gipuzkoa eta Nafarroako komunikazio praktikak neurtzeko tresna nagusien laburpen sinoptikoa. / Tabla 1: Resumen sintético de los principales instrumentos de medición de las prácticas comunicativas en Araba, Bizkaia, Gipuzkoa y Nafarroa

Epe laburreko helburua/Objetivo a corto plazo: CIES + IKUSIKER	**Kantar Media**	**Hekimen + Beha Analytics**	**IKUSIKER**	**EGM**	**CIES**	
User-centric	User-centric	Site-centric	User-centric	User-centric	User-centric	**Izaera/Carácter**
4 herrialde/ territorios 14-19 urte/años	3 herrialde/ territorios +4 urte/años	7 herrialde/ territorios	4 herrialde/ territorios[2] 11-23 urte/ años	3 herrialde/ territorios +14 urte/ años	4 herrialde/ territorios +14 urte/ años	**Unibertsoa/ Universo**
Zehazteke/ Por definir	Panela/ Panel 994[3]	Errolda bidezkoa. Euskarazko komunikabide digital gehienak/Censal. Mayoría de medios digitales en euskera	Panela/ Panel 3.000	Lagina/ Muestral 6.732[1]	Lagina/ Muestral 8.600	**Adierazgarritasuna/ Representatividad**
Observado/ Behatua	Declarado/ Aitortua	Observado/Behatua	Declarado/ Aitortua	Declarado/ Aitortua	Declarado/ Aitortua	**Tipo de dato/ Datu mota**
						Soziodemografikoak/Sociodemográficas
						Demolinguistikoak/Demolingüísticas
						Prentsa/Prensa
						Irratia/ Radio
						Telebista/Televisi ón
						Internet
						Ikus-entzunezkoak/ Audiovisual
						Sare sozialak/ Redes sociales
						Bestelakoak/Otras

Berde ilunez: datu kopuru handia/Verde oscuro: datos abundantes - Berde argiz: datu kopuru mugatua/Verde claro: datos escasos - Grisez: daturik ez/Gris: ausencia de dato

1 EAE solik/Solo CAPV.

2 2022-2023 ikasturtean hasi dira datuak biltzen Lapurdin, Nafarroa Beherean eta Zuberoan. Beraz 2023tik aurrera 7 lurraldeetara zabalduko da unibertsoa/ <?> En el curso 2022-2023 se ha comenzado con la recopilación de datos en Lapurdi, Nafarroa Beherea y Zuberoa, con lo que a partir de 2023 el universo se extenderá a los 7 territorios.

3 EAE solik/Solo CAPV.

2.4. Datuen fusioa oinarri duen azterketarako planteamendua

Orain arte ikusi dugunaren arabera, honakoa esan dezakegu: komunikazioan gertatzen ari diren aldaketak orain arte baino modu zehatzagoan aztertzea ahalbidetuko duten datuak lortzeko zailtasunak daude Euskal Herrian. Informazio ugari dago komunikabideetako audientzien ikerketari ekiteko, baina informazio horrek mugak ditu, bai bere orientazioagatik bai eta ikuspegi global faltagatik ere. Hutsune hori berori ageri da euskarazko komunikazio praktikak aztertu nahi badira.

Hala ere, badira bideak dagoen datuen ugaritasunetik abiatuta aurrerapausoak emateko. Ikuspegi horretatik abiatuta burutu da Datu Integralak (D^{i}) izendatutako proiektua, zera bilatuz: iturri ezberdinetako datuak fusionatu (*statistical matching* metodologia matematikoaren bidez), inkesta ezberdinek ematen dituzten ikuspegiak bateratzeko eta komunikazio ohituretan gertatzen ari diren aldaketak hobeto ulertzeko tresna iraunkor bat eraikitzeko.

Aurreragoko ataletan luze eta zabal kontatuko bada ere, labur formulatuta honela aurkez daiteke datuen fusioa oinarri duen azterketarako planteamendu honen estrategia: audientzia ikerketan egiten diren neurketa batzuek datu ugari baina ez na-

2.4. Planteamiento para el análisis basado en la fusión de datos

Por lo visto hasta ahora, podemos decir, por un lado, que en el País Vasco existen dificultades para obtener datos que permitan analizar, de forma más precisa que hasta ahora, los cambios que se están produciendo en la comunicación. Hay mucha información procurada por la investigación de audiencias, pero tiene limitaciones por su orientación y por la falta de visión global. Esta misma laguna se advierte para el análisis de las prácticas comunicativas en euskera.

Sin embargo, existen vías para avanzar dada la abundancia de datos existentes. Desde esta perspectiva se ha desarrollado el proyecto denominado Datos Integrales (D^{i}) buscando la fusión (mediante la metodología matemática de *statistical matching*) de datos extraídos de diferentes fuentes para aunar los puntos de vista que ofrecen las diferentes encuestas y construir una herramienta permanente para comprender mejor los cambios que se están produciendo en los hábitos de comunicación.

Si bien más adelante se abordará en profundidad, podemos ahora formular brevemente la estrategia de este planteamiento de análisis basado en la fusión de datos: si algunas de las mediciones que se realizan en la investigación de audiencias proporcionan datos

hikoak ematen badituzte, horiek bestelako neurketekin fusionatzea, beharrezkoa dugun ikuspegi globalagoa lortzeko. Kasu honetan bi datu-base fusionatzea proposatzen dugu: CIES enpresak egiten duen *Estudio de Audiencia de Medios de Comunicación en la C. A. de Euskadi y Navarra* (CIES, 2021) batetik, eta EHUk beste eragile batzuekin lankidetzan egiten duen *IKUSIKER* panela bestetik (UPV/EHU et al., 2023). Lehenak lagin handiko eta datu ugariko informazioa ematen du; bigarrenak propio begiratzen die komunikazio aztura batzuei (zehazki gazteen sareko kontsumoari). Biak fusionatuz, orain arte ezezaguna zen informazio bat lortzea izango da xedea.

Bi datu-base horiek 2.3 atalean deskribatu badira ere, hona hemen datu-iturri horien zer urtetako inkestekin lan egin erabakitzeko jarraitutako prozeduraren xehetasunak.

abundantes pero insuficientes, se trata de fusionarlos con otras mediciones para lograr la visión más global que precisamos. En este caso se propone fusionar dos bases de datos: el Estudio de Audiencia de Medios de Comunicación en la C. A. de Euskadi y Navarra (CIES, 2021) realizado por la empresa CIES, y el panel IKUSIKER que la UPV/EHU elabora en colaboración con otros agentes (UPV/EHU et al., 2023). El primero aporta información de gran tamaño muestral y de gran profusión de datos, mientras que el segundo atiende de forma específica a ciertos hábitos de comunicación (concretamente al consumo en la red entre las personas jóvenes). Fusionando ambos, el objetivo será obtener una información hasta ahora desconocida.

A pesar de que ambas bases de datos se han descrito en el apartado 2.3, a continuación se detalla el procedimiento seguido para decidir el año de encuestación de estas fuentes de datos.

2. taula: CIES eta IKUSIKER inkestetako datu-base zehatzen hautaketa eta fusiorako prestaketa. / Tabla 2: Selección de bases de datos específicas de las encuestas de CIES y de IKUSIKER y preparación para la fusión.

CIES 2021	CIES 2021
Tradizio handiko datu bilketa bat da, ia lau hamarkadatan egonkortasuna eta berrikuntza mantentzen jakin duena. Edozein urtetako datu-basea hautatzeko aukera irekitzen du horrek, salbu 2022ko inkestak —proiektu hau hasi zenean orainidik egin gabe zeudelako—; hortaz, bigarren datu-basearen ezaugarriei erreparatu zaie hautaketarako irizpide bila. IKUSIKERren kasuan 2020-2021eko inkestak erabili behar izango dira, hortaz, CIESen 2021eko urte naturaleko datu metatuak erabiltzea erabaki zen.	Se trata de una recopilación de datos de gran tradición que ha sabido mantener la estabilidad y la innovación durante casi cuatro décadas. Esto abre la posibilidad de elegir una base de datos de cualquier año, excepto las encuestas de 2022, que a la fecha de inicio de este proyecto no estaban realizadas, por lo que se ha buscado un criterio de selección atendiendo a las características de la segunda base de datos. En el caso de IKUSIKER será necesario utilizar las encuestas correspondientes al año 2020-2021, por lo que en CIES se decidió utilizar los datos acumulados del año natural 2021.
IKUSIKER 2020-2021	**IKUSIKER 2020-2021**
IKUSIKER panelaren kasuan ordea, panel pilotua izanik, garrantzitsua da kontuan hartzea oraindik hiru arazo nagusiri aurre egin behar diola. 1. Panelkide guztiek ez dituzte inkesta guztiak erantzuten. 2. Panelkide batzuek kode pertsonala gaizki erabiltzen dute (kodea gaizki idatzita dago batzuetan, ez dira gogoratzen bestetan, eta abar). 3. Galdera batzuk ez dira modu berean egiten DBH eta Batxilergoko panelean eta unibertsitatekoan. Ondorioz, falta diren balioen kopurua handia da —bai inkesta bakoitzean, bai datu-base osoan—. IKUSIKERren datu-basea hautatzeko, lehenik eta behin, 2019tik aurrerako zein ikasturtetako datuak erabiliko ziren erabaki behar izan zen. Datuen fusiorako proiektua hasi zenean 2021-2022 ikasturteko inkestak oraindik amaitu gabe zeudenez, 2020-2021 ikasturteko datuak erabiltzea erabaki zen. Ondoren, inkesta horietako bakoitzeko erantzun kopurua aztertu zen, ahalik eta gehien sartzeko asmoz. Lau hauek bete zuten baldintza: 1. Galdera soziodemografikoekin kaptaziorako/berriro lotzeko inkesta. 2. Instagramen kontsumoari buruzko inkesta. 3. Serieen kontsumoari buruzko inkesta. 4. Youtube eta Twitchen kontsumoari buruzko inkesta.	En el caso del panel IKUSIKER, al ser un panel piloto, es importante tener en cuenta que todavía tiene que hacer frente a tres problemas principales. 1. No todas las personas panelistas responden a todas las encuestas. 2. Algunas personas panelistas no utilizan bien el código personal (código mal escrito, olvido, etc.). 3. Algunas preguntas no se hacen del mismo modo en el panel de ESO y Bachiller que en el de la universidad. En consecuencia, el número de valores que faltan es elevado, tanto en cada encuesta como en toda la base de datos. La selección de la base de datos de IKUSIKER se realizó, en primer lugar, sobre la base de datos de los cursos a partir de 2019. Dado que a la fecha de inicio del proyecto de fusión de datos no habían finalizado las encuestas correspondientes al curso 2021-2022, se decidió utilizar los datos correspondientes al curso 2020-2021. Posteriormente, se analizó el número de respuestas de cada una de las encuestas, con el fin de incluir el mayor número posible de respuestas. Estas cuatro cumplieron el requisito: 1. Encuesta de captación/relanzamiento con preguntas sociodemográficas 2. Encuesta de consumo de Instagram. 3. Encuesta de consumo de series. 4. Encuesta de consumo de Youtube y Twitch.

Inkesta horiek elkarrekin lotzeko, lehen aipatutako panelkideen kode pertsonalak erabili ziren. Kode horiek identifikatzeko arazoak izan ziren kasu batzuetan eta horiek paneletik kanpo utzi behar izan ziren. Amaitzeko, inkesta guztiak estekatu ondoren DBH eta Batxilergoko panelean eta unibertsitatekoan kategoria desberdinak zituzten aldagai batzuk birkategorizatu behar izan ziren.
Prestaketa lan horien ondoren, 60 bat galderako informazioarekin osatutako datu-basea sortu zen.

Para relacionar estas encuestas se utilizaron los códigos personales de las personas panelistas anteriormente mencionadas. La identificación de estos códigos planteó en algunos casos problemas que obligaron a dejar registros fuera del panel. Por último, una vez enlazadas todas las encuestas, hubo que recategorizar algunas variables con categorías diferentes en el panel de ESO y Bachiller y en el de la universidad.
Tras estos trabajos preparatorios se creó una base de datos con información de unas 60 preguntas.

Bi datu-baseak hautatuak eta lanerako prestatuak izan eta gero, datuen fusioari ekitea baino ez da falta; lehenik, ariketari dagozkion metodologia eta teknikak ezagutuz eta, ondoren, CIES eta IKUSIKERren arteko fusioa helburu duen azterketa aplikatua burutuz.

Una vez seleccionadas las dos bases de datos y preparadas para el trabajo, pudo acometerse la fusión de los datos, primero conociendo las metodologías y técnicas propias del ejercicio y, posteriormente, realizando un análisis aplicado que buscaba la fusión entre CIES y IKUSIKER.

IKUSIKER panelaren izaera pilotuaren ondoriozko mugak azpimarratu behar dira hemen. Egindako zortzi inkestetatik lautara mugatu behar izan da fusioa, eta hainbat erregistro alboratu behar izan dira, panelisten erantzunak behar bezala lotu ezin direlako. Horrek fusioaren datuen kalitatea murriztu du. Murrizketa hori, beraz, erabilitako lehengaitik dator, eta ez, geroago ikusiko dugunez, aplikatutako metodologiatik; eta lortutako datuetatik egin dezakegun analisi mugatuan islatuko da.

Se hace necesario incidir en las limitaciones derivadas del carácter piloto del panel IKUSIKER. Haber tenido que limitar la fusión a cuatro encuestas de las ocho realizadas, así como descartar registros por no poder enlazar adecuadamente las diferentes respuestas de las personas panelistas, ha traído consigo una merma en la calidad de los datos de la fusión. Esta merma se deriva pues de la materia prima utilizada y no, como veremos posteriormente, de la metodología aplicada; y se verá reflejada en el alcance, limitado, del análisis que podemos realizar de los datos obtenidos.

Nolanahi ere, prozesuak balio izan du, hala bateratuko diren baseen ezaugarriei buruzko ondorioak ateratzeko, nola haien oinarrian dauden inkesten alderdi batzuk aldatzeko. CIESen eta IKUSIKERen 2023ko edizioetan

En todo caso el proceso ha servido tanto para extraer conclusiones sobre las características de las bases que se vayan a fusionar, como para modificar algunos

jada aplikatzen ari diren aldaketak dira, lehenengoaren kasuan aldagai berri bat sartuta eta bigarrenaren kasuan laginketa eta galdetegiak moldaketak ezarrita.

aspectos de las encuestas de las que derivan aquellas, modificaciones que ya están siendo aplicadas en las ediciones tanto de CIES como de IKUSIKER en 2023, con la introducción de una nueva variable en la primera y la modificación en el muestreo y en el cuestionario en la segunda.

Bigarren Atala: datuen fusioa

3. *Statistical matching*: aurrekariak

3.1. Aurrekari globalak

Datu-fusioaren (*statistical matchig*) jatorria 1960ko hamarkadaren erdialdean kokatzen da, *1966 US Tax File* eta *1967 Survey of Economic Opportunities* fusionatu zirenean (Okner, 1972). Fusio hori egiteko arrazoia honakoa izan zen: AEBko diru-sarrera pertsonalaren estimazio totala lortzea erraza izan arren, ez zegoen estatistika ofizialik diru-sarrera horren tamainaren banaketari buruz, ezta ohiko ezaugarri demografikoen araberako sailkapen gurutzaturik ere. Ondoren, 70eko hamarkadaren hasieran, AEBko inkesta sozialetan fusio-teknika desberdinak aplikatu ziren (Ruggles, 1974), baina teknika horiek gogor kritikatu ziren frogatu edo justifikatu ezin ziren susmoetan oinarritzen zirela argudiatuta (Kadane, 2001; Rodgers, 1984). Orduz geroztik eta gaur egun arte ekarpen ugari egin dira inkestak lotzeko tekniken inguruan (Bárcena & Tusell, 1999; Bello, 1993; Dempster et al., 1977; Muñoz & Villagarcía, 1997; Nordbotten, 1996; Rius, 1994; Rius et al., 1996; Rubin, 1986) , eta beste hainbat ekarpen *statistical matching* teknikei dagokienean (De Waal, 2015; De

Capítulo segundo: fusión de datos

3. *Statistical matching*: antecedentes

3.1. Antecedentes globales

Los orígenes de la fusión de datos (*statistical matching*) se remontan a mediados de los años 60, cuando fueron fusionados el *1966 US Tax File* y el *1967 Survey of Economic Opportunities* (Okner, 1972). El motivo de llevar esta operación a cabo fue que a pesar de la facilidad con la que se podía obtener una estimación total del ingreso personal en EEUU, no existían estadísticas oficiales sobre la distribución del tamaño de dicho ingreso ni clasificaciones cruzadas del ingreso personal por características demográficas típicas. Posteriormente, a principios de los años 70 se aplicaron diferentes técnicas de fusión a encuestas sociales en EEUU (Ruggles, 1974), pero estas técnicas fueron duramente criticadas por basarse en suposiciones no justificadas y no comprobables (Kadane, 2001; Rodgers, 1984). Desde entonces y hasta día de hoy ha habido numerosas nuevas contribuciones en las técnicas de enlace de encuestas (Bárcena & Tusell, 1999; Bello, 1993; Dempster et al., 1977; Muñoz & Villagarcía, 1997; Nordbotten, 1996; Rius, 1994; Rius et al., 1996; Rubin, 1986). Y otras muchas contribuciones en

Waal et al., 2011; D'Orazio et al., 2006; EUROSTAT, 2013; Moriarity & Scheuren, 2001; Rässler, 2002).

Statistical matching teknikak erabilgarriak izan daitezkeen egoeren adibideak ondokoak dira: aldagai komunak dituzten eta gainjartzen ez diren bi inkesten arteko lotura, *Big Data* inkesten datuekin edo datu administratiboekin lotzea, eta egozpen-balioak bilatzea talde batzuentzat aldagai bat baino gehiago falta direnean. *Statistical matching* teknikak erabiltzen dituzten erakundeen artean, estatistikako institutu nazionalen presentzia nabarmendu behar da (ISTAT, EUROSTAT, Statistics Netherlands, INE, edo EUSTAT esaterako), informazio estatistiko egokia, fidagarria eta ez-kontraesankorra izateko duten beharrak bultzatzen ditu horretara.

Hona hemen erakunde horiek egindako fusioen adibide batzuk: datu ekonomikoak eta datu sozialak, etxeetakoak, biztanleria-erroldaren datuak, diru-sarreren eta osasunari buruzkoak, eta abar. Horrez gain, datu-fusioa beste hainbat arlotan erabili daiteke, hala nola honakoetan:

- Gizartea eta politika
- Hezkuntza-ikerketa
- Zerbitzu publikoak
- Komunikabideen ikerketa
- Osasuna

lo respectivo a las técnicas de *statistical matching* (De Waal, 2015; De Waal et al., 2011; D'Orazio et al., 2006; EUROSTAT, 2013; Moriarity & Scheuren, 2001; Rässler, 2002).

Ejemplos de situaciones en las que las técnicas de *statistical matching* pueden ser útiles son: enlace de dos encuestas que no se superponen pero que cuentan con variables de fondo comunes, el enlace de Big Data con datos de encuestas o datos administrativos, y la búsqueda de valores de imputación cuando faltan para ciertos grupos algunas variables. Entre las instituciones que hacen uso de técnicas de *statistical matching* con el fin de contar con información estadística oportuna, confiable y no contradictoria, cabe destacar la presencia de los institutos nacionales de estadística (ISTAT, EUROSTAT, Statistics Netherlands, INE, EUSTAT, etc.).

Algunos ejemplos de fusiones realizadas por estas instituciones son: datos económicos y datos sociales, datos de hogares, datos del censo de población, datos de ingresos, datos de salud, etc. Además de ello, la fusión de datos también tiene aplicaciones en diferentes áreas como pueden ser:

- Sociedad y política.
- Investigación educativa.
- Servicios públicos.
- Investigación de medios de comunicación.
- Sanidad.

3.2. Aurrekariak Euskal Herrian: EUSTAT

EUSTATek erregistroen fusioan eginiko ibilbideari 1997. urtean ekin zion, Biztanleriaren Erregistro Estatistikoari dagozkion lehen lanak hasi zirenean. Une horretan, iturri bakoitza REP populazio oinarriarekin automatikoki fusionatzeko lehen metodoak diseinatu ziren EUSTATen. Hasierako fase batean, erregistroak fusionatzeko teknika deterministak erabili ziren, identifikazio-aldagai komunak erabiliz, bai sarrera-fitxategietan bai erreferentzia-fitxategietan ere (izen-deiturak, NAN/AIZ, jaiotze-data, posta-helbidea, eta abar). Teknika horiek modulu batean ezarri ziren fitxategi lauak fusionatzeko.

Hurrengo urteetan fusiorako probabilitate-teknikak garatu ziren. Teknika horietan, fusio-aldagaiei buruzko edozein akordiok edo desadostasunek emandako informazio guztia hartzen zen kontuan, eta horietako bakoitzak ahalmen erabakigarria zuen fusioan. Gainera, informazioa manipulatzeko metodo eta prozedurak gehitu ziren, bere garaian fitxategi lauetarako garatutakoak egokitzeko.

2006. urtean, EUSTAT metodo berriak erabiltzen hasi zen biztanleriaren zentsu-estatistikak egiteko, eta lehen pausoa 2006ko Biztanleriaren eta Etxebizitzen

3.2. Antecedentes en el País Vasco: EUSTAT

La experiencia de EUSTAT en la fusión de registros comenzó en el año 1997, cuando se iniciaron los primeros trabajos correspondientes al Registro Estadístico de Población (REP). En aquel momento se diseñaron en EUSTAT los primeros métodos para fusionar cada una de las fuentes de forma automática con la base poblacional del REP. En una primera fase inicial se emplearon técnicas deterministas de fusión de registros utilizando variables identificativas comunes, tanto a los ficheros de entrada como al de referencia (nombre, apellidos, DNI/NIE, fecha de nacimiento, dirección postal, etc.). Estas técnicas se implementaron en un módulo para fusionar ficheros planos.

En los años siguientes se desarrollaron técnicas probabilísticas de fusión, en las que se tenía en cuenta toda la información proporcionada por cualquiera de los acuerdos o desacuerdos sobre las variables a fusionar, y cada una de ellas tenía un poder decisivo diferente en la fusión. Además, se añadieron métodos y procedimientos de manipulación de la información que permitían adaptar los tradicionales, ya desarrollados en su momento, para los ficheros planos.

En el año 2006 EUSTAT comenzó a utilizar nuevos métodos en

Estatistika (BEE06) egitea izan zen. Estatistika horretan, administrazio-erregistroetatik eta zenbait estatistika-produktutatik jasotako informazioa konbinatu zen.

2006ko Biztanleria eta Etxebizitzen Estatistikaren abiapuntua EUSTATek 1997an abian jarritako Biztanleriaren Erregistro Estatistikoa eta 2001eko Biztanleria eta Etxebizitzen Errolda izan ziren. Populazio Erregistrotik lortutako populazio oinarritik abiatuta, populazioaren zentsu ezaugarriak eguneratu ziren, zentsuen ohiko esparru tematikoak birsortzeko helburuarekin.

Euskararen inguruko datu-iturrien fusioa egiteko, ondorengo urratsetan oinarritutako prozedura metodologikoa erabili zen (Morán Alaez et al., 2008):

1. Fusionatu beharreko iturria definitzea.
2. Aldagai bakoitzari dagozkion pisuak eta iturri bakoitzaren ontasunak esleitzea.
3. Iturri edo sarrera bakoitzerako datu-sarrera zehaztea.
4. Iturri edo sarrera bakoitzerako fusio eremuak identifikatzea.
5. Iturri edo sarrera bakoitzerako fusio eremuak normalizatzea.
6. Fusio determinista nagusia gauzatzea (bakarra/anizkoitza).

la elaboración de estadísticas censales de la población, siendo el primer paso la realización de la Estadística de la Población y Viviendas de 2006 (EPV06). En dicha estadística se combinó información procedente de registros administrativos y de diversos productos estadísticos.

El punto de partida de la Estadística de Población y Viviendas de 2006 fue el Registro Estadístico de Población puesto en marcha por EUSTAT en el año 1997 y el Censo de Población y Viviendas de 2001. A partir de la base poblacional obtenida del Registro de Población se actualizaron las características censales de la población con el objetivo de reproducir los tradicionales ámbitos temáticos de los censos.

Para realizar la fusión de las fuentes relativas al euskera se siguió una metodología basada en los siguientes pasos (Morán Alaez et al., 2008):

1. Definición de la fuente a fusionar.
2. Asignación de los pesos a cada variable y las bondades para cada fuente.
3. Determinación de la entrada de datos para cada fuente o entrada.
4. Identificación de los campos de fusión para cada fuente o entrada.
5. Normalización de los campos de fusión para cada fuente o entrada.

7. Bigarren mailako fusio determinista gauzatzea (bakarra/anizkoitza).
8. Fusio probabilistikoa gauzatzea (bakarra).
9. Fusioan parte hartzen duten erregistroak berrikustea eta xehetasunez aztertzea eta, hala badagokio, errepikatzea.

Horren ondoren, 2007an, EUSTATek mintegi bat antolatu zuen, William E. Yancey-k gidatua. Bertan, erregistroen fusioarekin erlazionatutako teoria azaldu zen: definizioa, hainbat fusio-mota eta horiei dagozkien metodologiak, fusioa erabiltzeko testuinguruak eta fusioaren erabilera ezberdinak (Yancey, 2007).

Urte batzuk geroago, 2010ean, Laura Otero Francok egindako *Unitate ekonomikoen erregistro administratiboak batu probabilitate-teknikak erabilita* proiektua gauzatu zuen EUSTATek (Otero Franco, 2010). Proiektu horretan, Fellegik eta Sunterrek proposatutako probabilitate-bilketaren metodologia aztertu zen (Fellegi & Sunter, 1969), eta unitate ekonomikoetako erregistro administratiboak tratatzeko egokitu. Gainera, aplikazio bat programatu zen SAS pakete estatistikoan administrazio-unitate ekonomikoen bi erregistro automatikoki elkartzeko. Azkenik, 2014. an beste *statistical matching* proiektu bat ere burutu zen, *Enlace de encuestas: Cuestio-*

6. Ejecución de la fusión determinista principal (única/múltiple).
7. Ejecución de la fusión determinista secundaria (única/múltiple).
8. Ejecución de la fusión probabilística (única).
9. Revisión y análisis detallado de los registros que intervienen en la fusión y en su caso, repetición de la misma.

En 2007 EUSTAT organizó un Seminario presentado por William E. Yancey en el que se explicó la teoría relacionada con el fusión de registros: definición, diferentes tipos de fusión y sus correspondientes metodologías, contextos en los que utilizar la fusión y los diferentes usos de la misma (Yancey, 2007).

Años más tarde, en 2010, EUSTAT llevó a cabo el proyecto "*Unitate ekonomikoen erregistro administratiboak batu probabilitate-teknikak erabilita*" realizado por Laura Otero Franco (Otero Franco, 2010). En dicho proyecto se analizó la metodología de recogida probabilística propuesta por Fellegi y Sunter (Fellegi & Sunter, 1969) y se adaptó para el tratamiento de los registros administrativos de las unidades económicas. Además, se programó una aplicación en el paquete estadístico SAS para poder unir dos registros de unidades económicas administrativas automáticamente. Por último, en 2014 también se realizó otro proyecto

nes metodológicas y práctica con R-StatMatch izenekoa, Inés Garmendia Navarrok egina (Garmendia Navarro, 2014). Proiektu horretan, datuen fusio motak azaltzeaz gain, *Hot-Deck* metodoak aplikatu ziren.

3.3. Statistical matching audientzia ikerketan

Duela urte batzuetatik hona, audientziak neurtzeko agentziek iturri desberdinetako datuak biltzen dituzte, ikuspegi osoagoa lortzeko.

Une honetan, paradigma berri baten beharra du audientzien neurketak. Estatu espainiarreko AIMCk -zeregin honetan aritzen direnetatik Europako aintzindarietakoa- honako ezaugarriak atxikitzen dizkio neurketen paradigma berriari (Santiago & AIMC, 2017):

- Makinen zentsuetan oinarritzea, pertsonek osatuetan baino gehiago (aztarna digitala).
- Laginez eta panelez baliatzea, erabiltzaileen aldagai soziodemografikoak ezagutu ahal izateko.
- Zentsu eta lagin datuak integratzea.

Hala, audientzien neurketari dagokionez, hainbat aurrerapauso egin dira Espainiako estatuan. Hona hemen horietako batzuk:

de *statistical matching* titulado "*Enlace de encuestas: Cuestiones metodológicas y práctica con R-StatMatch*" realizado por Inés Garmendia Navarro (Garmendia Navarro, 2014). En dicho proyecto, además de explicar los diferentes tipos de fusiones de datos, se hizo uso de los métodos *Hot- Deck*.

3.3. Statistical matching en los estudios de audiencia

Desde hace unos años, las agencias de medición de audiencias recopilan datos de diferentes fuentes para obtener una visión más completa.

En estos momentos la medición de audiencias precisa de un nuevo paradigma según la AIMC, una de las pioneras en este terreno en Europa, que señala las siguientes premisas para este nuevo paradigma (Santiago & AIMC, 2017):

- Basado en "censos" de máquinas, no personas (huella digital).
- Necesidad de muestras/paneles de las que poder extraer las variables sociodemográficas necesarias.
- Técnicas de integración de datos censales y muestrales.

Así, en el terreno de la medición de audiencias en el Estado español, se han dado diferentes avances, alguno de los cuales son:

- En 2008 la AIMC anunció la fusión de las diferentes en-

- 2008an, EGMren txostenerako egiten dituzten inkesta ezberdinen arteko fusioa iragarri zuen AIMCk: multimedia eta monomedia inkestak fusionatuz, datu bakarra lortzea zen helburua (ODEC & QUINAO, 2009).
- 2013an *Proyecto multiverso, Big Media Data* proiektua aurkeztu zen, bertan EGM eta Comscoreren datuen arteko fusioa planteatuta: bi sistemen panelen datuak (*user-centric*) eta Comscorek neurtutako webguneen datuak (*site-centric*) elkartuz.
- 2015ean *SKO, Video Integration Model: Building the Factory* proiektua abiatu zuen Kantar Mediak, sektoreko beste eragile batzuekin batera (TNS, Comscore, GfK eta Nielsen). Proiektu honek ere errolda-datuak eta panelak fusionatzea izan zuen xede *(Beck, 2015).*
- 2016an *Extended TV y Total Vídeo: la nueva medición Cross Media* proiektua aurkeztu zuten Kantar Mediak eta Comscorek *(Nafría & Vásques, 2016)*. Proiektu honetan Kantar Mediaren telebista paneleko eta Comscoreren panelak zein sareko zentsu datuak fusionatu ziren.
- 2019an AIMC eta Comscorek lankidetza hitzarmena sinatu zuten multiplataforma neurketa egiteko (Unicuestas elaboradas para el informe EGM, con el objetivo de obtener un único dato a partir de las encuestas multimedia y monomedia (ODEC & QUINAO, 2009).
- En 2013 se presentó el *Proyecto multiverso, Big Media Data*, en el que se planteaba la fusión de datos entre EGM y Comscore, uniendo los datos de los paneles de ambos sistemas (*user-centric*) y *los* datos de las webs medidas por Comscore (*site-centric*).
- En 2015 Kantar Media puso en marcha el proyecto *SKO, Video Integration Model: Building the Factory* en colaboración con otros agentes del sector (TNS, Comscore, GfK y Nielsen). Este proyecto también buscaba fusionar datos censales y paneles (Beck, 2015).
- Kantar Media y Comscore presentaron en 2016 el proyecto *Extended TV y Total* Vídeo: la nueva medición Cross *Media* (Nafría & Vásques, 2016). En este proyecto se fusionaron los paneles del panel de televisión de Kantar Media y Comscore, así como los datos censales de la red.
- En 2019 AIMC y Comscore firmaron un acuerdo de colaboración para la medición multiplataforma (*Unified Digital Measurement -UDM*). De esta forma se fusionaban tanto los datos muestrales de una como los datos censales de la otra. Las caracte-

fied Digital Measurement - UDM). Horren ondorioz, bataren eta bestearen lagin zein zentsu datuak fusionatzen hasi ziren. Ezaugarri teknikoak AIMCk lehiaketara aurkeztutako pleguan eta Comscorek aurkeztutako proiektuan zehazten dira *(Comscore, 2019)*.

Ezagutza, beraz, badago; alabaina, orain arteko aurrerapenak publizitatearen merkatuaren ikuspegitik egin dira, eta horrek zenbait inplikazio ditu. Batetik, merkatu estatala fokatzen da, euskal errealitatea, askotan, lauso geratzen delarik —horrek ondorioak ditu hizkuntzaren aldagaiaren neurketan (zentsu eta panel gehienetan ez da kontuan hartzen); lurraldetasunean (Euskal Herria, komunikazio espazio bat izan arren ez da unitate analisi, eta sarritan *Zona Norte* delako espazioan urtuta geratzen da); edota bertako datuen bazterketan (CIESen datuak ez dira kontuan hartzen, euskal gizarteko komunikazio praktiken erregistro zehatzena izan arren)—.

4. Metodologia

4.1. Datu-fusioa: sarrera

Datu-fusio (*statistical matching*) terminoa, oro har, xede-populazio bera duten bi datu-iturri (gehienetan laginak) edo gehiago erabiltzen dituzten metodoak

rísticas técnicas se especifican en el pliego presentado por el AIMC a concurso y en el proyecto presentado por Comscore al mismo (Comscore, 2019).

El conocimiento pues ya existe y está disponible. Sin embargo, los avances realizados hasta el momento se han realizado desde el punto de vista del mercado de la publicidad, lo que se traduce en diversas implicaciones. Por un lado, el punto de vista es el del mercado estatal, en el que la realidad vasca queda a menudo difuminada. Esto tiene consecuencias en la medición de la variable lingüística (que no se tiene en cuenta en la mayoría de los censos y paneles); en la territorialidad (el País Vasco, a pesar de ser un espacio de comunicación, no es considerado una unidad de análisis, y muchas veces se queda diluido dentro del espacio de Zona Norte); o en la exclusión de los datos disponibles en nuestro entorno (los datos del CIES quedan fuera, a pesar de ser el registro más exhaustivo de las prácticas de comunicación en la sociedad vasca).

4. Metodología

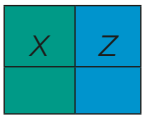
4.1. Fusión de datos: Introducción

El término fusión de datos (*statistical matching*), en general, se refiere a una serie de métodos que utilizan dos (o más) fuentes de datos disponibles (generalmente

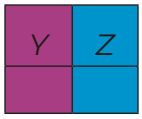
izendatzeko erabiltzen da; beti ere, helburua datu-iturri bakar batean batera behatzen ez diren aldagaien arteko erlazioa aztertzea izanik.

Datu-fusioaren oinarrizko egituran *A* eta *B* datu-baseak daude, *X* eta *Y* aldagai espezifikoekin, hurrenez hurren; eta ***Z*** aldagai komunen beste multzo bat, bi iturrietan beha daitekeena. Fusioaren helburua *X* eta *Y* aldagaien arteko erlazioa ikertzea da.

A datu-iturria:



B datu-iturria:



4.2. Datu-fusio motak

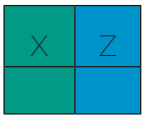
X eta *Y* aldagaien arteko erlazioa aztertzeko hainbat metodo daude. Horrenbestez, kasu bakoitzerako fusio-metodo egokiena aukeratzeko, kontuan hartu beharreko hiru ardatz azalduko dira jarraian banan-banan:

- Laginetan erabilitako inferentzia-mota.
- Fusioaren helburua.
- Erabili nahi den hurbilketa mota.

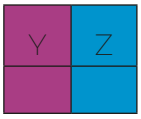
muestras), referidas a la misma población objetivo, con el fin de estudiar la relación entre las variables no conjuntamente observadas en una sola fuente de datos.

En la estructura básica de fusión de datos, se tienen dos bases de datos *A* y *B* con variables específicas *X* e *Y*, respectivamente, y otro conjunto de variables comunes ***Z*** observadas en ambas fuentes. El objetivo de la fusión consiste en investigar la relación entre las variables no conjuntamente observadas *X* e *Y*.

Fuente de datos A:



Fuente de datos B:



4.2. Tipos de fusión de datos

Para estudiar la relación entre las variables no conjuntamente observadas *X* e *Y* existen diferentes métodos. Para elegir el método de fusión más adecuado para cada caso hay tres cuestiones a tener en cuenta:

- El tipo de inferencia realizada sobre las muestras.
- El objetivo de la fusión.
- El tipo de aproximación que se quiere utilizar.

Laginetan erabilitako inferentzia-mota

Bi datu-base fusionatu nahi direnean, funtsezkoa da laginen gainean erabilitako inferentzia-mota kontuan hartzea. Fusio-metodo gehienek onartzen dute *A* eta *B* populazio infinitu berdinetik aukeratutako behaketa independente eta berdin-berdin banatuen zorizko laginak direla. Hau da, suposatzen dute *X*, *Y* eta ***Z*** zorizko aldagai talde baten emaitza independenteak direla, eta aldagai horien banaketa bateratuak eredu jakin bat jarraitzen duela (eredu batean oinarritutako inferentzia deritzo honi).

Praktikan, ordea, datu-base gehienak populazio finitu berean egindako laginketa konplexuetatik (estratifikatuak, klusterrekin osatuak, eta abar) ateratako inkestak dira. Kasu horietan, populazio finituko unitate bakoitzerako *X*, *Y* eta ***Z***-rako onartutako balioak balio finkotzat hartzen dira, eta ez zorizko aldagaien emaitzatzat. Arbitrariotasun-mekanismo bakarra populaziotik lagina aukeratzeko erabiltzen den probabilitate-irizpidea da. Hautapen-irizpideak laginean sartzeko π_a probabilitate ez-nulu bat esleitzen dio populazio-unitate bakoitzari ($0< \pi_a \leq 1$). Eta π_a probabilitate horiek inferentziaren oinarria dira, laginеko indibiduo bakoitzaren *pisua* izenarekin ezagutzen denarekiko zuzenki proportzionalak

Tipo de inferencia realizada sobre las muestras

Cuando se quieren fusionar dos bases de datos es primordial tener en cuenta el tipo de inferencia realizada sobre las muestras. La mayoría de métodos de fusión asumen que *A* y *B* son muestras aleatorias de observaciones independientes e idénticamente distribuidas seleccionadas de la misma población infinita. Es decir, suponen que *X*, *Y* y ***Z*** son resultados independientes de un grupo de variables aleatorias cuya distribución conjunta sigue un determinado modelo (denominada *inferencia basada en un modelo*).

La realidad, sin embargo, es que la mayoría de bases de datos son encuestas provenientes de muestreos complejos (estratificados, con clústeres, etc.) realizadas en la misma población finita. En este caso, los valores asumidos para *X*, *Y* y ***Z*** para cada unidad de la población finita son tomados como valores fijos y no resultados de variables aleatorias. El único mecanismo de arbitrariedad es el criterio de probabilidad utilizado para elegir la muestra de la población. El criterio de selección asigna a cada unidad de la población una probabilidad no- nula π_a ($0< \pi_a \leq 1$) de ser incluida en la muestra. Y esas probabilidades π_a son la base de la inferencia, pues son directamente proporcionales a lo que se conoce como el *peso* de cada individuo de la muestra

baitira (*d_a lagin-pisua)*. Hau da:

$$d_a = \frac{1}{\pi_a}$$

Inferentzia horri diseinu batean oinarritutako inferentzia deritzo (D'Orazio, 2013).

Fusioaren helburua

Fusioaren helburua erabakigarria da erabiliko den metodoa aukeratzeko orduan. Horren arabera, datuak fusionatzeko metodoak bi talde handitan banatzen dira: makro metodoak eta mikro metodoak.

Makro metodoak

Makro metodo deitzen direnen helburua aldagai espezifikoen intereseko parametroen estimazio zuzenak egitea da. Adibidez, kontingentzia-taulak edo *X* eta *Y* aldagaien arteko korrelazio-koefizienteak.

Mikro metodoak

Mikro metodo deitutakoen helburua berriz bestelakoa da: *X*, *Y* eta ***Z*** aldagai guztien informazioa izango duen datu-base sintetikoa sortzea hain zuzen ere.

Kasu honetan honelako itxura izango luke *A* eta *B* datu-iturrien arteko fusioaren emaitzak:

X	*Y*	***Z***

(*peso muestral d_a*). Es decir:

$$d_a = \frac{1}{\pi_a}$$

Esta inferencia es conocida *como inferencia basada en un diseño* (D'Orazio, 2013).

Objetivo de la fusión

El objetivo de la fusión es determinante a la hora de elegir el método que se va a utilizar. Dependiendo de éste, los métodos de fusión de datos se dividen en dos grandes grupos: métodos macro y métodos micro.

Métodos macro

El objetivo de los *métodos macro* es elaborar estimaciones directas de los parámetros de interés de las variables específicas. Por ejemplo, tablas de contingencia o coeficientes de correlación entre las variables *X* e *Y*.

Métodos micro

El objetivo de los *métodos micro* es crear una base de datos sintética que contenga información de todas las variables *X*, *Y* y ***Z***.

En este caso el resultado de la fusión entre *A* y *B* tendría el siguiente aspecto:

X	*Y*	***Z***

Datu-base sintetiko hori bi modutan lor daiteke. Lehena osatzen da *A* eta *B* kateatuz, eta, ondoren, falta diren balioak betez —hau da, *X B*-n eta *Y A*-n—. Bigarrena osatzeko modua datu-iturrietako batetik abiatzen da: datu-baseetako bat hartzen da (adibidez, *A*) eta falta den aldagaiaren informazioa egozten zaio (*Y*), beste datu-basean dagoen informazioa erabiliz (*B*); kasu honetan, *A* datu-base *hartzailea* dela esaten da eta *B* datu-base *emailea* —normalean, *emailea* erregistro gehien dituena izaten da, eta *hartzailea*, berriz, gutxien dituena—.

Esta base de datos sintética se puede obtener de dos maneras diferentes. La primera, concatenando *A* y *B* y posteriormente rellenando los valores faltantes, es decir, *X* en *B* e *Y* en *A*. Y la segunda, tomando una de las bases de datos (por ejemplo *A*) e imputándole la información de la variable faltante (*Y)* usando la información disponible en la otra base de datos (*B*). En este caso, se dice que *A* toma el papel de la base de datos *receptora* y *B* el de la base de datos *donante*. Generalmente la base de datos *donante* suele ser la que más registros tiene, y por lo contrario, la *receptora*, la que menos.

Erabili nahi den hurbilketa mota

Kontuan hartu beharreko azken puntua erabiliko den hurbilketa mota da. Bai *makro metodo*etan, baita *mikro metodo*etan ere, bi hurbilketa mota erabil daitezke: *hurbilketa parametrikoa* eta *hurbilketa ez-parametrikoa*. *Hurbilketa parametrikoak* berekin dakar eredu esplizitu bat hautatzea *X*, *Y* eta ***Z***-ren baterako banaketarako; eta, aldiz, *hurbilketa ez- parametrikoa* malguagoa da, eta ez dago eredu zehatz bat aurresuposatu beharrik. Gainera, *mikro metodo*etan metodo mistoak ere erabil daitezke, bi hurbilketa moten abantailak konbinatuz.

Jadanik aipatu den bezala, bai CIES eta bai IKUSIKER laginke-

Tipo de aproximación a utilizar

La última cuestión a tener en cuenta es el tipo de aproximación que se va a utilizar. Tanto en los *métodos macro* como en los *métodos micro* se pueden utilizar dos tipos de aproximaciones: *aproximación paramétrica* y *aproximación no-paramétrica*. La *aproximación paramétrica* implica la elección de un modelo explícito para la distribución conjunta de *X*, *Y* y ***Z*** y, por lo contrario, la *aproximación no-paramétrica* es más flexible, y no requiere presuponer ningún modelo concreto. Además, en los métodos *micro* también existe la posibilidad de utilizar métodos mixtos, los cuales combinan las ventajas de ambos tipos de aproximaciones.

ta estratifikatuetatik ateratako inkestak dira; beraz, lagin horietarako erabilitako inferentzia *diseinu batean oinarrituta* dago. Gainera, proiektu honen helburua aldagai guztien informazioa duen datu-base bat lortzea da, hau da, helburu *mikroa* du. Azkenik, aldagai kopurua handia dela kontuan hartuta, oso zaila da eredu estatistiko zehatz bat aurresatea, eta, horregatik, *hurbilketa ez-parametrikoa* erabiltzea erabaki da.

Hurrengo taulan datu-baseak fusionatzeko orduan kontuan hartu beharreko hiru puntuak laburtzen dira, eta laranja-kolorez markatzen dira CIES eta IKUSIKER datu-iturrien fusioari dagozkionak.

Como ya se ha mencionado, tanto CIES como IKUSIKER son encuestas provenientes de muestreos estratificados, por lo que la inferencia realizada para esas muestras está *basada en un diseño*. Además, el objetivo de este proyecto es conseguir una base de datos con la información de todas las variables, es decir, tiene un objetivo *micro*. Por último, teniendo en cuenta el número tan elevado de variables es muy complejo presuponer un modelo estadístico concreto, y por ello se ha optado por utilizar una *aproximación no–paramétrica*.

En la siguiente tabla se resumen las tres cuestiones a tener en cuenta a la hora de fusionar bases de datos, y se marcan en color naranja cuáles son las que corresponden a la fusión de CIES e IKUSIKER.

3. taula: Datu-baseak fusionatzerakoan kontuan hartu beharreko ezaugarriak eta gure fusioari dagokiona. / Tabla 3: Cuestiones a tener en cuenta al fusionar bases de datos y la correspondiente a nuestra fusión.

	Inferentzia / Inferencia	**Helburua / Objetivo**	**Hurbilketa / Aproximación**
	Eredu batean oinarritua Basada en modelo	Makro Macro	Parametrikoa Paramétrica
CIES + IKUSIKER	Diseinu batean oinarritua Basada en diseño	Mikro Micro	Ez-parametrikoa No-paramétrica

Horrenbestez, CIESen eta IKUSIKERren fusioaren ezaugarri zehatzak kontuan hartuta, fusioa egiteko lehendabiziko ideia *Hot-Deck* metodoak erabiltzea izan zen, *Naïve hurbilketa* aplikatuz; izan ere, metodo horiek

Considerando las características concretas de la fusión de CIES e IKUSIKER, la primera aproximación considerada para llevar a cabo la fusión fue utilizar métodos *Hot-Deck* con una *aproximación Naïve*, ya que estos métodos

ez-parametrikoak dira, helburu *mikroa* dute eta *diseinu batean oinarritutako inferentzian* zein *eredu batean oinarritutako inferentzian* erabil daitezke. Segidan, metodo horiek eta horien ezaugarrien azalpena.

4.2.1. Hot-Deck metodoak

Hot-Deck metodoak inkesten loturarako metodo erabilienak izan dira (Garmendia Navarro, 2014). Metodo hauek *ez-parametrikoak* dira, eta helburu *mikroa* dute. Fusioa gauzatzeko, datu-base *emaile* bat eta *hartzaile* bat ezartzeko prozedura erabiltzen dute, eta falta diren aldagaiak datu-base *hartzaile*an egozten dira, datu-base *emaile*aren informazioa erabiliz. Beraz, hauen ezaugarri nagusi kontsideratzen da egotzitako balioak beti direla errealak.

Metodo honen barruan hiru mota daude, eta bakoitzak metodologia bat du *emaile*en eta *hartzaile*en arteko erlazioak esleitzeko.

Random Hot-Deck

Metodo honetan indibiduoak estratu homogeneotan (dohaintza-klaseak) taldekatzen dira aldagai kategoriko komun baten edo gehiagoren arabera —hala nola sexua, adina, eskualdea, eta abar—. Dohaintza-klase horien barruan, datu-base *hartzaile*aren erregistro bakoitzerako datu-ba-

son *no-paramétricos* con un objetivo *micro*, y pueden utilizarse para muestras obtenidas tanto de *inferencia basada en un diseño* como de *inferencia basada en un modelo*. A continuación, se explica en detalle los métodos y sus características.

4.2.1. Métodos Hot- Deck

Los métodos *Hot-Deck* han sido los más utilizados para el enlace de encuestas (Garmendia Navarro, 2014). Estos son métodos *no – paramétricos* con un objetivo *micro* y utilizan el procedimiento de establecer una base de datos *donante* y una *receptora*, e imputar en la base de datos *receptora* las variables faltantes utilizando la información de la base de datos *donante*. Su característica principal es, por lo tanto, que los valores imputados siempre son reales.

Dentro de estos métodos hay tres tipos y cada uno de ellos sigue una metodología diferente para asignar las relaciones entre los *donantes* y los *receptores*.

Random Hot-Deck

En este método los individuos se agrupan en estratos homogéneos (clases de donación) dependiendo de una o más (las elegidas) variables categóricas comunes, como pueden ser sexo, edad, región, etc. Dentro de esas clases de donación, para cada registro de la base de datos

se *emaile*aren erregistro bat aukeratzen da ausaz eta horren informazioa egozten zaio. *Emaile* bakoitza behin baino gehiagotan aukeratu daiteke.

Rank Hot-Deck

Metodo honetan datu-base *hartzaile*aren erregistro bakoitzerako, datu-base *emaile*aren erregistro *hurbileneko* informazioa egozten da, aldagai jarraitu komun bakarraren banaketa metagarriaren funtzioaren ehunekoen arabera kalkulatutako distantzia baten arabera.

Distance Hot-Deck

Metodo honetan indibiduoak dohaintza-klaseetan taldekatzen dira, *Random Hot-Deck* metodoan bezalaxe. Ondoren, datu-base *hartzaile*aren erregistro bakoitzerako, datu-base *emaile*aren erregistro *hurbileneko* informazioa egozten da, aldagai kategoriko komunen azpimultzo baten bidez kalkulatutako distantzia baten arabera.

4.2.2. Naïve hurbilketa duten Hot-Deck metodoak

Laginketa konplexuetako inkesten datuak bateratzeko *Naïve hurbilketa*ren oinarria *Hot-Deck metodo mikro ez-parametrikoak* aplikatzea da. Inferentzia lagin *hartzaile*aren gainean egiten da, haren unitateei lotutako *lagin-pisuak* kontuan hartuta. *Pisuak* lotu-

receptora, se elige al azar y se imputa un registro de la base de datos *donante*, pudiéndose seleccionar cada *donante* más de una vez.

Rank Hot-Deck

En este método, para cada registro de la base de datos *receptora*, se imputa la información del registro "más cercano" de la base de datos *donante*, de acuerdo a una distancia calculada sobre los puntos porcentuales de la función de distribución acumulativa de la única variable continua común.

Distance Hot-Deck

En este método los individuos se agrupan en clases de donación, de la misma manera que en *Random Hot-Deck*. Después, para cada registro de la base de datos *receptora*, se imputa la información del registro "más cercano" de la base de datos *donante*, de acuerdo a una distancia calculada con un subconjunto (el seleccionado) de variables categóricas comunes.

4.2.2. Métodos Hot-Deck con *aproximación Naïve*

La *aproximación Naïve* para la fusión de datos de encuestas provenientes de muestreos complejos tiene como base aplicar los *métodos micro no–paramétricos Hot-Deck*. La inferencia se lleva a cabo sobre la muestra *receptora* considerando los *pesos muestrales* asociados a las unidades de la misma.

ran erabil daitezke edo ez. Bereziki:

Weighted Random Hot-Deck

Dohaintza-klaseen barruan, *B* datu-baseko *emaileak* aukeratzeko, *lagin-pisuekiko* proportzionala den probabilitatea erabil daiteke.

Weighted Rank Hot-Deck

A eta *B* unitateen *pisuak* erabil daitezke aldagai jarraitu komun bakarraren banaketa metagarriaren funtzioaren ehuneko puntuak estimatzeko.

Distance Hot-Deck-en kasuan, ezin dira erabili *pisuak* loturan, eta, beraz, *Weighted Distance Hot-Deck* ez da existitzen. Hala ere, laginaren diseinu-aldagaien arabera sor daitezke dohaintza-klaseak.

Los *pesos* pueden ser utilizados o no en la unión. En particular:

Weighted Random Hot-Deck

Dentro de las clases de donación, los *donantes* de la base de datos *B* pueden ser seleccionados con una probabilidad proporcional a los *pesos muestrales*.

Weighted Rank Hot-Deck

Se pueden utilizar los *pesos* de las unidades de *A* y *B* a la hora de estimar los puntos porcentuales de la función de distribución acumulativa de la única variable continua común.

En el caso de *Distance Hot-Deck*, no se pueden utilizar los *pesos* en la unión como tal, por lo que no existe *Weighted Distance Hot-Deck*. Sin embargo, es posible crear las clases de donación en función de las variables de diseño de la muestra.

En la siguiente tabla se resumen los *métodos Hot-Deck con aproximación Naïve*:

4. taula: *Naïve hurbilketa* duten *Hot-Deck* metodoen ezaugarrien laburpena. / Tabla 4: Resumen de las características de los métodos *Hot-Deck* con aproximación *Naïve*.

Metodoa / Método	Random Hot-Deck	Rank Hot-Deck	Distance Hot-Deck
Estratifikazio komuna Estratificación común	✓	x	✓
Aldagai komunak Variables comunes	Kategorikoak Categóricas	Jarraitua (bat) Continua (una)	Kategorikoak Categóricas
Distantzien kalkulua Cálculo de distancias	x	x	✓
Erregistroen *pisuen* erabilera Uso de los *pesos* de los registros	✓	✓	x

Lehenago aipatu den bezala, hasieran metodo horietako bat erabiltzea pentsatu zen, baina bakar batek ere ez zituenez CIES eta IKUSIKER datu-iturrien fusioan beharrezkoak ziren ezaugarri guztiak betetzen, azkenean proposamen metodologiko berri bat garatzea erabaki zen —metodo bakoitzaren abantailak konbinatuz, ahalik eta metodo osatuena lortzeko—.

4.3. Proposamen metodologikoa

Proposamen metodologiko berriak bost urrats ditu:

1. urratsa: datu-baseen arteko estratifikazio komuna identifikatu.

2. urratsa: aldagai komunak aurkitu, harmonizatu eta birsailkatu.

3. urratsa: indibiduoen arteko distantziak kalkulatu.

4. urratsa: emaile-hartzaile erlazioak ezarri, distantziak eta *pisuak* kontuan hartuta.

5. urratsa: emaile-hartzaile erlazioak erabili fusionatutako datu-basea lortzeko.

Lehenengo urratsa datu-baseen arteko estratifikazio komuna identifikatzea da. Horretarako, beharrezkoa da jakitea zer aldagaitatik abiatuta estratifikatu zen populazioa. Aldagai horiek kasu bakoitzean zeintzuk di-

Como se ha mencionado anteriormente, en un principio se pensó utilizar alguno de estos métodos, pero dado que ninguno resolvía todas las cuestiones necesarias en la fusión de CIES e IKUSIKER se decidió finalmente desarrollar una nueva propuesta metodológica, combinando las ventajas de cada uno de ellos para conseguir un método lo más completo posible.

4.3. Propuesta metodológica

La nueva propuesta metodológica, por lo tanto, consta de los siguientes 5 pasos:

Paso 1: Identificar la estratificación común entre las bases de datos.

Paso 2: Encontrar, armonizar y recategorizar las variables comunes.

Paso 3: Calcular las distancias entre los individuos.

Paso 4: Establecer las relaciones *donante-receptor* teniendo en cuenta las distancias y los *pesos*.

Paso 5: Utilizar las relaciones *donante-receptor* para conseguir la base de datos de la fusión.

El primer paso es identificar la estratificación común entre las bases de datos. Para ello es necesario saber a partir de qué variables se estratificó la población. Una vez conocidas dichas variables

ren ezagutu ondoren, horietako zeintzuk diren berdinak begiratu behar da. Behin identifikatuta, aldagai horien arabera estratu desberdinak sortuko dira (dohaintza-klaseak), informazioa estratu bereko erregistroen artean bakarrik transferitu ahal izateko.

Dohaintza-klaseak sortu ondoren, datu-baseek zer beste aldagai komun dituzten ikusi behar da. Kasu batzuetan, aldagaiak ez dira berdin-berdinak izango edo ez dituzte kategoria berak izango; beraz, kasu horietan, aldagai horiek harmonizatu eta birsailkatu egin beharko dira, bi datu-baseetan bat etor daitezen.

Aldagaiak identifikatu eta eraldatu ondoren (beharrezkoa izanez gero), loturarako zeintzuk erabiliko diren erabaki behar da. Hautatuak *X* e ta *Y* aldagai espezifikoak hoberen adierazten dituztenak izan behar dira. Gainera, egiaztatu behar da aldagai horien banaketa marjinalak bat datozela bi datu-baseetan; horrek esan nahi du aldagaiak populazioan modu berean banatuta egon behar direla. Banaketa marjinalak bat datozen edo ez ikusteko, aldagai kategorikoen kasuan, antzekotasun-ezaren indizea (*tvd*), gainjartze-indizea (*overlap*), Bhattacharyya koefizientea (*bhatt*) eta Hellingerren distantzia (*hell*) erabiltzen dira. Aipatutako indize guztiek 0 eta 1 arteko balioak hartzen dituzte, eta ez dago balio zehatzik al-

en cada caso, hay que mirar cuáles de ellas son iguales. Una vez identificadas, dependiendo de estas variables se crearán diferentes estratos (clases de donación), para así poder transferir información solamente entre registros del mismo estrato.

Una vez creadas las clases de donación, hay que ver de qué otras variables comunes disponen las bases de datos. En algunos casos, las variables no serán exactamente iguales o no tendrán las mismas categorías. Por lo tanto, habrá que armonizar y recategorizar esas variables para que sean coincidentes.

Tras identificar y transformar las variables (en caso de ser necesario), hay que decidir cuáles se van a utilizar para la unión. Las elegidas deben ser las que mejor representen a las variables específicas *X* e *Y*. Además, hay que comprobar que las distribuciones marginales de estas variables sean concordantes en ambas bases de datos. Esto quiere decir que las variables deben estar igualmente distribuidas en la población. Para mirar hasta qué punto son concordantes las distribuciones marginales, en el caso de las variables categóricas se utilizan el índice de disimilitud (*tvd*), el índice de superposición (*overlap*), el coeficiente de Bhattacharyya (*bhatt*) y la distancia de Hellinger (*hell*). Todos los índices mencionados anteriormente toman valores entre 0 y 1 y si bien

dagaien arteko komunztadura onartzeko; hala ere, Agrestiren arauari jarraituz, bi banaketak bat datozela onartzen da baldin eta *tvd* ≤0.06 bada, eta, beraz, *overlap* ≥0.94 bada (*tvd* = 1-*overlap* baita). Gainera, bi banaketa gertu daudela esaten da baldin eta *hell* ≤0.05 bada (D'Orazio, 2013).

Bestalde, analisi grafikoa ere egin daiteke, barra-grafikoak, sektoreak, eta abar erabiliz. Aldiz, aldagaiak jarraituak badira, erreferentzia estatistiko deskribatzaileak erabiltzen dira, hala nola, minimoa, maximoa, mediana, desbideratze estandarra eta abar. Analisi grafikoa ere egin daiteke, qqplots, histogramen eta abarren bidez.

Lotura egiteko erabiliko diren aldagai komunak erabaki eta gero, aldagai horiek baliatuta, datu-base *emaile*aren erregistroen eta estratu bereko datu-base *hartzaile*aren erregistroen arteko distantziak kalkulatu behar dira. Horretarako, *Gower*ren distantzia erabili da, distantzia horrek aldagai kuantitatiboak eta aldagai kualitatiboak kontuan hartzeko aukera ematen baitu (Anand, 2020). Oro har, honela definitzen da *Gower*ren distantzia, x_1 eta x_2 indibiduoetarako *p* aldagaien gainean kalkulatuta:

$$D_{Gower}(x_1, x_2) = 1 - \frac{1}{p}\sum_{j=1}^{p} s_j(x_1, x_2)$$

es cierto que no hay unos valores concretos para aceptar la concordancia entre variables, siguiendo la regla de Agresti se consideran concordantes si *tvd* ≤0.06 y por lo tanto si *overlap* ≥0.94 (ya que *tvd* = 1- *overlap*). Además, se considera que dos distribuciones están cerca si *hell* ≤0.05 (D'Orazio, 2013).

Por otro lado, también es posible hacer análisis gráfico mediante gráficos de barras, sectores, etc. Por el contrario, si las variables son continuas se utilizan estadísticos descriptivos como el mínimo, el máximo, la mediana, la desviación estándar, etc. También es posible hacer análisis gráfico mediante qqplots, histogramas, etc.

Una vez decididas las variables comunes que se van a utilizar para la unión, utilizando esas variables se procede a calcular las distancias entre los registros de la base de datos *donante* y la base de datos *receptora* del mismo estrato. Para ello, se ha utilizado la distancia de *Gower*, ya que esta distancia permite tener en cuenta tanto variables cuantitativas como variables cualitativas. En general, la distancia de *Gower* calculada sobre *p* variables para dos individuos x_1 e x_2 se define de la siguiente manera (Anand, 2020):

$$D_{Gower}(x_1, x_2) = 1 - \frac{1}{p}\sum_{j=1}^{p} s_j(x_1, x_2)$$

non $s_j(x_1,x_2)$ x_1 eta x_2 indibiduoek j aldagaian duten antzekotasuna adierazten duen, $j \in \{1,....p\}$ izanik. x_1 eta x_2 indibiduoek j aldagaian duten antzekotasuna modu honetan kalkulatzen da:

donde $s_j(x_1,x_2)$ indica la similitud entre los individuos x_1 y x_2 en la variable j y $j \in \{1,....p\}$. La similitud entre los individuos x_1 y x_2 en la variable j se calcula de la siguiente manera:

- Aldagaia kuantitatiboa bada:

$$s_j(x_1,x_2) = 1 - \frac{|y_{1j} - y_{2j}|}{R_j}$$

- Si la variable es cuantitativa:

$$s_j(x_1,x_2) = 1 - \frac{|y_{1j} - y_{2j}|}{R_j}$$

- Aldagaia kualitatiboa bada:

$$\begin{cases} y_{1j} = y_{2j} bada \rightarrow s_j(x_1,x_2) = 1 \\ y_{1j} \neq y_{2j} bada \rightarrow s_j(x_1,x_2) = 0 \end{cases}$$

- Si la variable es cualitativa:

$$\begin{cases} Si\ y_{1j} = y_{2j} \rightarrow s_j(x_1,x_2) = 1 \\ Si\ y_{1j} \neq y_{2j} \rightarrow s_j(x_1,x_2) = 0 \end{cases}$$

non y_{1j} eta y_{2j}, x_1 eta x_2 indibiduoek j aldagaian hartzen dituzten balioak diren, hurrenez hurren, eta R_j j aldagaiaren balioek hartzen duten tartea den.

donde y_{1j} y y_{2j} son los valores que toman los individuos x_1 y x_2 en la variable j, respectivamente, y R_j es el rango de la variable j.

Distantziak kalkulatu ondoren, estratu bakoitzerako distantzia-matrizea sortzen da, honako itxura duena:

Después de calcular las distancias, se crea la matriz de distancias para cada estrato, la cual tiene la siguiente forma:

Emaileak/Hartzaileak / *Donantes/receptores*	1 (wr_1)	...	q (wr_q)
1(wd_1)	d_{11}	...	d_{1q}
2(wd_2)	d_{21}	...	d_{2q}
...	...	...	...
p(wd_p)	d_{p1}	...	d_{pq}

Lehenengo zutabean *emaileak* zerrendatuta ageri dira, bakoitza bere wd_i *pisua*rekin eta lehenengo errenkadan *hartzaileak* zerrendatuta ageri dira, bakoi-

En la primera columna aparecen los *donantes* enumerados cada uno con su *peso* wd_i correspondiente, y en la primera fila aparecen los *receptores* enumerados

tza bere wr_j *pisua*rekin, non $i \in \{1, ..., p\}$ eta $j \in \{1 ,..., q\}$ diren, eta p eta q estratu horretako *emaile* eta *hartzaile* kopurua diren, hurrenez hurren. Gainera, d_{ij}-ek *i emaile*aren eta *j hartzaile*aren arteko distantzia adierazten du. *Hartzaile* guztien *pisuen* baturak guztira *wr* ematen du, eta *emaile* guztien *pisu*en baturak guztira *wd* ematen du. Tribiala da *wr= wd* estratu horri dagokion populazio osoaren berdina dela.

Distantzia-matrizea sortu ondoren, azken urratsaren helburua *emaile-hartzaile* loturak ezartzea da. Horretarako, w_{ij} aldagai hau definitzen da: «*i emaile*aren eta *j hartzaile*aren informazioarekin bat egin ondoren sortutako erregistroaren *pisua*». Aldagai horrek 0 balioa hartzen du *i emaileak j hartzaile*ari informaziorik ematen ez badio, edo 0 baino balio handiagoa, *i emaileak j hartzaile*ari w_{ij} *pisua* duen informazioa ematen badio. Aldagai hori horrela definituta, hurrengo murrizketak planteatzen dira:

cada uno con su respectivo *peso* wr_j, donde $i \in \{1, ..., p\}$ y $j \in \{1 ,..., q\}$, siendo p y q el número de *donantes* y *receptores* en ese estrato, respectivamente. Además, d_{ij} indica la distancia entre el *donante i* y el *receptor j*. La suma de los *pesos* de todos los *receptores* da un total de *wr* y la suma de los *pesos* de todos los *donantes* da un total de *wd*. Es trivial que *wr= wd* es igual al total de la población correspondiente a ese estrato.

Una vez creada la *matriz de distancias*, el último paso tiene como objetivo establecer los vínculos *donante- receptor*. Para ello, se define la siguiente variable w_{ij}: «*Peso* del registro creado tras la fusión con la información del *donante i* y el *receptor j*». Esta variable toma valor 0 si el *donante i* no le traspasa información al *receptor j*, o un valor mayor que 0, si, por el contrario, el *donante i* le traspasa información al *receptor j* con un *peso* de w_{ij}. Definida así esta variable, se crean las siguientes restricciones:

$$\bullet w_{11} + w_{21} + ... + w_{p1} = wr_1$$

$$\bullet w_{12} + w_{22} + ... + w_{p2} = wr_2$$

...

$$\bullet w_{1q} + w_{2q} + ... + w_{pq} = wr_q$$

$$\bullet w_{11} + w_{12} + ... + w_{1q} = wd_1$$

$$\bullet w_{21} + w_{22} + ... + w_{2q} = wd_2$$

...

$$\bullet w_{p1} + w_{p2} + ... + w_{pq} = wd_p$$

Sistema horrek *pxq* ezezagun eta *p+q* murrizketa ditu; beraz, sistema bateragarri indeterminatua da, hau da, soluzio posible bat baino gehiago ditu. Horregatik, edozein irtenbide hartu beharrean, metodologia-proposamen honen helburua irtenbide optimoa aurkitzea da. Kasu honetan, indibiduoen arteko distantziak neurtzen direnez, helburua da d_{ij} distantzia eta w_{ij} *pisu*en biderkadura guztien arteko batura ahalik eta txikiena izatea. Hau da:

Este sistema tiene *pxq* incógnitas y *p+q* restricciones, por lo que es un sistema compatible indeterminado, es decir, tiene más de una solución posible. Por ello, en lugar de tomar cualquiera de las posibles soluciones, esta propuesta metodológica trata de encontrar la solución óptima. En este caso, al estar midiendo distancias entre individuos, el objetivo es conseguir que la suma de todas las distancias d_{ij} multiplicadas por los *pesos* w_{ij} sea la mínima posible. Es decir:

$$(H.F.)/(F.O.)\ min\ Z = d_{11}w_{11}+d_{21}w_{21}+...+d_{p1}w_{p1}+...+d_{1q}w_{1q}+d_{2q}w_{2q}+...+d_{pq}w_{pq}$$

Helburu-funtzio (*H.F.*) hori eta lehenago planteatutako murrizketak kontuan hartuta optimizazio-problema bat lortzen da, Simplex Algoritmoaren bidez ebazten dena. Problema horren emaitza w_{11}, w_{12}, ..., w_{1q}, w_{21}, w_{22}, ..., w_{2q}, ... w_{p1}, w_{p2}, ..., w_{pq} *pisu*en balioak dira. Orain, lortutako *pisu* guztiak erabiliz, estratu bakoitzerako datu-base bat sortzen da, non erregistro bakoitzerako *i emaile*aren aldagai guztien informazioa, *j hartzaile*aren aldagai guztien informazioa eta w_{ij} *pisua* adierazten duen azken zutabea dauden.

Amaitzeko, estratu bakoitzerako fusionatutako datu-baseak lortu ondoren, guztiak batzen dira, fusioaren azken datu-basea lortzeko.

Teniendo en cuenta esta función objetivo (F.O.) y las restricciones planteadas anteriormente, se tiene un problema de optimización, el cual se resuelve mediante el Algoritmo Simplex. La solución de este problema son los valores de los *pesos* w_{11}, w_{12}, ..., w_{1q}, w_{21}, w_{22}, ..., w_{2q}, ... w_{p1}, w_{p2}, ..., w_{pq}. Ahora, utilizando todos los *pesos* conseguidos, se crea una base de datos para cada estrato en la que para cada registro se tiene la información de todas las variables del *donante i*, la información de todas las variables del *receptor j*, y una última columna que indica el *peso* w_{ij}.

Para terminar, una vez conseguidas las bases de datos fusionadas para cada estrato, se unen todas ellas, consiguiendo así la base de datos final de la fusión.

4.4. EUSTAT eta AIMCren lanei eginiko ekarpenak

Aurretik aipatu den bezala, EUSTAT eta AIMC oso lagungarriak izan dira proiektu hau aurrera eramateko. Biek eman duten informazioa ezinbestekoa izan da datu-fusioaren kontzeptua ulertzeko eta hura gauzatzeko metodoak ezagutzeko. Hala ere, garrantzitsua da adieraztea proiektu honetan ez dela ezaguna zen metodologietako bat ere ez aplikatu; aitzitik, urrats berriak eman dira metodologiaren planteamenduan. Horregatik, jarraian labur-labur azaltzen da EUSTAT eta AIMCren fusioen funtzionamendua eta horiekiko egindako ekarpenak.

Alde batetik, EUSTATek erregistro-lotura (*record linkage*) deritzona erabiltzen du nagusiki. Metodo horren helburua da unitate berak identifikatzea eskura dauden datu-iturrietan. Datuen integrazioaren oinarria unitateak haien identifikatzaileen bidez lotzea da; unitatearen identifikatzailea errorerik gabe erregistratutako kode bakarra bada (adibidez, identifikazio-zenbaki pertsonala), orduan lotura zehatza egin daiteke. Aldiz, lotura-kodeak erroreak baditu edo ez bada bakarra, eta unitatea zenbait aldagai gakoren bidez identifika badaiteke (izena, abizena, adina, sexua, eta abar), erregistroen lotura probabilistikoa aplikatzen da. Hau da, zenbait datu-iturritako bi erregistroren artean, uni-

4.4. Aportaciones respecto a los trabajos de EUSTAT y AIMC

Como se ha mencionado en el capítulo de introducción, tanto EUSTAT como AIMC han sido de gran ayuda para poder llevar a cabo este proyecto. Ambos han facilitado información que ha sido indispensable a la hora de entender el concepto de fusión de datos y conocer los diferentes métodos para llevarla a cabo. Sin embargo, es importante señalar que este proyecto no ha consistido simplemente en aplicar una de las metodologías ya existentes, sino que, por el contrario, ha dado nuevos pasos en el planteamiento de la metodología. Por ello se explica brevemente a continuación el funcionamiento de las fusiones de EUSTAT y AIMC, así como y las aportaciones respecto a estas.

Por un lado, EUSTAT utiliza principalmente lo que es conocido como enlace de registros (*record linkage*). El objetivo de este método es identificar las mismas unidades en las diferentes fuentes de datos disponibles. La integración de los datos se basa en unir las unidades mediante identificadores de los mismos; si el identificador de la unidad es un código único (por ejemplo, un número de identificación personal) registrado sin errores, entonces se puede hacer el enlace exacto. Sin embargo, si el código de enlace tiene errores o no es único y la unidad puede ser identificada por varias variables cla-

tate bera izateko probabilitatea estimatzen da, aldagai gakoetan behatutako balioak kontuan hartuta (D'Orazio, 2013).

Bestalde, AIMCk datuak fusionatzen ditu bost inkestetatik lortutako informazioa oinarritzat hartuz: EGM Radio, EGM Prensa, EGM Publicaciones, EGM TV eta EGM Multimedia. Lehenengo lauen kasuan, irratiaren, prentsaren, aldizkarien eta telebistaren kontsumoari buruz soilik galdetzen da, hurrenez hurren; EGM Multimedian, ordea, lau medio horien eta beste euskarri batzuen kontsumoari buruz galdetzen da.

Laginak handiak izan ohi dira: 2020ko edizioan EGM Radio-n 79.100 lagun lagin teorikoan, EGM Prensa-n 75.100, EGM Revistas-en 51.900, EGM TV-n 43.000, eta EGM Multimedia-n beste 30.000; orotara, beraz, 279.000 galdetegi urte bakarrean. Horrek zera ahalbidetu du: irratiari buruzko informazioa, 129.100 inkestatatik ateratzea (irratikoa gehi multimedia); prentsari buruzkoa, 105.000tatik (prentsakoa gehi multimedia); aldizkariei buruzkoa, 81.900tatik (aldizkarietakoa gehi multimedia); eta telebistei buruzkoa 73.000tatik (telebistakoa gehi multimedia) (Asociación para la Investigación de Medios de Comunicación AIMC, 2022).

Fusioaren helburua da medio eta elementu guztien informazioa (nombre, apellido, edad, sexo, etc.), se aplica el enlace de registros probabilístico. Es decir, se estima la probabilidad de que dos registros en diferentes fuentes de datos se refieran a la misma unidad, teniendo en cuenta los valores observados en las variables clave (D'Orazio, 2013).

Por otro lado, AIMC realiza fusiones de datos con información proveniente de cinco encuestas diferentes: EGM Radio, EGM Prensa, EGM Revistas, EGM TV y EGM Multimedia. En el caso de las primeras cuatro se pregunta exclusivamente sobre el consumo de radio, prensa, revistas y televisión, respectivamente. En la quinta, EGM Multimedia se pregunta sobre el consumo de los cuatro medios y otros soportes.

Las muestras suelen ser grandes: en la edición de 2020 la muestra teórica de EGM Radio estaba compuesta por 79.100 personas, la de EGM Prensa por 75.100, a de EGM Revistas por 51.900, la de EGM TV por 43.000, y la de EGM Multimedia por otras 30.000. Esto supone 279.000 cuestionarios en un solo año. Ello permite obtener información de 129.100 encuestas (radio más multimedia) en el caso de la radio, de 105.000 (prensa más multimedia) en el de la prensa, de 81.900 (revistas más multimedia) en el de las revistas y de 73.000 (televisión más multimedia) en el de la prensa televisión (Asociación para la Investigación de Medios de Comunicación AIMC, 2022).

mazioa duen fitxategi bakarra lortzea. Horretarako, proiektu honetan proposatutakoaren antzeko metodologia erabiltzen da. Lehenik, orekatze-matrize berriak eraikitzen dira, hau da, datu-baseetarako estratifikazio berriak sortzen dira. Horren ondoren, estratuak sortzen dira planteatutako estratifikazioak erabiliz, eta estratu horien barruan, indibiduoen arteko distantziak kalkulatzen dira (haiek proposatutako distantzia-neurri bat erabiliz). Ondoren, informazioa transferitzeko distantziak ordenatzen dira, distantzia txikienetik hasita, eta horrela hurrenez hurren, informazio guztia transferitzen den arte (ODEC & QUINAO, 2009). Hona hemen gakoa: horrek ez du bermatzen transferitutako informazioaren distantzia totala ahalik eta txikiena izango denik. Horregatik, gure ekarpen metodologikoa da Simplex Algoritmoa erabiltzea problemaren irtenbide optimoa lortzeko, hau da, fusiorako ahalik eta distantzia total txikiena eskuratzeko.

Jarraian, bien arteko aldea ilustratzeko adibide gisa, AIMCk erabiltzen duen teknika jarraituta distantzia-matrizearen ebazpena nolakoa izango litzatekeen, eta optimizazioa erabiliz nola ebatziko litzatekeen.

El objetivo de la fusión es conseguir un solo fichero con la información de todos los medios. Para ello se utiliza una metodología muy similar a la propuesta en este proyecto. Primero, se construyen unas matrices de equilibraje nuevas, es decir, nuevas estratificaciones para las bases de datos. Tras ello, se crean estratos utilizando las estratificaciones planteadas y dentro de esos estratos se calculan las distancias entre los individuos (utilizando una medida de distancia propuesta por ellos mismos). Después, se ordenan las distancias para transferir la información empezando desde los individuos que están a menos distancia, y así sucesivamente hasta que se haya trasferido toda la información (ODEC & QUINAO, 2009). El problema de todo esto, sin embargo, es que no garantiza que la distancia total de la información transferida sea la mínima posible. Por ello, nuestra aportación metodológica es utilizar el Algoritmo Simplex para conseguir la solución óptima del problema, es decir, la menor distancia total posible.

Con el fin de ilustrar la diferencia entre ambos métodos, se ofrece un ejemplo de cómo sería la solución de la matriz de distancias con la ordenación que utiliza AIMC y cómo es la solución utilizando optimización.

Har bedi honako distantzia-matrizea:

Sea la siguiente matriz de distancias:

Emaileak/Hartzaileak / *Donantes/Receptores*	1 (20)	2 (25)	3 (55)
1 (10)	0.1	0.25	0.5
2 (15)	0	0.3	0.2
3 (50)	0.15	0.55	0.35
4 (25)	0.45	0.5	0.05

Honek 4 *emaile* eta 3 *hartzaile* ditu. Matrizean, alde batetik, 1etik 4ra zerrendatutako *emaileak* daude, parentesi artean laginean dagozkien *pisu*ekin; eta, bestetik, 1etik 3ra zerrendatutako *hartzaileak*, laginean dagozkien *pisu*ekin; eta, azkenik, *emaile*en eta *hartzaile*en arteko distantziak.

La cual tiene 4 *donantes* y 3 *receptores*. En la matriz se observan por un lado los *donantes* enumerados del 1 al 4 con sus respectivos *pesos* en la muestra entre paréntesis, los *receptores* enumerados del 1 al 3 también con sus respectivos *pesos* en la muestra y las distancias entre los *donantes y receptores*.

Distantzien ordena erabiliz, problema honela ebatziko litzateke:

Utilizando la ordenación de distancias, el problema se resolvería de la siguiente manera:

1. Distantziarik txikiena d_{21}= 0 da; beraz, 2. *emaileak* 1. *hartzaile*ari informazioa transferitzen dio 15eko *pisu*arekin (w_{21}= 15), *emailea* desagertu egiten da —kolore gorriz adieraziko dugu desagertzen den errenkada edo

1. La menor distancia es d_{21}= 0, por lo que el *donante* 2 le transfiere información al *receptor* 1 con un *peso* de 15 (w_{21}= 15), el *donante* desaparece —indicamos en color rojo la fila o columna que desaparece— porque ya ha

Emaileak/Hartzaileak / *Donantes/Receptores*	1 (5)	2 (25)	3 (55)
1 (10)	0.1	0.25	0.5
2 (15)	0	0.3	0.2
3 (50)	0.15	0.55	0.35
4 (25)	0.45	0.5	0.05

zutabea— bere informazio guztia transferitu baitu, eta *hartzaileak* bere horretan jarraitzen du, oraindik ere 5eko *pisua* behar duelako.

transferido toda su información y el *receptor* permanece porque todavía necesita un *peso* de 5.

2. Distantziarik txikiena orain d_{43}= 0.05, da; beraz, 4. *emaileak* 3. *hartzaile*ari informazioa transferitzen dio 25eko *pisu*arekin (w_{43}= 25), 4. *emailea* desagertu egiten da eta 3. *hartzailea* mantendu egiten da 30ko *pisu*arekin.

2. La menor distancia ahora es d_{43}= 0.05, por lo que el *donante* 4 le transfiere información al *receptor* 3 con un *peso* de 25 (w_{43}= 25), desaparece el *donante* 4 y se mantiene el *receptor* 3 con un *peso* de 30:

Emaileak/Hartzaileak / *Donantes/Receptores*	1 (5)	2 (25)	3 (30)
1 (10)	0.1	0.25	0.5
2 (15)	0	0.3	0.2
3 (50)	0.15	0.55	0.35
4 (25)	0.45	0.5	0.05

3. Distantziarik txikiena orain d_{11}= 0.1 da; beraz, 1. *emaileak* 1. *hartzaile*ari informazioa transferitzen dio 5eko *pisu*arekin (w_{11}=5), 1. *hartzailea* desagertu egiten da eta 1. *emailea* mantendu egiten da 5eko *pisu*arekin.

3. La distancia mínima es ahora d_{11}= 0.1, por lo que el *donante* 1 transfiere información al *receptor* 1 con un *peso* de 5 (w_{11}=5), el *receptor* 1 desaparece y el *donante* 1 se mantiene con un *peso* de 5.

Emaileak/Hartzaileak / *Donantes/Receptores*	1 (5)	2 (25)	3 (30)
1 (5)	0.1	0.25	0.5
2 (15)	0	0.3	0.2
3 (50)	0.15	0.55	0.35
4 (25)	0.45	0.5	0.05

4. Distantziarik txikiena orain d_{12}= 0.25 da; beraz, 1. *emaileak* 2. *hartzaile*ari informazioa transferitzen dio 5eko *pisu*arekin (w_{12}=5), 1. *emailea* desagertu egiten da eta 2. *hartzailea* mantentzen da 20ko *pisu*arekin.

4. La distancia mínima es ahora d_{12}= 0.25, por lo que el *donante* 1 transfiere información al *receptor* 2 con un *peso* de 5 (w_{12}=5), el *donante* 1 desaparece y el *receptor* 2 se mantiene con un *peso* de 20.

Emaileak/Hartzaileak / *Donantes/Receptores*	1 (5)	2 (25)	3 (30)
1 (5)	0.1	0.25	0.5
2 (15)	0	0.3	0.2
3 (50)	0.15	0.55	0.35
4 (25)	0.45	0.5	0.05

5. Distantziarik txikiena orain d_{33}= 0.35 da; beraz, 3. *emaileak* 3. *hartzaile*ari informazioa transferitzen dio 30eko *pisuarekin* (w_{33}= 30), 3. *hartzailea* desagertu egiten da eta 3. *emailea* mantentzen da 20ko *pisu*arekin.

5. La distancia mínima es ahora *d*33= 0.35, por lo que el *donante* 3 transfiere información al *receptor* 3 con un *peso* de 30 (*w*33= 30), el *receptor* 3 desaparece y se mantiene el *donante* 3 con un *peso* de 20.

Emaileak/Hartzaileak / *Donantes/Receptores*	1 (5)	2 (20)	3 (30)
1 (5)	0.1	0.25	0.5
2 (15)	0	0.3	0.2
3 (25)	0.15	0.55	0.35
4 (25)	0.45	0.5	0.05

6. Azkenik, azkenengo *pisua* w_{32}=20 da; beraz, bai 3. *emailea* eta bai 2. *hartzailea* desagertu egiten dira.

6. Por último, el último *peso* es w_{32}=20, y tanto el *donante* 3 como el *receptor* 2 desaparecen.

Orduan, *pisu* ez-nuluak hurrengoak dira: w_{11}=5, w_{12}=5, w_{21}=1, w_{32}=20, w_{33}=30 eta w_{43}=25. Beraz, distantzia eta *pisu*en arteko biderketaren batura kalkulatuz, helburu-funtzioaren emaitza honakoa da: $Z = 5 \cdot 0.1 + 5 \cdot 0.25 + 15 \cdot 0 + 20 \cdot 0.55 + 30 \cdot 0.35 + 0.05 \cdot 25 = 24.5$.

Hala ere, distantziak antolatu beharrean Simplex Algoritmoa erabil dezakegu, problema beste modu batera ebatzita. Horrela, lortutako *pisu* ez-nuluak honako hauek lirateke: w_{12}=10, w_{22}=15, w_{31}=20, w_{33}=30 eta w_{43}=25. Ondorioz, distantzia eta *pisu*en arteko biderketaren batura kalkulatuz, helburu-funtzioaren emaitza honakoa da: $Z = 10 \cdot 0.25 + 15 \cdot 0.3 + 20 \cdot 0.15 + 30 \cdot 0.35 + 25 \cdot 0.05 = 21.75$.

Beraz, nahiz eta bi erantzunak baliozkoak izan planteatutako problemarako, Simplex Algoritmoari esker distantziaren balio minimoa lortzen da, hau da, ebazpen optimoa.

4.5. Fusioaren baliozkotzea

Fusioaren datu-basea lortu ondoren, lortutako emaitza baliozkotu behar da. Horretarako, Rässlerek honakoak proposatzen ditu (Rässler, 2002):

- Begiratu egotzitako aldagaiek (hots, datu-base *emaile*koek) banaketa marjinala mantentzen duten, datu-base *emailea* erreferentzia gisa

Por lo tanto, los *pesos* no-nulos conseguidos son los siguientes: w_{11}=5, w_{12}=5, w_{21}=1, w_{32}= 20, w_{33}= 30 y w_{43}= 25. Por lo que, si calculamos la suma del producto de las distancias y los *pesos*, el resultado de la función objetivo es $Z= 5 \cdot 0.1 + 5 \cdot 0.25 + 15 \cdot 0 + 20 \cdot 0.55 + 30 \cdot 0.35 + 0.05 \cdot 25 = 24.5$.

Sin embargo, si en lugar de ordenar las distancias podemos utilizar el Algoritmo Simplex y resolver el problema de manera diferente. Así, los *pesos* no-nulos obtenidos serían como siguen: w_{12}=10, w_{22}=15, w_{31}=20, w_{33}=30 y w_{43}= 25. Por lo que, si se calcula la suma del producto de las distancias y los *pesos*, el resultado de la función objetivo es $Z= 10 \cdot 0.25 + 15 \cdot 0.3 + 20 \cdot 0.15 + 30 \cdot 0.35 + 25 \cdot 0.05 = 21.75$.

Por lo que, aunque ambas respuestas son válidas para el problema planteado, gracias al Algoritmo Simplex, se consigue el valor mínimo de la distancia, esto es, la solución óptima.

4.5. Validación de la fusión

Una vez conseguida la base de datos de la fusión, hay que validar el resultado obtenido. Para ello, Rässler propone lo siguiente (Rässler, 2002):

- Comprobar que las variables imputadas (es decir, las de la base de datos *donante*) conservan la distribución marginal, tomando la base de datos

hartuta. Hau da, egotzitako aldagaiak berdin banatuta egotea datu-base *emaile*an eta fusionatutako datu-basean.

- Begiratu egotzitako aldagaiek lotura-aldagaiekin duten baterako banaketa mantentzen duten, datu-base *emailea* erreferentzia gisa hartuta. Hau da, egotzitako aldagaiak berdin banatuta egotea lotura-aldagaien arabera datu-base *emaile*an eta fusionatutako datu-basean.

Azterketa hori egiteko, antzekotasun-indizea, Bhattacharyyaren koefizientea eta beste erabil daitezke aldagai kategorikoentzat eta aurretik aipatutako estatistiko deskribatzaileak aldagai jarraituentzat.

5. Proposamen metodologikoaren aplikazioa

Kapitulu hau bi zati nagusitan banatzen da. Lehenengo zatian, CIES eta IKUSIKER datu-baseetan proposamen metodologikoa nola aplikatu den azaltzen da; bigarren zatian, berriz, lortutako emaitzak baliozkotzen dira.

5.1. Metodologiaren aplikazioa

Aurreko kapituluan aipatu den bezala, sortutako metodologiak

donante como referencia. Es decir, que las variables imputadas estén igualmente distribuidas en la base de datos *donante* y en la base de datos de la fusión.

- Comprobar que las variables imputadas mantienen la distribución conjunta con las variables de unión, tomando la base de datos *donante* como referencia. Es decir, que las variables imputadas estén igualmente distribuidas respecto a las variables de unión en la base de datos *donante* y en la base de datos de la fusión.

Para realizar este análisis se pueden utilizar el índice de similitud, el coeficiente de Bhattacharyya, etc. explicados anteriormente para las variables categóricas, así como los estadísticos descriptivos también mencionados para las variables continuas.

5. Aplicación de la propuesta metodológica

Este capítulo se divide en dos partes principales. En la primera parte se explica el proceso de cómo se ha aplicado la propuesta metodológica a las bases de datos de CIES e IKUSIKER, y en la segunda parte se hace la validación de los resultados obtenidos.

5.1. Aplicación de la metodología

Como se ha mencionado en el capítulo anterior, la metodolo-

bost urrats nagusi ditu. Jarraian, CIES eta IKUSIKER datu-baseetarako bost urrats horiek nola aplikatu diren zehazten da.

5.1.1. Lehen urratsa: datu-baseen arteko estratifikazio komuna identifikatu

Lehenengo urratsaren helburua bi datu-baseen arteko estratifikazio komuna identifikatzea da. Horretarako, lehenik fusioaren unibertsoa zehaztu behar da.

Datu-baseen kapituluan azaldu den bezala, IKUSIKER panelaren unibertsoa Araba, Bizkaia, Gipuzkoa eta Nafarroako 11 eta 23 urte bitarteko ikasle gazteek osatzen dute (bi urteko taldeetan banatuta), eta CIESen unibertsoa lau probintzia horietako 14 urtetik gorako biztanleek (bost urteko taldeetan). Horrenbestez, bien arteko populazio komuna 14 eta 23 urte bitarteko gazteak dira; baina CIES bost urteko taldeetan banatzen denez, fusiorako unibertso komuna lau probintzietako 14-19 urte bitarteko gazteak dira. Bestalde, kontuan hartu behar da IKUSIKER panelak Derrigorrezko Hezkuntzako, Batxilergoko eta unibertsitateko ikasleen datuak baino ez dituela jasotzen, Lanbide-Heziketako ikasle gazteak eta ez ikasleak alde batera utzita, eta, beraz, unibertsoak ez datozela guztiz bat. Hala ere, lau lurralde historiko

gía creada consta de cinco pasos principales. A continuación, se detalla cómo se han aplicado estos pasos para las bases de datos de CIES e IKUSIKER.

5.1.1. Primer paso: identificar la estratificación común entre las bases de datos

El objetivo del primer paso es identificar la estratificación común entre ambas bases de datos. Para ello, primero hay que fijar cuál es el universo de la fusión.

Como se ha explicado en el capítulo de bases de datos, el universo del panel IKUSIKER son las y los jóvenes estudiantes de entre 11 y 23 años (en grupos de dos años) de Araba, Bizkaia, Gipuzkoa y Nafarroa, y el universo de CIES son las personas mayores de 14 años (en grupos quinquenales) de las mismas cuatro provincias. Por ello, la población coincidente entre ambas son las personas de entre 14 y 23 años. Sin embargo, dado que CIES divide la población en grupos quinquenales, el universo común para la fusión son las y los jóvenes de entre 14 y 19 años de las cuatro provincias. Es necesario también tener en cuenta que IKUSIKER solo recoge datos de estudiantes de Educación Secundaria Obligatoria, Bachiller y universidad, dejando a un lado a jóvenes estudiantes de Formación Profesional y a no estudiantes, por lo que los universos no son totalmente coincidentes. Si consideramos que en este rango de edad en el conjunto

horietan gazteen %94,8 ikasleak dira, eta, beraz, ikasle ez direnak kontuan ez hartzeagatik dagoen desbiderapena oso txikia da. Handiagoa da ordea Lanbide-Heziketako ikasleak ez sartzeak eragindakoa, adin tarte horretako gazteen %16,8 baitira. Etorkizunean, IKUSIKER datu-bilketak fusioaren bide honetatik jarraitzen badu, LHko panelkideak eta ikasle ez diren gazte panelkideak sartu ahal izango ditu laginean. Edonola ere, aplikaziorako laginetan zehaztu den unibertsoak 1604 panelkide biltzen ditu IKUSIKER datu-basean, eta 430 gazte CIESen. Horrek esan nahi du IKUSIKER datu-base *emailea* izango dela eta CIES, berriz, *hartzailea*.

Fusiorako erabiliko den unibertsoa zehaztu ondoren, laginen arteko estratifikazio komuna bilatu da, dohaintza-klaseak sortu ahal izateko. IKUSIKER panelaren kasuan, biztanleria sexuaren, adinaren (bi urteko taldeetan) eta probintziaren arabera banatuta dago, eta hori bat dator CIESeko sexuaren eta probintziaren arteko banaketarekin. Aitzitik, adin-taldeak ez datoz bat.

Hasieran zortzi dohaintza-klase sortzea erabaki zen, populazioa sexuaren eta probintziaren arabera banatuta, baina horrek bi arazo nagusi sortzen zituen. Bat: dohaintza-klaseak handiegiak ziren. Bi: CIES bost urteko

de las cuatro provincias el 94,8% de las y los jóvenes son estudiantes, la desviación por no tener en cuenta a los y las no estudiantes es pequeña. Sí puede ser mayor al no incluir en la muestra a estudiantes de Formación Profesional, que son el 16,8% de la población de ese tramo de edad. Aun así, se ha decidido tomar la muestra de estas y estos jóvenes estudiantes como representantes de la toda la población juvenil. En un futuro, si IKUSIKER sigue por este camino será capaz de introducir en la muestra a panelistas de FP y jóvenes no estudiantes. Así las cosas, el universo mencionado suma un total de 1.604 panelistas en IKUSIKER, y en CIES un total de 430 jóvenes. Lo que implica que IKUSIKER será la base de datos *donante* y CIES la *receptora*.

Una vez concretado el universo que se va a utilizar para la fusión, se procede a buscar la estratificación común entre las muestras para poder crear las clases de donación. En el caso de IKUSIKER, la población está divida por sexo, edad (grupos de dos años) y provincia, lo cual coincide con la división por sexo y provincia de CIES. Sin embargo, en cuanto a grupos de edad no son coincidentes.

En un principio se decidió crear ocho clases de donación dividiendo la población solo por sexo y provincia, pero esto planteaba dos problemas principales. El primero es que las clases de donación eran demasiado grandes y el

5. taula: CIES eta IKUSIKER datu-baseen arteko estratifikazio komunaren banaketa./ Tabla 5: Distribución de la estratificación común entre las bases de datos de CIES y de IKUSIKER.

Estratua/ Estrato	Sexua/ Sexo	Adin-taldea/ Grupo de edad	Probintzia/ Provincia	Unibertso totala/ Total universo
1	Neska/ Chica	14-15	Araba	3104
2	Neska/ Chica	16-17	Araba	2963
3	Neska/ Chica	18-19	Araba	2884
4	Neska/ Chica	14-15	Gipuzkoa	7199
5	Neska/ Chica	16-17	Gipuzkoa	6985
6	Neska/ Chica	18-19	Gipuzkoa	6765
7	Neska/ Chica	14-15	Bizkaia	10483
8	Neska/ Chica	16-17	Bizkaia	10200
9	Neska/ Chica	18-19	Bizkaia	9799
10	Neska/ Chica	14-15	Nafarroa	6737
11	Neska/ Chica	16-17	Nafarroa	6799
12	Neska/ Chica	18-19	Nafarroa	6890
13	Mutila/Chico	14-15	Araba	3284
14	Mutila/Chico	16-17	Araba	3232
15	Mutila/Chico	18-19	Araba	3181
16	Mutila/Chico	14-15	Gipuzkoa	7583
17	Mutila/Chico	16-17	Gipuzkoa	7421
18	Mutila/Chico	18-19	Gipuzkoa	7304
19	Mutila/Chico	14-15	Bizkaia	11050
20	Mutila/Chico	16-17	Bizkaia	11023
21	Mutila/Chico	18-19	Bizkaia	11029
22	Mutila/Chico	14-15	Nafarroa	7269
23	Mutila/Chico	16-17	Nafarroa	7107
24	Mutila/Chico	18-19	Nafarroa	7173

Azken zutabean estratu bakoitzeko populazio totalak ageri dira. / En la última columna aparecen los totales poblacionales de cada estrato.

taldeen arabera ponderatuta zegoenez, adinaren araberako desoreka zegoen. Hau da, 14-15 urteko pertsonen kopurua, 18-19 urtekoen aldean, oso txikia zen. Horregatik, CIES bi urteko taldeetan birponderatzea erabaki zen; eta hori dela-eta, sexua, adina (bi urteko taldeetan) eta probintzia dira estratifikazio komunerako erabilitako aldagaiak.

Bi laginen arteko estratifikazio komuna identifikatu ondoren, dohaintza-klaseak sortu dira aldagai horietatik abiatuta. Bi sexu kategoria, hiru adin-talde (bi urtekoak) eta lau probintzia daudenez, guztira 24 estratu sortu dira (ikusi 5. taula).

5.1.2. Bigarren urratsa: aldagai komunak aurkitu, harmonizatu eta birsailkatu

Dohaintza-klaseak sortu ondoren, hurrengo urratsa aldagai komunak identifikatzea da, estratuen barruko distantziak kalkulatu ahal izateko.

Guztira, bost aldagai komun erabili dira CIESen eta IKUSIKER panelaren artean:

- **EDA**: adina.
- **HABI**: habitata.
- **SOZ**: gune soziolinguistikoa.

segundo que CIES, al estar ponderado por grupo quinquenal, estaba distribuido de manera descompensada respecto a la edad. Esto es, el número de individuos de 14-15 años comparado con el de 18-19 años era muy bajo. Por ello, se decidió reponderar CIES en grupos de dos años, siendo así el sexo, la edad (en grupos de dos años) y la provincia las variables utilizadas para la estratificación común.

Una vez identificada la estratificación común entre ambas muestras, se crean las clases de donación partiendo de esas variables. Por lo tanto, como hay dos sexos, tres grupos de edad (de dos años) y cuatro provincias, hay un total de 24 estratos (ver tabla 5).

5.1.2. Segundo paso: localizar, armonizar y recategorizar las variables comunes

Una vez creadas las clases de donación, el siguiente paso es identificar las variables comunes para poder calcular las distancias dentro de los estratos.

En total se han utilizado cinco variables comunes entre CIES e IKUSIKER:

- **EDA**: Edad.
- **HABI**: hábitat.
- **SOZ**: zona sociolingüística.

- **KONTS**: euskarazko kontsumoa bezperan.

- **POS161A**: euskara ezagutza.

EDA eta *POS161A* aldagaiak jadanik existitzen ziren eta bi datu-baseetan kategoria berberak zituzten. *HABI* CIESen lehendik zegoen aldagai bat zen, baina ez IKUSIKERren, eta, beraz, IKUSIKERren sortu behar izan da herriaren informazioa ematen duen *HERRI* aldagaiaren bidez. Azkenik, *SOZ* eta *KONTS* aldagaiak bi datu-baseetan sortu behar izan dira: *SOZ* bai IKUSIKERren eta bai CIESen eremu geografikoa adierazten duen *ZONA* aldagaiaren bidez sortu da, eta *KONTS* galderak erantzundako pertsonak bezperan euskarazko medioren bat kontsumitu ote duen zehazten duten hainbat aldagairen bidez sortu da.

Ondoko taulan bildu dira bost aldagaiak, haien esanahia eta hartzen dituzten kategoriak zehaztuta:

- **KONTS**: consumo en euskera el día anterior.

- **POS161A**: Conocimiento de euskera.

EDA y *POS*161*A* eran variables ya existentes y con las mismas categorías en ambas bases de datos. *HABI* era una variable ya existente en CIES, pero no en el panel de IKUSIKER y, por lo tanto, en este panel se ha tenido que crear mediante la variable *HERRI* que da la información del municipio. Por último, las variables *SOZ* y *KONTS* han tenido que ser creadas en ambas bases de datos: *SOZ* se ha creado tanto en el panel de IKUSIKER como en CIES mediante la variable *ZONA* que representa la zona geográfica y *KONTS* se ha creado con variables que determinan si el individuo ha consumido algún medio en euskara la víspera.

En la tabla que sigue se detallan las cinco variables con las especificaciones sobre su significado y las categorías que incluyen.

6. taula: CIES eta IKUSIKER datu-baseen arteko aldagai komunen sailkapena. / Tabla 6: Clasificación de las variables comunes entre las bases de datos de CIES y de IKUSIKER.

Aldagaia/ Variable	**Esanahia/ Significado**	**Kategoriak/ Categorías**
EDA	Adina / Edad	14-19 urte/años
HABI	Habitata / Hábitat	1: <5.000 biztanle / habitantes 2: 5.000-10.000 biztanle / habitantes 3: 10.000-50.000 biztanle / habitantes 4: >50.000 biztanle / habitantes 5: Kapitala / Capital
SOZ	Gune soziolinguistikoa / Zona sociolingüística	1: % 0-9 2: % 10-19 3: % 20-29 4: % 30-39 5: % 40-49 6: % 50-59 7: % 60-69 8: % 70-79 9: % 80-89
KONTS	Euskarazko kontsumoa bezperan / Consumo en euskara en la víspera	0: Ez / No 1: Bai / Sí
POS161A	Euskara ezagutza / Conocimiento de euskara	1: Ez du ulertzen / No entiende 2: Ulertzen du / Entiende 3: Hitz egiten du / Habla 4: Hitz egin eta irakurtzen du / Habla y lee 5: Hitz egin, irakurri eta idazten du / Habla, lee y escribe

Proposamen metodologikoaren kapituluan aipatu den bezala, garrantzitsua da egiaztatzea lotura-aldagaien banaketa marjinala bat datorrela bi datu-baseetan. Kasu honetan aldagai guztiak kategorikoak direnez, banaketa marjinalen arteko komunztadura begiratzeko, *R* programako *StatMatch* paketearen *comp.prop* funtzioa erabili da, *tvd*, *Bhatt*, *overlap* eta *Hell* balioak itzultzen dituena. Hauek dira aldagai bakoitzarentzat lortutako emaitzak:

Como se ha mencionado al explicar la metodología, es importante comprobar que la distribución marginal de las variables de unión es concordante en ambas bases de datos. Como en este caso todas las variables son categóricas, para mirar la concordancia entre las distribuciones marginales se ha utilizado la función *comp.prop* del paquete *StatMatch* del programa *R* que devuelve los valores de *tvd*, *Bhatt*, *overlap* y *Hell*. Estos son los resultados obtenidos para cada una de las variables:

7. taula: CIES eta IKUSIKER datu-baseen arteko lotura-aldagaien banaketa marjinala bat datorrela egiaztatzeko ariketa./ Tabla 7: Ejercicio de comprobación de la concordancia de las distribuciones marginales de las variables de las bases de datos de CIES y de IKUSIKER.

Aldagaia / Variable	***tvd***	***overlap***	***Bhatt***	***Hell***
EDA	0.23	0.77	0.97	0.18
HABI	0.13	0.87	0.98	0.12
SOZ	0.28	0.72	0.93	0.26
KONTS	0.02	0.98	0.99	0.02
POS161A	0.25	0.75	0.94	0.24

Erantzunen taulan ikus daitekeenez, *KONTS* aldagaiak bakarrik betetzen ditu aurrez finkatutako baldintzak, hau da, *tvd* ≤ 0.06, *overlap* ≥ 0.94 eta *hell* ≤ 0.05. Horrek esan nahi du CIES eta IKUSIKERren *KONTS* aldagaiaren banaketa marjinalen arteko antzekotasun-indizea 0.02 dela, gainjartze-indizea 0.98 eta Hellingerren distantzia 0.02, eta, beraz, bat datozela *KONTS* aldagaiaren banaketa marjinalak CIES eta IKUSIKER datu-basetan. Gainerako aldagaien kasuan, finkatutako tarte horietatik ateratzen direnez, ez da egiaztatzen bi datu-baseetan banaketa marjinal konkordanteak dituztenik. Hala ere, bost horiek lotura-aldagai gisa erabiltzea erabaki da, bi arrazoirengatik: lehenengoa, loturarako aldagai bakarra erabiltzea arriskutsuegia delako — horrek esan nahi baitu *emaileak* eta *hartzaileak* haren arabera baino ez direla lotzen, eta, beraz, guztiz desberdinak izan dai-

Como se puede observar en la tabla de respuestas, solo la variable *KONTS* cumple los rangos fijados anteriormente, es decir, *tvd* ≤ 0.06, *overlap* ≥0.94 y *hell* ≤0.05. Esto implica que el índice de disimilitud entre las distribuciones marginales de la variable *KONTS* en CIES e IKUSIKER es 0.02, el índice de superposición 0.98 y la distancia de Hellinger 0.02, y que, por lo tanto, las distribuciones marginales de la variable *KONTS* en las bases de datos de CIES y de IKUSIKER son concordantes. En el caso de las demás variables, al salirse de estos rangos, no se verifica que tengan distribuciones marginales concordantes en las dos bases de datos. Aun así, se ha decidido utilizar las cinco como variables de unión por dos motivos. El primero, porque la utilización de una única variable para la unión es demasiado arriesgada ya que implica que los *donantes* y *receptores* solo se enlazan dependiendo

tezkela beste aldagai komunetan—; eta, bigarrena, lortutako emaitzek ez dutelako esan nahi banaketa marjinalak guztiz bat ez datozenik —eta, beraz, proiektu pilotua izanik, nahiago izan da lotura-aldagai gehiago erabili, lortuko diren emaitzak ez direla guztiz zehatzak izango onartuta—.

de ésta, y por lo tanto, pueden ser completamente diferentes en las otras variables comunes. Y el segundo, porque los resultados conseguidos no implican que las distribuciones marginales sean completamente discordantes. Y, por lo tanto, siendo un proyecto piloto, se ha preferido usar más variables de enlace asumiendo que los resultados que se van a obtener no serán del todo exactos.

5.1.3. Hirugarren urratsa: indibiduoen arteko distantziak kalkulatu

Distantziak kalkulatzeko lotura-aldagaiak aukeratu ondoren, distantziak kalkulatu behar dira. Jarraian, distantziak nola kalkulatzen direneko adibide bat azaltzen da.

Etsenplu gisa, lehenengo estratutik CIESeko eta IKUSIKERreko indibiduo bana hartuko dugu. 1. estratukoak direnez, indibiduo hauek Arabako 14-15 urteko neskak dira. Hurrengo taulan, bi indibiduo horien lotura-aldagaien informazioa eta aldagai horietako bakoitzerako indibiduoen arteko antzekotasunaren kalkulua ageri dira.

5.1.3. Tercer paso: calcular las distancias entre los individuos

Después de elegir las variables de unión con las que se van a calcular las distancias, se procede al cálculo de distancias. A continuación, se da un ejemplo de cómo se calculan las distancias.

Ejemplo. Vamos a coger del primer estrato un individuo de CIES y otro de IKUSIKER. Al ser del estrato 1, estos individuos son chicas de 14-15 años de Araba. En la siguiente tabla viene dada la información de las variables de unión de estos dos individuos y el cálculo de la similitud entre los individuos para cada una de estas variables.

8. taula: CIES eta IKUSIKER datu-baseetako indibiduoen arteko distantzien kalkulurako adibidea. / Tabla 8. Ejemplo del cálculo de distancias entre los individuos de las bases de datos de CIES y de IKUSIKER.

Aldagaia / Variable	**CIES-en indibiduoa (x_1) / Individuo de CIES (x_1)**	**IKUSIKERren indibiduoa (x_2) / Individuo de IKUSIKER (x_2)**	**$s_j(x_1,x_2)$**
EDA	14	14	$1-\frac{\lvert 14-14\rvert}{1}=1$
HABI	1	5	$1-\frac{\lvert 1-5\rvert}{4}=0$
SOZ	5	3	$1-\frac{\lvert 5-3\rvert}{8}=\frac{3}{4}$
KONTS	0	0	$1-\frac{\lvert 0-0\rvert}{1}=1$
POS161A	2	5	$1-\frac{\lvert 2-5\rvert}{4}=\frac{1}{4}$

Beraz, bi indibiduo horien arteko distantzia honako hau da:

Y por lo tanto la distancia entre estos dos individuos viene dada por:

$$D_{Gower}(x_1,x_2)=1-\frac{1}{5}\left(1+0+\frac{1}{4}+\frac{3}{4}+1\right)=\frac{2}{5}$$

Distantzia hori estratu bereko *emaile* eta *hartzaile* guztien artean kalkulatu behar da. Horretarako, *R* programako *StatMatch* paketeko *Gower* liburutegiko *gower.dist* funtzioa erabili da.

Esta distancia se debe calcular entre todos los *donantes* y *receptores* de un mismo estrato. Para ello, se ha utilizado la función *gower.dist* de la librería *Gower* del paquete *StatMatch* del programa *R*.

5.1.4. Laugarren urratsa: emaile-hartzaile erlazioak ezarri, distantziak eta pisuak kontuan hartuta

Estratu bakoitzerako distantzia guztiak kalkulatu ondoren, distantzia-matrizea sortzen da, eta horren arabera planteatzen da optimizazio-problema. Ondorengo lerroetan, matrize horren eta optimizazio-problemaren

5.1.4. Cuarto paso: establecer las relaciones donante-receptor teniendo en cuenta las distancias

Una vez calculadas todas las distancias para cada estrato, se crea la matriz de distancias y se plantea el problema de optimización en base a ésta. A continuación, se da un ejemplo de dicha matriz y del planteamien-

planteamenduaren adibide bat azaltzen da.

Adibidea

Aurreko urratsean bezala, adibide honetan lehen estratuko informazioarekin emango da azalpena. Estratu horretan 77 erregistro *emaile* daude (IKUSIKER datu-basekoak) eta 7 erregistro *hartzaile* (CIESekoak). Beraz, 77x7 distantziari buruzko informazioa, 7 *hartzaileen pisuak* (lehenengo lerroan) eta 77 *emaileen pisuak* (lehenengo zutabean) dituen matrizea daukagu. Hartara, honako forma du matrizeak:

to del problema de optimización.

Ejemplo

Como en el paso anterior, en este ejemplo se va a tratar con la información del primer estrato. En este estrato hay 77 registros *donantes* (es decir, de IKUSIKER) y 7 registros *receptores* (es decir, de CIES). Por lo tanto, se tiene una matriz con la información de 77x7 distancias y los 7 *pesos* de los *receptores* (en la primera fila) y los 77 de los *donantes* (en la primera columna). Es decir, la matriz tiene la siguiente forma:

9. taula: CIES eta IKUSIKER datu-baseetako *emaile-hartzaileen* erlazioen kalkulurako adibidea. / Tabla 9. Ejemplo del cálculo de relaciones *donante-receptor* de las bases de datos de CIES y de IKUSIKER

Emaileak/Hartzaileak/ Donantes/Receptores	584.6	401.89	308.12	584.6	308.12	452.92	463.75
40.31	0.2	0.0	0.4	0.2	0.4	0.0	0.0
40.31	0.0	0.2	0.6	0.0	0.6	0.2	0.2
40.31	0.2	0.0	0.4	0.2	0.4	0.0	0.0
...	...	...	...	...	...	...	...
40.31	0.2	0.0	0.4	0.2	0.4	0.0	0.0

Beraz, ebatzi beharreko optimizazio-problemaren planteamendua hau da:

(H.F.) min Z= 0.2w_{11}+ 0 .4w_{13}+ 0.2w_{14}+ 0.4w_{15}+ 0.2w_{22}+ 0.6w_{23}+ 0.6w_{25}+ 0.2w_{26}+ 0.2w_{27}+ ...+ 0.2w_{771}+ 0.4w_{773}+ 0.2w_{774}+ 0.4w_{775}.

Por lo tanto, el planteamiento del problema de optimización a resolver es el siguiente:

(F.O.) min Z= 0.2w_{11}+ 0.4w_{13}+ 0.2w_{14}+ 0.4w_{15}+ 0.2w_{22}+ 0.6w_{23}+ 0.6w_{25}+ 0.2w_{26}+ 0.2w_{27}+ ...+ 0.2w_{771}+ 0.4w_{773}+ 0.2w_{774}+ 0.4w_{775}.

Murrizketak:

Restricciones:

- $w_{11}+w_{21}+w_{31}+...+w_{771}= 584.6$
- $w_{12}+w_{22}+w_{32}+...+w_{772}= 401.89$
- $w_{13}+w_{23}+w_{33}+...+w_{773}= 308.12$
- $w_{14}+w_{24}+w_{34}+...+w_{774}= 584.6$
- $w_{15}+w_{25}+w_{35}+...+w_{775}= 308.12$
- $w_{16}+w_{26}+w_{36}+...+w_{776}= 452.92$
- $w_{17}+w_{27}+w_{37}+...+w_{777}= 463.75$
- $w_{11}+w_{12}+w_{13}+w_{14}+w_{15}+w_{16}+w_{17}= 40.31$

Problema hau (eta estratu bakoitzekoa) ebazteko *R* programako *lpSolve* liburutegiko *lp* funtzioa erabili da. Lortutako emaitza *wij pisu* guztien balioak izan dira, non $i \in \{1,2,...,77\}$ eta $j \in \{1,...,7\}$. *Pisu* horiek erabiliz, estratu bakoitzaren fusionatutako datu-baseak lortu dira.

Para resolver este problema (y los de cada uno de los estratos) se ha utilizado la función *lp* de la librería *lpSolve* del programa *R*. El resultado obtenido han sido los valores de todos los *pesos* w_{ij} donde $i \in \{1,2,...,77\}$ y $j \in \{1,...,7\}$. Utilizando estos *pesos* se han conseguido las bases de datos de fusión de cada estrato.

5.1.5. Bosgarren urratsa: emaile-hartzaile erlazioak erabili fusionatutako datu-basea lortzeko

Estratu bakoitzerako fusionatutako datu-baseak lortu ondoren, fusionatutako datu-base osoa lortzea da azken urratsa. Horretarako, datu-baseak bata bestearen azpian jarri izan dira *R* programaren *rbind* funtzioa erabiliz, eta, hala, fusionatutako datu-basea sortu da. Datu-base horretan, guztira, 2.010 erre-

5.1.5. Quinto paso: utilizar las relaciones donante-receptor para obtener la base de datos de la fusión

Una vez conseguidas todas las bases de datos de fusión para cada estrato el último paso es obtener la base de datos total de la fusión. Para ello solo ha habido que colocar una base de datos debajo de la otra utilizando la función *rbind* del programa *R*, creando así la base de datos fusionada. Esta base de datos contiene la información de

gistroren 260 galdera ingururi buruzko informazioa dago, 6.198 aldagaitan banatuta.

Inkestetatik ateratako datuak direnez, errore-marjina eta konfiantza-tartea zehaztea beharrezkoa da. Kasu honetan, bai CIES eta bai IKUSIKER ausazko laginketa estratifikatuetatik datozen laginak direnez —eta ez ausazko laginketa sinpleetatik—, laginen diseinuak errore marjinalean eragiten du, eta, beraz, lagin bakoitzari bere diseinuan oinarritutako balio desberdina dagokio. Hori dela eta, fusioaren errore marjinalaren kalkulua ez da kontu tribiala, eta tentuz aztertu behar da. Literatura zientifikoan oraindik ez dugu konponbide egokirik aurkitu, eta hainbat aukera lantzen hasiak gara —hala nola balio handiagoa eskaintzen duen fitxategiko errore-marjinala ontzat ematea, bat egindako fitxategia inkesta gisa tratatzea eta dagozkion balioak estimatzea, eta beste batzuk—. Hala ere, edozein konponbide-proposamen aurreratu aurretik, gai horri buruz gehiago ikertzea proposatzen dugu.

5.2. Fusioaren baliozkotzea

Fusionatutako datu-basea lortu ondoren, hura baliozkotzen da. Horretarako, bi baliozkotze-mota erabili dira. Lehena, metodologiaren kapituluan azaldu den baliozkotze teknikoa, eta bigarrena, gaian adituak diren pertsonen baliozkotzea.

un total de 2.010 registros de alrededor de 260 preguntas divididas en 6.198 variables.

Tratándose de datos originalmente extraídos de encuestas, es necesario determinar el margen de error y el intervalo de confianza con el que vamos a trabajar. En este caso, como tanto CIES como IKUSIKER son muestras provenientes de muestreos aleatorios estratificados y no de muestreos aleatorios simples, el diseño de las muestras influye en el error marginal, y por lo tanto cada muestra ofrece un valor distinto basado en su propio diseño. Por esa razón el cálculo del error marginal de la fusión no es una cuestión trivial, y debe ser abordada. No hemos encontrado aún en la literatura científica una solución satisfactoria, y estamos trabajando diferentes alternativas, como dar por buenos los valores de error del fichero matriz que ofrezca un valor más alto, tratar el fichero fusionado como una encuesta y estimar los valores correspondientes, y otras. Sin embargo, antes de adelantar cualquier propuesta de solución, nos proponemos investigar más sobre esta cuestión.

5.2. Validación de la fusión

Tras conseguir la base de datos fusionada, se procede a hacer la validación de ésta. Para ello se han utilizado dos tipos de validaciones. La primera, la validación técnica que se ha explicado en el capítulo de metodología, y la segunda la validación de personas expertas en el tema.

5.2.1. Baliozkotze teknikoa

Baliozkotze mota honetan bi azterketa hauek egin dira:

- **Egotzitako aldagaiek banaketa marjinala mantentzen ote duten egiaztatu, datu-base *emailea* erreferentzia gisa hartuta.**

Propietate hau berehalakoa da; izan ere, sortutako metodologiak ziurtatzen du egotzitako datuek informazio guztia modu berean gordetzen dutela datu-base *email*ean eta fusionatutako artxiboan. Hala ere, hori betetzen dela ikusteko, *comp. prop* funtzioa erabili da berriro, IKUSIKERrek fusionatutako datu-baseari egotzitako aldagaien banaketa marjinalak bat datozela egiaztatzeko.

Aldagaien azpimultzo baterako lortutako emaitzak ageri dira jarraian (gainerako aldagaien emaitzak ia berdinak dira):

5.2.1. Validación técnica

En este tipo de validación se han hecho los siguientes dos análisis:

- **Conocer si las variables imputadas conservan la distribución marginal, tomando como referencia la base datos *donante*.**

Esta propiedad es inmediata, ya que la metodología diseñada asegura que los datos imputados conservan toda su información distribuida de la misma manera en el archivo fusionado que en el *donante*. Aun así, para ver que esto se cumple, se ha vuelto a utilizar la función *comp. prop* para comprobar que las distribuciones marginales de las variables imputadas de IKUSIKER a la base de datos de fusión sean concordantes.

En la siguiente tabla aparecen los resultados obtenidos para un subconjunto de variables, siendo los resultados del resto de las variables prácticamente idénticos:

Aldagaia / Variable	***tvd***	***overlap***	***Bhatt***	***Hell***
InstagramG1	0.00064	0.99936	0.99999	0.00084
InstagramG2	0.00003	0.99996	1	0.00005
InstagramG3	0.00015	0.99984	0.99999	0.00018
InstagramG3_Beste	0.00042	0.99958	0.99999	0.00055
InstagramG4	0.00042	0.99958	0.99999	0.00057
InstagramG5	0.00004	0.99995	1	0.00005
InstagramG5_Beste	0.00049	0.9995	0.99999	0.00061
InstagramG6	0.000005	0.99999	1	0.00005

Ikus daitekeenez, aldagai guztiek *tvd* ≤ 0.06, *overlap* ≥ 0.94 eta *hell* ≤ 0.05 betetzen dute. Beraz, egotzitako aldagaiek banaketa marjinalak mantentzen dituztela egiaztatzen da.

- **Egotzitako aldagaiek lotura-aldagaiekin duten baterako banaketa mantentzen ote duten egiaztatu, datu-base *emailea* erreferentzia gisa hartuta.**

Propietate hau CIES eta IKUSIKERren arteko lotura-aldagaien banaketa marjinalen komunztaduraren mende dago erabat. Alde batetik, aipatu behar da egotzitako aldagaiek estratifikazio-aldagai komunekin batera mantentzen dutela baterako banaketa —hau da, *EDA* (bi urteko taldeetan), *SEX* eta *PROV* aldagaiekin—. Izan ere, metodologia honek ziurtatzen du dohaintza-klase bereko indibiduoen artean soilik transferitzen dela informazioa. Bestalde, lotura-aldagai batzuen banaketa marjinalak ez datozenez guztiz bat, propietate hori ez da guztiz betetzen.

Jarraian, egotzitako aldagaien azpimultzo baten eta lotura-aldagaien arteko baterako banaketetan lortutako emaitzak ageri dira (gainerako erantzunak oso antzekoak dira):

Como se puede observar para todas las variables *tvd* ≤0.06, *overlap* ≥ 0.94 y *hell* ≤0.05. Por lo tanto, se verifica que las variables imputadas conservan las distribuciones marginales.

- **Analizar si las variables imputadas conservan la distribución conjunta con las variables de unión, en referencia a la base de datos *donante*.**

Esta propiedad, depende completamente de la concordancia de las distribuciones marginales de las variables de unión de CIES e IKUSIKER. Por un lado, cabe mencionar que las variables imputadas mantienen la distribución conjunta con las variables comunes de estratificación, es decir, *EDA* (en grupos de dos años), *SEX* y *PROV,* ya que esta metodología asegura que solo se transfiere información entre individuos de la misma clase de donación. Por otro lado, como las distribuciones marginales de algunas de las variables de unión no son totalmente concordantes, esto hace que esta propiedad no se cumpla al completo.

En las siguientes tablas aparecen los resultados obtenidos para las distribuciones conjuntas entre un subconjunto de variables imputadas y las variables de unión, siendo el resto de repuestas muy similares a las que se muestran:

EDA lotura-aldagaia: *Variable de unión: EDA*

Aldagaia / Variable	***tvd***	***overlap***	***Bhatt***	***Hell***
InstagramG1	0.31346	0.68654	0.73872	0.51115
InstagramG2	0.28853	0.71147	0.93547	0.25402
InstagramG3	0.28965	0.71035	0.93022	0.26416
InstagramG3_Beste	0.30634	0.69336	0.77943	0.46965
InstagramG4	0.31132	0.68868	0.81176	0.43387
InstagramG5	0.28932	0.71068	0.93699	0.25101
InstagramG5_Beste	0.46172	0.53828	0.58309	0.64568
InstagramG6	0.28815	0.71185	0.94075	0.24341

HABI lotura-aldagaia: *Variable de unión: HABI*

Aldagaia / Variable	***tvd***	***overlap***	***Bhatt***	***Hell***
InstagramG1	0.39108	0.60892	0.67849	0.56702
InstagramG2	0.11095	0.88905	0.98433	0.12517
InstagramG3	0.12591	0.87409	0.97756	0.14979
InstagramG3_Beste	0.44212	0.55788	0.68027	0.56545
InstagramG4	0.29889	0.70111	0.81544	0.42961
InstagramG5	0.09958	0.90042	0.99017	0.09916
InstagramG5_Beste	0.32286	0.67714	0.75065	0.49934
InstagramG6	0.07527	0.92573	0.9926	0.08602

POS161A lotura-aldagaia: *Variable de unión: POS161A*

Aldagaia / Variable	***tvd***	***overlap***	***Bhatt***	***Hell***
InstagramG1	0.30537	0.69463	0.75749	0.49245
InstagramG2	0.27555	0.72445	0.92527	0.27337
InstagramG3	0.27919	0.72081	0.90773	0.30376
InstagramG3_Otro	0.28779	0.712	0.79248	0.45554
InstagramG4	0.29497	0.70503	0.82455	0.41887
InstagramG5	0.27403	0.72597	0.93332	0.25822
InstagramG5_Otro	0.24155	0.75845	0.8031	0.44373
InstagramG6	0.27415	0.72585	0.93741	0.25019

SOZ lotura-aldagaia: *Variable de unión: SOZ*

Aldagaia / Variable	***tvd***	***overlap***	***Bhatt***	***Hell***
InstagramG1	0.37787	0.62213	0.67294	0.57189
InstagramG2	0.31622	0.68378	0.90219	0.31274
InstagramG3	0.33151	0.66849	0.88991	0.33181
InstagramG3_Beste	0.47272	0.52728	0.59616	0.63549
InstagramG4	0.36634	0.63366	0.74971	0.50026
InstagramG5	0.32599	0.674	0.91384	0.29353
InstagramG5_Beste	0.39554	0.60446	0.64809	0.59322
InstagramG6	0.31355	0.68645	0.92255	0.27829

KONTS lotura-aldagaia: *Variable de unión: KONTS*

Aldagaia / Variable	***tvd***	***overlap***	***Bhatt***	***Hell***
InstagramG1	0.06408	0.93592	0.95391	0.21469
InstagramG2	0.02486	0.97514	0.99833	0.0408
InstagramG3	0.02835	0.97165	0.99532	0.06843
InstagramG3_Beste	0.04268	0.95732	0.97592	0.15519
InstagramG4	0.04819	0.9518	0.97837	0.14708
InstagramG5	0.0343	0.96569	0.99609	0.06252
InstagramG5_Beste	0.02061	0.97939	0.99934	0.02566
InstagramG6	0.10896	0.89104	0.93717	0.25066

Ikus daitekeenez, egotzitako aldagai gehienek *KONTS* lotura-aldagaiarekin baterako banaketa mantentzen dute, espero zitekeen bezala, aldagai horrek oso antzeko banaketa marjinalak baititu CIES eta IKUSIKER datu-baseetan. Gainerako aldagaiei dagokienez, aniztasuna dago. Adibidez, *HABI* aldagaiak nahiko ongi mantentzen du baterako banaketa *InstagramG2*, *InstagramG3*, *InstagramG5* eta *InstagramG6* aldagaiekin, baina ez

Como se puede observar, la mayoría de las variables imputadas mantienen su distribución conjunta con la variable de unión *KONTS*, como era de esperar ya que esta variable tiene distribuciones marginales muy similares en las bases de datos de CIES y de IKUSIKER. En cuanto a las demás variables, se observa diversidad. Por ejemplo, la variable *HABI* mantiene bastante bien la distribución conjunta con variables como *InstagramG2*, *Ins-*

InstagramG1 eta *InstagramG3_Beste* aldagaiekin. *EDA*, *POS161A* eta *SOZ* aldagaiek, ordea, ez dute oso ondo mantentzen baterako banaketa egotzitako aldagaiekin. Ondorioz, ezin da egiaztatu baliozkotze propietate hau betetzen denik; eta, beraz, hori kontuan hartu behar da aldagai horiek aztertu eta gurutzatzean. Hurrengo fusioei begira, ezinbestekoa izango da banaketa marjinal konkordanteak dituzten lotura-aldagaiak aurkitzea. Horretarako, funtsezkoa izango da aldagai komun gehiago dituzten datu-baseetatik abiatzea. Hori dela eta, CIESen eta IKUSIKER panelean lotura-aldagai gisa erabili ahalko diren aldagai berriak sartzeko proposamenak egin dira jadanik.

tagramG3, *InstagramG5* e *InstagramG6*, pero no lo hace en absoluto con variables como *InstagramG1* e *InstagramG3_Beste*. Sin embargo, las variables *EDA*, *POS161A* y *SOZ* no mantienen especialmente bien la distribución conjunta con las variables imputadas. Por lo tanto, no se puede verificar que se cumpla esta propiedad, por lo que hay que tener esto en cuenta a la hora de hacer el análisis y cruzar estas variables. De cara a próximas fusiones será imprescindible encontrar variables de unión que tengan distribuciones marginales concordantes y para ello será fundamental partir de bases de datos con más variables en común. Por ello, se han hecho ya propuestas tanto en CIES como en IKUSIKER para introducir nuevas variables que puedan ser utilizadas posteriormente como variables de unión.

5.2.2. Pertsona adituen baliozkotzea

Proiektu hau, parte-hartzaileentzat guztiz berriak diren metodologiak lantzen dituen heinean, proiektu pilotua izan da. Hori dela eta, jada adierazia izan den gisan, zenbait erakunderen laguntza jaso da —datuak fusionatzeko lehen urratsak egiten hasiak ziren hala nola AIMC eta EUSTAT—. Proiektua aurrera ateratzeko oso lagungarria izan zen informazioa partekatu zutenez, erakunde horiekin harremanetan jartzea erabaki zen fusioa egin ondoren, proposamen me-

5.2.2. Validación de personas expertas

Este proyecto, al tratar metodologías completamente nuevas para el equipo participante, ha sido un proyecto piloto. Por ello, se ha recabado ayuda de varias entidades que ya habían empezado a dar los primeros pasos en fusión de datos, como AIMC y EUSTAT, aportando información que ha sido de gran ayuda para poder sacar el proyecto adelante. Por esta razón, tras realizar la fusión, se decidió contactar con estas entidades para explicar la propuesta me-

todologikoa azaltzeko eta haren baliozkotasuna egiaztatzeko.

Alde batetik, AIMCko Jose Andrés Gabardorekin, EGMren datu-baseen fusioetan lan egin duen matematikariarekin, bideokonferentzia bat egin zen uztailean. Gabardok sortutako metodologia baliozkoa eta egokia dela baieztatu zuen, eta, gainera, arreta berezia jarri zuen metodologian optimizazio problema bat erabiltzeari, onartu baitzuen aurretik ez zuela inoiz ikusi datu-fusioan. Bestalde, urrian bilera presentziala egin zen EUSTATeko bi matematikarirekin, Marina Ayestaran eta Marta Mas. Horiek ere berretsi egin zuten fusioaren baliotasuna eta fusio-kasu honetarako egokitasuna.

todológica y contrastar la validez de la misma.

Por un lado, en julio se realizó una videoconferencia con el matemático de AIMC José Andrés Gabardo, quien trabaja con fusiones de bases de datos de EGM. Gabardo confirmó que la metodología creada es válida y adecuada y además mostró especial interés en el uso de un problema de optimización en la metodología, ya que reconoció no haberlo visto nunca antes en una fusión de datos. Por otro lado, en octubre se hizo una reunión presencial con dos matemáticas de EUSTAT, Marina Ayestaran y Marta Mas, que también corroboraron la validez de la fusión y su adecuación para un caso de fusión como este, en el que el número de variables comunes entre las bases de datos es muy reducido.

Hirugarren atala: emaitzen analisia

Liburuaren hasieran azaldu den moduan, nagusiki euskarazko komunikazio praktikak hobeto ezagutzeak ekarri gaitu proiektu honetara. Horrek baldintzatu du fusionatu diren datu-baseen aukeraketa, eta horrek gidatu du analisia ere.

Bestalde, jadanik azaldu dugu, 2.4 atalean, lanaren mugak zein diren. Muga nagusia erabilitako jatorrizko datuen izaeratik dator. Egindako balidazioak erakusten du fusioa, bere horretan, egokia izan dela. Alabaina, beste zerbait ere erakutsi du: datu sendoak lortzeko, fusionatuko diren artxiboek ezaugarri jakin batzuk izatea komeni da. Gurea proiektu pilotu bat izanik, eta fusionatu diren artxiboen ezaugarriengatik, lortu diren emaitza zehatzak tentuz hartu beharra dago. Erabilitako datu-baseetako bat (IKUSIKER panela) garatzen ari den tresna bat da. Fusioa hala ere balidatua izan da, baina datuak interpretatzeko unean zuhur jokatzera behartuta gaude.

Hori dela medio, oso analisi mugatua aurkezten da hemen, fusioak ahalbidetzen duena erakusteko asmoarekin eta ez komunikazio praktiken analisi sakon bat egiteko helburuarekin. Xede hori hurren-

Capítulo tercero: análisis de los resultados.

Tal y como se ha explicado al comienzo del libro, es la búsqueda de un mejor conocimiento de las prácticas de comunicación en euskera lo que ha motivado este proyecto. Ello ha condicionado tanto la elección de las bases de datos fusionadas como el análisis.

Por otro lado, ya se han señalado en el apartado 2.4 los límites de este trabajo. La limitación principal proviene de la naturaleza de los datos originales utilizados. La validación realizada muestra que la fusión ha sido, en sí misma, adecuada. Pero también ha demostrado que para obtener datos sólidos conviene que los archivos a fusionar tengan unas características determinadas. Siendo este un proyecto piloto, y por las características de los archivos que se han fusionado, es precisa la cautela a la hora de analizar los resultados concretos obtenidos de la fusión. Concretamente una de las bases de datos utilizadas (panel IKUSIKER) es una herramienta aun en desarrollo. La fusión ha sido validada, pero estamos obligados a ser prudentes en el momento de interpretar los datos.

Por ello, aquí se presenta un análisis muy limitado, con la intención de mostrar lo que permite la fusión y no con el de hacer un análisis en profundidad de las prác-

go edizio batzuetarako utziko dugu.

6. Komunikabideen kontsumoaren marko orokorra Araba, Bizkaia, Gipuzkoa eta Nafarroan

Azken hamar urtean gertatu diren aldaketak modu zehatzagoan ikusteko datu interesgarriak ematen dizkigute jarraian datozen 2. eta 3. grafikoek.

Duela hamar urte teknologia berriei lotutako euskarrien kontsumoan sortzen ziren alde handienak adin taldeen artean: Internet, oro har, eta prentsa digitala. Euskarri tradizionalak berdinzaleagoak ziren ordea, haien arteko alde nabarmenekin bada ere (telebista oso berdinzalea eta egunkariak ez hainbeste). Gaur egun iraulita agertzen da panorama: Interneteko erabileretan berdintzen dira adin talde gehienak (nagusienak kenduta), eta euskarri tradizionaletan bereizten. Hauetan joera argia dago: zenbat eta gazteagoa izan, orduan eta gutxiago kontsumitu. Salbuespena prentsa digitala da: gailurra tarteko adin taldeetan hartzen du kontsumoak (35-44 urte), gazteen zein nagusienen artean asko jaitsiz. Bera ere tarteko medioa dela esan genezake (egitura tradizionala baina euskarri berria); edonola ere, joerari erreparatuta adierazi daiteke 2011n publikoa gazteagoa zela, eta orain adin ertaineko jendearen artean kontsumitzen dela gehien. Gazteenak, aldiz, gero

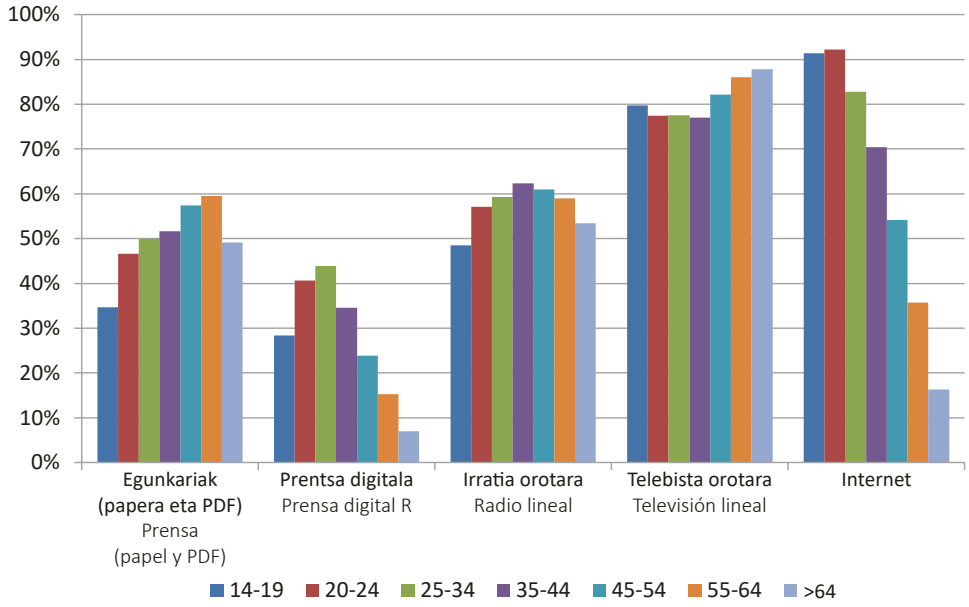
ticas comunicativas. Eso quedará para próximas ediciones.

6. El marco general del consumo de medios en Araba, Bizkaia, Gipuzkoa y Nafarroa

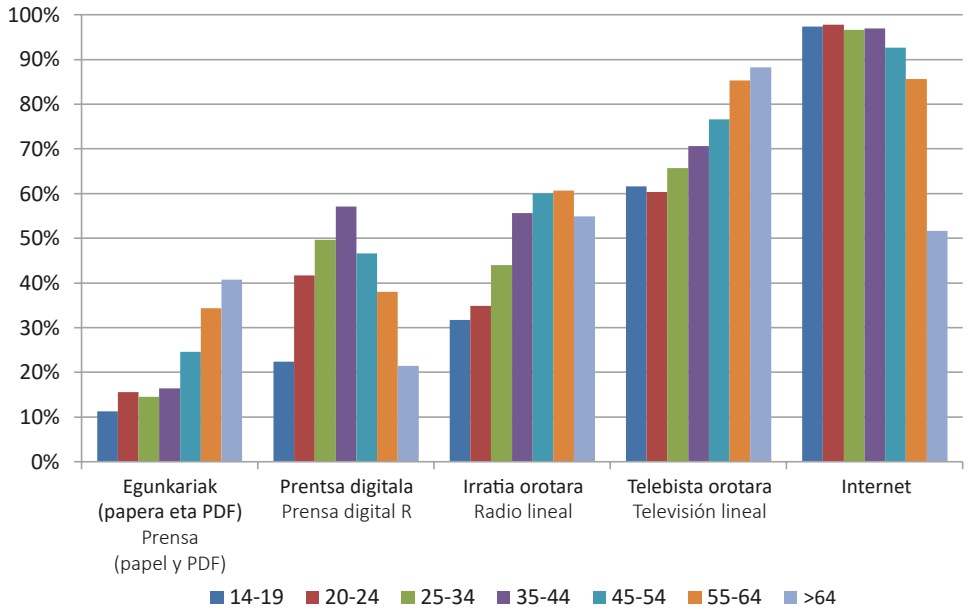
Los gráficos 2 y 3 ofrecen datos que ilustran algunos cambios producidos en la última década.

Hace diez años las mayores diferencias entre los grupos de edad se observaban en el consumo de soportes ligados a las nuevas tecnologías: internet en general, y prensa digital. Los soportes tradicionales eran, por el contrario, más igualitarios; aunque con notables diferencias entre ellos (la televisión más igualitaria y la prensa menos). El panorama se presenta hoy muy diferente: la mayoría de los grupos de edad (excluidos los mayores) se igualan en los accesos a Internet, y se diferencian en relación a los soportes tradicionales. En estos aparece claramente una tendencia: cuanto más joven se es, menos consumo se hace. La excepción es la prensa digital, donde el consumo alcanza su punto álgido en los grupos de edad intermedios (35-44 años), descendiendo notablemente tanto entre las personas más jóvenes como entre las más mayores. También se puede decir que es un medio intermedio desde el punto de vista técnico (estructura tradicional de prensa pero soporte tecnológico nuevo), pero la tendencia muestra un desplazamiento de las edades de

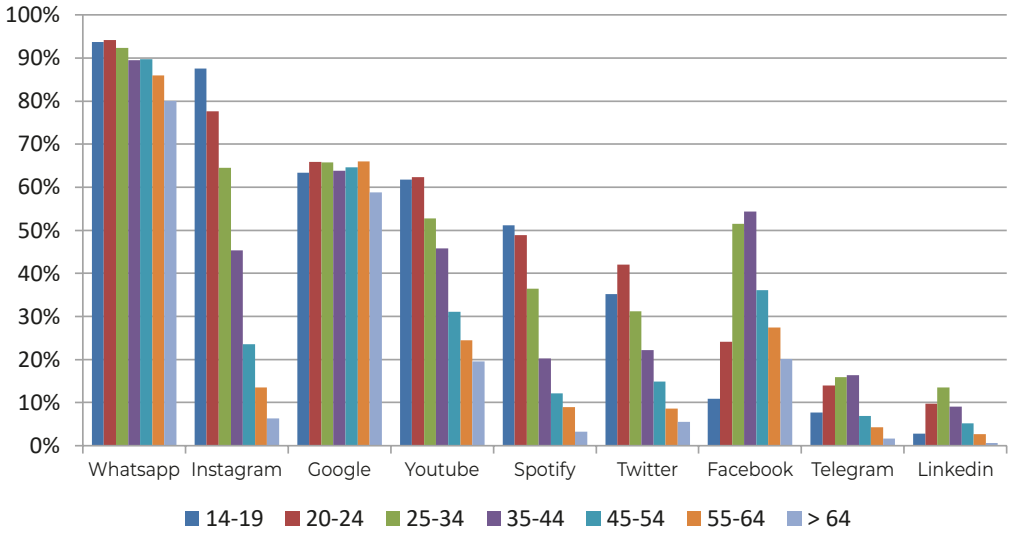
2. grafikoa: Komunikabide tradizionalen eta interneten eguneroko kontsumoak (CIES, 2011) / Gráfico 2: Consumo diario de medios tradicionales e internet (CIES, 2011)



3. grafikoa: Komunikabide tradizionalen eta interneten kontsumoa bezperan (CIES, 2021) / Gráfico 3: Consumo diario de medios tradicionales e internet en la víspera (CIES, 2021)



4. grafikoa: Interneteko erabilera bezperan (CIES, 2021) / Gráfico 4: usos de internet en la víspera (CIES, 2021)



eta gutxiago sartzen dira euskarri horretara.

Interneteko erabilera batzuetara bagatoz, hemen ere ikusten da euskarri batzuk modu berdintsuagoan zabaltzen direla adin talde gehientsuenetan, beste batzuk oso diskriminatzaileak diren bitartean.

Whatsappda, 4. grafikoan ikusten denez, adin talde guztietan gehien erabiltzen den sareko erreferentzia. Google bilatzailea ere oso berdinzalea da zentzu berean. Gainerako sare sozialetan, berriz, oso alde handiak agertzen dira taldeen artean: Youtube, Spotify eta, bereziki, Instagram, gazteenen sareak dira ezbairik gabe[6]. Twitterren 20-24 urteetako

mayor consumo desde 2011 hasta ahora, siendo la población de mediana edad la que más consume. Por el contrario, las personas más jóvenes cada vez acceden menos a este soporte.

Si nos fijamos en algunos de los usos diarios de internet (Gráfico 4), vemos que también hay diferencias significativas por grupos de edad. Whatsapp es la plataforma en la que con mayor igualdad confluyen todas las generaciones. En la misma línea se encuentra el buscador de Google. El resto de plataformas aparece muy marcada por la edad: Youtube, Spotify y, sobre todo, Instagram son los medios de la gente más joven.[6] Twitter tiene mayor presencia en el grupo de 20 a 24 años, y

6 Ez ditugu oraindik CIESen Twitch, TikTok eta beste sare batzuen datuak.

6 No disponemos aún de datos de Twitch y de TikTok en las encuestas de CIES.

gazteak nabarmentzen dira, eta Facebook, Telegram zein Linkedinen erabiltzaileen batez besteko adinak gora egiten du.

7. *Online* eta *offline* kontsumoak 14 eta 19 urte bitarteko populazioan

Artxiboen fusiotik ateratako datuei helduz, medio tradizionalei lotutako *offline* kontsumoen eta sareko euskarriei lotutako *online* kontsumoen arteko erlazioa azter dezakegu.

IKUSIKER panela batez ere —baina ez esklusiboki— ikus-entzunezko edukien kontsumoan oinarritzen denez, eduki horien kontsumoak telebista orokorraren kontsumoarekin alderatzea erabaki dugu. Lehenengo datua IKUSIKER panelean biltzen da, bigarrena CIESen txostenetan.

Hurrengo hiru tauletan, sareko ikus-entzunezkoen plataforma ezberdinetan (Youtube, Twitch eta Instagram) egunero emandako denbora eta telebista jeneralista bezperan ikusi izanaren datuak alderatzen dira, populazioaren datu estimatuetan. Lehen datua IKUSIKERen panelekoa da, eta bigarrena CIESen inkestakoa.

Facebook, Telegram y Linkedin la tienen en grupos de edades más altas.

7. Consumo *offline* y consumo *online* entre la población de 14 a 19 años

Entrando ya a los datos extraídos de la fusión de archivos, podemos analizar la relación existente entre los consumos offline, ligado a los medios tradicionales, y los consumos online, ligado a los soportes en red.

Dado que el panel IKUSIKER se centra sobre todo –aunque no exclusivamente- en el consumo de contenidos audiovisuales, hemos optado por comparar los consumos de estos contenidos con el consumo de televisión generalista. El primero de los datos aparece en el panel de IKUSIKER, el segundo en el Estudio de Audiencia de Medios de CIES.

Las tres tablas que siguen comparan el tiempo diario dedicado a las diferentes plataformas audiovisuales de la red (Youtube, Twitch e Instagram) con los datos de haber visto la televisión generalista la víspera, en datos estimados de población. El primer dato proviene del panel IKUSIKER, el segundo de la encuesta de CIES.

10. taula: Youtube ikusten egunero emandako denbora (fusionatutako artxiboa, 2021). / Tabla 10: Tiempo diario dedicado al consumo de contenidos en Youtube (archivo fusionado, 2021)

		<30 min	30-60 min	60-90 min	90-120 min	2-3 ordu/ horas	3-4 ordu/ horas	>4 ordu/ horas
Telebista jeneralistaren ohiko kontsumoa/ Consumo diario de televisión generalista	**Ez/No**	9.338	14.605	6.756	2.025	891	541	65
	Bai/Sí	6.049	15.479	4.977	1.375	927	554	46

11. taula: Twitch ikusten egunero emandako denbora (fusionatutako artxiboa, 2021). / Tabla 11: Tiempo diario dedicado al consumo de contenidos en Twitch (archivo fusionado, 2021)

		<30 min	30-60 min	60-90 min	90-120 min	2-3 ordu/ horas	3-4 ordu/ horas	>4 ordu/ horas
Telebista jeneralistaren ohiko kontsumoa/ Consumo diario de televisión generalista	**Ez/No**	5.766	3.732	1.311	393	597	186	221
	Bai/Sí	4.615	4.334	1.337	472	289	29	0

12. taula: Instagram ikusten egunero emandako denbora (fusionatutako artxiboa, 2021). / Tabla 12: Tiempo diario dedicado al consumo de contenidos en Instagram (archivo fusionado, 2021)

		< 15 min	15-30 min	30-45 min	45-60 min	60-90 min	90-120 min	2-3 ordu/ horas	3-4 ordu/ horas	>4 ordu/ horas
Telebista jeneralistaren ohiko kontsumoa/ Consumo diario de televisión generalista	**Ez/ No**	5.534	8.999	9.975	14.297	9.507	13.205	5.833	2.029	1.778
	Bai/ Sí	3.719	7.412	10.419	10.569	11.519	8.345	5.146	1.610	1.233

Taula hauei Chi Karratuaren testa aplikatu zaie, eta hiruetan agertzen da telebista jeneralista ikusi izanaren eta aztertutako plataformetan emandako denboraren arteko harreman esanguratsua. Halaber, erregresio logistiko sinplearen eredu bat sortu dugu, harreman horren norabidea ikusi eta kuantifikatzeko, eta honakoa da emaitza: oro har, telebista jeneralista ikusten duten gazteen artean joera

Hemos aplicado el test de Chi Cuadrado a los datos, y encontramos una relación entre el consumo de televisión generalista y el tiempo dedicado a cada de una de las plataformas analizadas. No se trata pues de variables independientes, tomadas de dos en dos (consumo diario de televisión y tiempo dedicado a cada plataforma). A continuación, para poder cuantificar la relación entre estas variables hemos creado

handiagoa dago Youtuben, Twitchen zein Instagramen egunero denbora gehiago emateko. Harremana ez da erabat lineala, eta bereziki sareko kontsumoa oso altua denean emaitza aldatu egiten dela dirudi.

Datu hauek analisi sakonagoa merezi dute. Hala ere, hipotesi interesgarri bat planteatzeko oinarria ematen digute fusiotik eratorritako datuek: sareko ikusentzunezkoen kontsumoek ez dute ohiko telebista kontsumoa ordezkatzen; pilatu egiten zaizkio. Badirudi gazte batzuek ikusentzunezko gehiago kontsumitzen dituztela, sarean zein ohiko modu linealean, eta beste batzuek gutxiago.

Hipotesi hau aztertu beharko da sakontasun gehiagorekin fusioaren hurrengo edizioetan. Modu berean, fusioak ahalbidetuko digu gazteen offline eta online kontsumo perfilak identifikatu ahal izatea.

un modelo de regresión logística simple. El resultado es que las personas jóvenes que ven diariamente televisión generalista tienen mayor probabilidad de pasar más tiempo consumiendo contenidos de Youtube, Twitch o Instagram, que aquellas que no consumen televisión. Esta relación no es lineal, y se modifica en el caso de los consumos muy prolongados de las plataformas en red.

Estos datos precisan de un análisis más profundo. Sin embargo, las evidencias que ya disponemos permiten avanzar la hipótesis de que existe una correlación positiva entre el consumo diario de televisión generalista y el consumo de contenidos audiovisuales en red. No se trataría pues de una mera sustitución de unas prácticas por otras.

Esta hipótesis habrá de ser analizada en mayor profundidad en las próximas ediciones de la fusión de datos. Asimismo, la fusión nos permitirá identificar perfiles de jóvenes consumidores en base a su comportamiento tanto offline como online.

8. 14-19 urte bitarteko populazioaren euskarazko kontsumo patroiak

Urteetan zehar CIESen datuen inguruan egindako azterketek, euskarazko komunikabideen erabilerari lotuta ageri diren aldagai batzuk identifikatzea ahalbidetu digute. Horrela,

8. Los patrones de consumo en euskera entre la población de 14 a 19 años

El análisis de los datos de CIES que hemos realizado durante años nos ha permitido encontrar algunas variables significativas a la hora de explicar el mayor o menor consumo de medios en

euskararen ezagutzak zeresan handia duela dioen (eta sarritan gehiegi erabiltzen den) hipotesiaren aldean, gaur egun badakigu, jakin, hizkuntza profila ezagutza soila baino askoz aldagai garrantzitsuagoa dela. Hizkuntza profiltzat zera hartzen dugu: Eusko Jaurlaritzako Hizkuntza Politikarako sailburuordetzak zein Soziolinguistika Klusterrak beren lanetan erabili ohi duten BILA indizeak ezartzen duena. Aldagai hau oso egokia da hizkuntzaren berreskurapenaren testuinguruan, hizkuntza mugikortasuna neurtzen baitu: haurtzaroan euskararen erabateko ezagutzagabetik ezagutzara doan bidea, edota alderantzizkoa, ezagutzatik galeraraino doana. Aldagaiaren kategoria ezberdinak eraikitzeko beste bi aldagairen gurutzaketa egiten da: hizkuntza gaitasuna (ezagutza) eta eskuratutako lehen hizkuntza. Bi aldagai horien gurutzaketatik zazpi kategoria eratortzen dira:

- **Euskaldun zaharrak**: euskaraz ondo mintzatu eta lehen hizkuntza euskara soilik izan dutenak.

- **Jatorrizko elebidunak**: euskaraz ondo mintzatu eta lehen hizkuntzatzat euskara eta beste bat, aldi berean, izan dituztenak.

- **Euskaldun berriak**: lehen hizkuntzen artean euskara

euskera. Así, frente a la hipótesis básica (y a la que en ocasiones se recurre en exceso) de que el conocimiento de la lengua es un factor determinante a la hora de explicar el consumo de dichos medios, hoy en día sabemos que el perfil lingüístico es una variable más precisa que el conocimiento en sí; entendiendo por perfil lingüístico aquel que se recoge en la variable BILA, utilizada tanto por la Viceconsejería de Política Lingüística del Gobierno Vasco como por el Clúster de Sociolingüística en sus respectivos análisis. Esta variable, muy apropiada en un contexto de recuperación de la lengua, mide la movilidad lingüística: el tránsito del desconocimiento inicial del euskera en la primera infancia hacia el conocimiento posterior; y el tránsito inverso, desde el conocimiento inicial como primera lengua hacia su pérdida. Para su construcción se realiza un cruce de las variables de competencia lingüística y de primera lengua, de la que se derivan siete categorías:

- **Euskaldun zaharrak:** personas que hablan bien euskera y cuya primera lengua fue solo el euskera.

- **Jatorrizko elebidunak**: personas que hablan bien euskera y cuya primera lengua fue el euskera junto con otra.

- **Euskaldun berriak**: personas que hablan bien euskera y

izan ez, eta gaur egun ondo mintzatzen direnak.

- **Partzialki euskaldun berriak**: lehen hizkuntzen artean euskara izan ez, eta gaur egun elebidun hartzaileak direnak.

- **Partzialki erdaldunduak**: lehen hizkuntza euskara izanik gaur egun elebidun hartzaileak direnak.

- **Guztiz erdaldunduak**: lehen hizkuntza euskara izanik gaur egun euskara ulertzen ez dutenak.

- **Erdaldun zaharrak**: lehen hizkuntzen artean euskara izan ez, eta gaur egun ulertu ere egiten ez dutenak.

Azken urteetan CIESen eragindako aldaketei esker aldagai hau berregitea posible izan da. Horren arabera, euskaraz hitz egiteko gaitasuna duten pertsonen artean ere oso jokabide ezberdinak aurkitzen ditugu euskarazko medioen kontsumoari dagokionez. Horrela, gazteen kasuan —gainerako adin taldeen kasuan bezalatsu— euskara lehen hizkuntza soila dutenen kontsumo maila gainerako euskaldunek dutenaren oso gainetik dago: bereziki euskara bigarren hizkuntza duten euskaldun berrien aldean. Horiek, kontsumoari dagokionez, hurbilago daude erdaldun zacuya primera lengua no fue el euskera.

- **Partzialki euskaldun berriak**: personas que no hablan bien euskera pero lo entienden, y cuya primera lengua no fue el euskera.

- **Partzialki erdaldunduak**: personas que tuvieron el euskera como primera lengua y ahora lo entienden, pero no lo hablan bien.

- **Guztiz erdaldunduak**: personas que tuvieron el euskera como primera lengua y ahora no lo entienden.

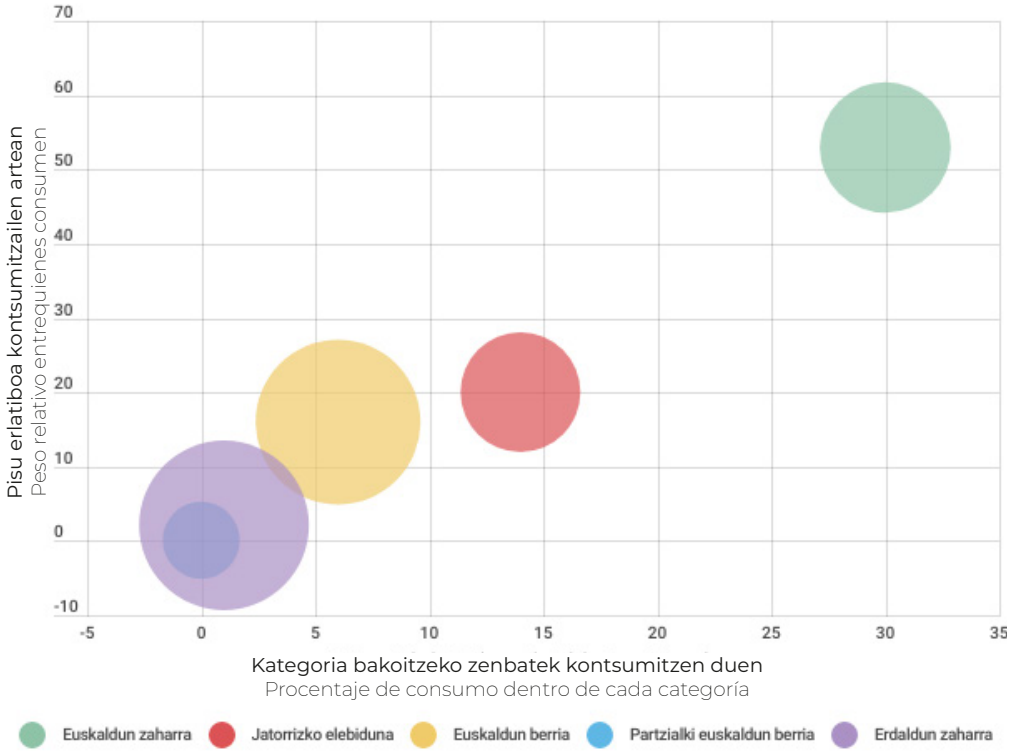
- **Erdaldun zaharrak**: personas que no entienden euskera cuya primera lengua no fue el euskera.

La modificación del cuestionario de CIES en los últimos años nos permite construir ahora esta variable, según la cual las personas que hablan bien el euskera tienen un comportamiento muy diferente en cuanto al consumo de contenidos en euskera, dependiendo de su perfil de BILA. Así en el caso de las y los jóvenes (al igual que en el resto de grupos de edad), quienes tienen el euskera como primera lengua muestran un nivel de consumo muy superior a aquellas personas vascohablantes que lo tienen como segunda lengua. Estas personas, *euskaldun berriak*, se encuentran, en cuanto a consumo de medios tradiciona-

harrengandik euskaldun zaharrengandik baino. Horixe ikusten dugu 5. grafikoan.

les en euskera, mucho más cerca de quienes no hablan euskera que de las y los *euskaldun zaharrak*. Es lo que muestran los datos del Gráfico 5.

5. grafikoa: Egunero euskarazko komunikabide tradizionalen bat kontsumitzen duten gazteak (CIES, 2021). / Gráfico 5: Jóvenes que consumen diariamente algún medio tradicional en euskera (CIES, 2021)



Ardatz horizontalak zera erakusten du: kategoria bakoitzaren barnean, beti ere 14-19 urte bitarteko populazioan, ehuneko zenbatek kontsumitzen duen egunero euskarazko komunikabideren bat: euskaldun zaharren artean % 30 dira. Portzentaje hori % 6ra jaisten da euskaldun berrien artean, horiek erdaldun zaharrengandik askoz

El eje horizontal muestra el porcentaje, dentro de cada categoría, de personas de entre 14 y 19 años que consumen diariamente algún medio tradicional en euskera: el 30 % de las y los jóvenes *euskaldun zaharrak* lo hace. Este porcentaje baja al 6 % en el caso de las y los *euskaldun berriak*, más cercanos, en cuanto a consumo, de las y los castella-

hurbilago daudelarik. Bi lehen hizkuntza dituzten pertsonak bestalde (haietako bat euskara izanik, jatorrizko elebidunak alegia) beste euskaldunen bi taldeen artean daude[7]. Ondorioz, euskaraz egunero zerbait kontsumitzen duten pertsonak euskaldun zaharrak direla esan ahal dugu, ardatz bertikalak erakusten duen moduan.

IKUSIKER panelak ez ditu, orain arte, CIESek hizkuntza ezagutzari buruz eta lehen hizkuntzari buruz dituen galderak. Beraz ezin dugu zuzenean panel horretatik daturik atera. Alabaina, inkesta batetik besterako datuak estrapolatzea ahalbidetzen digu fusio prozesuak, horrela grafikoa berreraikiz sareko kontsumoei dagokienez. Hori da 6. grafikoan erakusten dena.

Grafikoa aurrekoaren ezberdina da, honakoan jatorrizko elebidunak baitira sarean euskaraz gehien kontsumitzen dutenak. Datu esanguratsua izan liteke. Izan ere, euskaldun zaharren eta jatorrizko elebidunen artean — lehen hizkuntza euskara soilik, ala euskara eta beste bat izateaz gain— habitatari dagokion baldintzapena ere aintzat hartzekoa izan daitekeelako. EUSTA-

nohablantes monolingües. El grupo de personas con dos lenguas maternas (una de ellas el euskera) se sitúa a caballo entre los otros dos grupos de vascohablantes.[7] Como consecuencia, más de la mitad de las personas que consumen diariamente son *euskaldun zaharrak*, tal y como muestra el eje vertical.

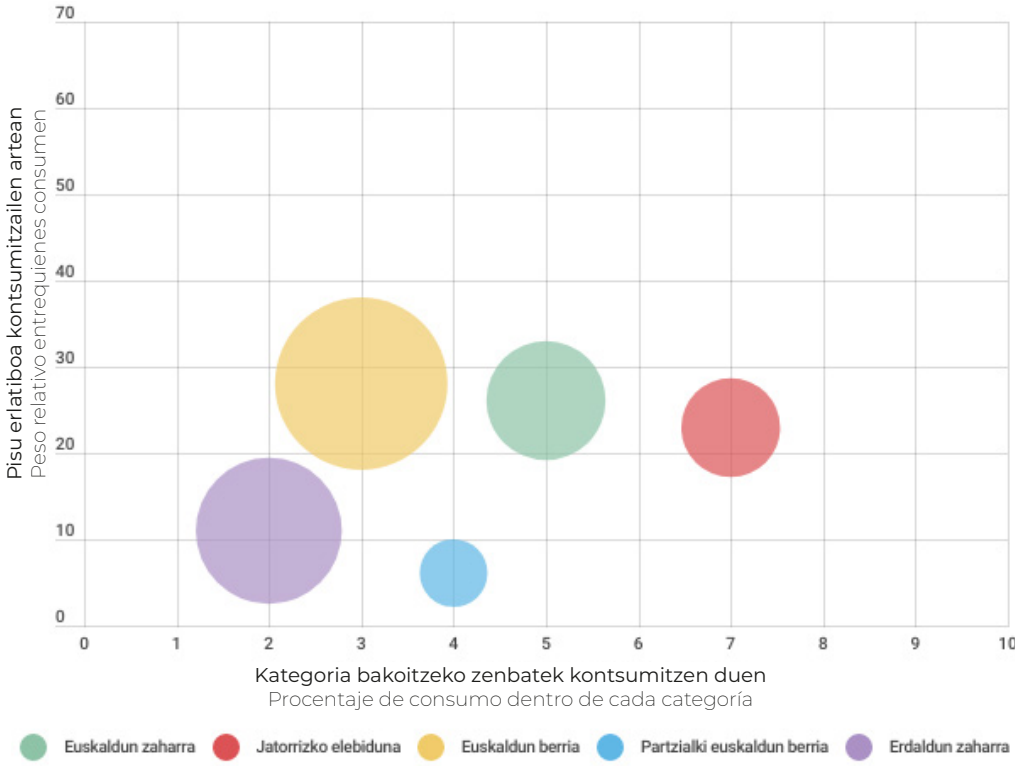
El panel IKUSIKER no ha incluido, hasta ahora, las mismas variables sobre competencia lingüística y primera lengua que la encuesta de CIES, por lo que no podemos extraer directamente los datos del mismo. Sin embargo el proceso de fusión nos permite extrapolar los datos de una encuesta a otra, con lo que es posible reconstruir el gráfico para observar los patrones de consumo en euskera en soportes en red. Es lo que se muestra en el Gráfico 6.

El gráfico es diferente al anterior, puesto que en lo relativo al consumo de medios en red son las y los jóvenes nativos bilingües quienes muestran un mayor consumo en euskera. Podría ser un dato importante, puesto que otra diferencia entre euskaldun zaharrak y jatorrizko elebidunak, además de que unos tienen solo el euskera como primera lengua y otros tie-

7 Burbuilen tamainak kategoria bakoitzeko pertsonen kopuru absolutua erakusten du. Partzialki erdaldunduen eta guztiz erdaldunduen kategoriak kendu ditugu, hots, euskara lehen hizkuntza izan arren asko edo guztiz galdu dutenen kategoriak, 14-19 urte bitarteko populazioan oso talde txikia direlako.

7 El tamaño de las burbujas indica el número absoluto de personas en cada categoría. Hemos suprimido las categorías de *partzialki erdaldunduak y guztiz erdaldunduak*, es decir, personas que tuvieron el euskera como su o entre sus lenguas maternas pero que lo han olvidado parcial o totalmente, ya que constituyen, entre la población de 14 a 19 años, dos grupos numéricamente muy reducidos.

6. grafikoa: Egunero euskarazko sareko komunikabideren bat kontsumitzen duten gazteak (fusionatutako artxiboa, 2021). / Gráfico 6: Jóvenes que consumen diariamente algún medio en red en euskera (Archivo fusionado, 2021)



Ten datuen arabera[8], lehenak nagusiki euskaldunen portzentaje handia duten udalerrietan bizi dira (% 68 euskaldunen portzentajea populazioaren erditik gorakoa duten udalerrietan), eta bigarrenak eremu urbanoagoetan eta ez hain euskaldunetan bizi dira (% 85 populazioaren erdia baino gutxiago euskalduna duten udalerrietan).

Hala ere, kontuan hartzekoa da kontsumo baxuko datuekin lan egiten ari garela, eta honen dos primeras lenguas, es que los primeros habitan en su mayoría, según datos del EUSTAT,[8] en zonas con alto porcentaje de vascohablantes (68 % en municipios con más de la mitad de población vascohablante), mientras que los segundos habitan más en zonas más urbanas y con menor presencia del euskera (85 % en zonas con menos de la mitad de población vascohablante).

Sin embargo hemos de tener en cuenta estamos tratando de

8 2016ko populazio erroldatik ateratako datuak dira, Nafarroa gabekoak.

8 Son datos extraídos del Censo de Población de 2016 y no incluyen a Nafarroa.

rren ondorioz taldeen artean agertzen diren aldeak esangura mugatukoak izan litezkeela. Datuak, beraz, tentu handiz irakurtzekoak dira, haiek osa ditzaketen informazio zehatzagoak izan bitartean.

Bestalde, komunikabide tradizionaletan ikusi dugun beste joera bat ere ageri da hemen: euskaldun berriak hurbilago daude, kontsumoari dagokionez, erdaldun zaharrengandik gainerako euskaldunengandik baino. Honakoan ere, zuhurtzia berdinaz irakurri beharreko datua da hori.

niveles generales de consumo bajos, por lo que las variaciones entre grupos que se observan en el gráfico podrían no ser muy significativas y entrar dentro de los márgenes de error. Ello nos obliga a interpretar los datos con cautela a la espera de otros datos más precisos que complementen a éstos.

También aparece la tendencia, ya observada con los medios tradicionales, a que las y los *euskaldun berriak* estén más cerca de quienes desconocen el euskera que de quienes lo tienen como primera lengua. Pero aquí hemos de aplicar la misma sobriedad en la interpretación.

Laugarren atala: ondorioak eta aurrera begirakoak

Di - Datu Integralak proiektuaren berehalako helburua bi datu-base bateratzea zen (CIESen urteko inkestarena eta IKUSIKER panelarena), euskarazko komunikazio-praktiketan gertatzen ari diren aldaketen ikuspegi zabalagoa eta globalagoa lortzeko. Bi datu-baseek, banan-banan hartuta, aldaketa horiek ulertzeko planteatzen dituzten mugek ezartzen zuten fusioaren beharra. Haietako bakoitzak eremu desberdinak hartzen ditu, komunikabide tradizionaletan oinarritzen da lehena eta sareko praktiketan bigarrena. Hala ere, bi iturriek lan egiten dute Araba, Bizkaia, Gipuzkoa eta Nafarroako biztanleek osatutako unibertso berean; kontsumo eta komunikazio-praktikak ikertzen dituzte; eta hizkuntzari buruzko informazioa ematen dute. Horrek, bien mikrodatuak eskuratzearekin batera —lehenengoaren kasuan, CIESen lankidetza ordainezinari esker; eta bigarrenaren kasuan, datu-base propioa izateari esker—, posible egin du fusionatzea.

Horretarako, audientzien ikerkuntzaren arloan zein beste esparru batzuetan garatutako metodo matematikoak aplikatu dira. Hitz gutxitan esanda,

Capítulo cuarto: conclusiones y líneas de futuro

El proyecto Di - Datos Integrales tenía como objetivo inmediato realizar la fusión de dos bases de datos (la de la encuesta anual de CIES y la del panel IKUSIKER) para poder obtener una visión más amplia y global de los cambios que se están produciendo en las prácticas comunicativas en euskera. La necesidad de la fusión venía determinada por las limitaciones que ambas bases de datos, tomadas individualmente, plantean a la hora de entender esos cambios. Cada una de ellas aborda ámbitos diferentes, la primera más centrada en los medios de comunicación tradicionales, y la segunda en las prácticas en red. Sin embargo ambas fuentes trabajan sobre un mismo universo constituido por la población de Araba, Bizkaia, Gipuzkoa y Nafarroa; indagan sobre consumos y prácticas comunicativas; y ofrecen información sobre la lengua. Ello, junto con el acceso a los microdatos de ambas —gracias, en el caso de la primera, a la inestimable colaboración de CIES; y en el caso de la segunda a ser una base propia— las convierte en susceptibles de ser fusionadas.

Con ese fin se han aplicado métodos matemáticos desarrollados tanto en el ámbito de la investigación de audiencias como en otros

Naïve hurbilketa duten *Hot-Deck* metodoak oinarri hartuta, proposamen metodologiko bat garatu da. Horrek ekarri du, lehenik eta behin, datu-base *emaile* bat eta *hartzaile* bat ezartzea, datu-base *emaile*aren informazioa erabiliz falta diren aldagaiak datu-base *hartzaile*an egozteko. IKUSIKERen datu-basea ezarri da *emaile* gisa, aztertutako unibertsoko erregistro gehien zituena baita. Kontuan hartu behar da IKUSIKER paneleko unibertsoa 12 eta 21 urte bitarteko ikasle-populazioa dela, eta CIESekoa, berriz, 14 urtetik gorako biztanleria osoa (bosturteko taldetan taldekatua bada ere); hortaz 14 eta 19 urte bitarteko gazteek osatzen dutela unibertso partekatua. Segidan, datu-baseen arteko estratifikazio komuna identifikatu ondoren, aldagai komunak kokatu, harmonizatu eta birsailkatu dira. Gero, indibiduoen arteko distantziak kalkulatu dira. Horren ostean fusioaren datu-basea lortu da, dohaintza-*emaileak* eta *hartzaileak* erlazionatuz. Azkenik, aplikatutako metodoa baliozkotu da.

Prozesuan zehar optimizazio problema bat detektatu da, *emaileak hartzaileekin* erlazionatzean. Gure erreferentziazko erakundeek (EUSTAT eta AIMC) egindako fusio-prozesuetan planteatu ez den arazoa da, eta proiektu honetan modu originalean ebatzi da.

ámbitos. Dicho en pocas palabras, se ha desarrollado una propuesta metodológica basada en métodos Hot-Deck con aproximación Naïve. Esto ha implicado, en primer lugar, establecer una base de datos *donante* y una *receptora*, e imputar en la base de datos *receptora* las variables faltantes utilizando la información de la base de datos *donante*. Se ha establecido la base de datos de IKUSIKER como *donante*, al ser la que contaba con mayor número de registros del universo analizado. Debe tenerse en cuenta que siendo el universo del panel IKUSIKER la población estudiantil de entre 12 y 21 años, y el de CIES toda la población mayor de 14 años (aunque agrupada en grupos quinquenales), el universo compartido está formado por las y los jóvenes de entre 14 y 19 años. Seguidamente, tras identificar la estratificación común entre las bases de datos se ha procedido a localizar, armonizar y recategorizar las variables comunes. Con posterioridad se ha llevado a cabo un cálculo de las distancias entre los individuos. Tras lo cual, relacionando individuos *donantes* y *receptores* se ha conseguido la base de datos de la fusión. Finalmente se ha procedido a la validación del método aplicado.

Durante el proceso se ha detectado un problema de optimización a la hora de relacionar *donantes* con *receptores*. Se trata de un problema no planteado en los procesos de fusión realizados por nuestras entidades de referencia

Lanketa horren ondorioz, 2.010 erregistro eta 6.198 aldagaitan banatutako 260 galderako datu-base sintetikoa lortu da.

Prozesuan zehar, erabilitako datu-baseei dagozkien mugak detektatu dira. Hala, fusioa berez arrakastatsua izan den arren, muga horiek, dagokion atalean azaldu den moduan, datu-base sintetikoaren azterketaren irismena ere mugatuta geratu da euskarazko komunikazio-praktikei dagokienez. Alde horretatik, azken produktu baten aurrean baino gehiago, bide baten abiapuntuaren aurrean gaude. Bide horrek, fusionatu beharreko datu-baseetan beharrezko zuzenketak aplikatu ondoren —2023an jadanik gauzatzen ari da hori—, medio tradizionalen kontsumoetatik sareko praktiketarako trantsizioa nola mamitzen ari den eta horrek euskarazko komunikazioan nola eragin dezakeen jakiteko jarraipen nahiko zehatza egitea ahalbidetuko digu.

Etorkizuneko fusioetarako kontuan hartu beharreko bestelako elementuak ere identifikatu dira. Horietako bat, ondoren fusionatu nahi diren datuak dituzten inkesten aldagai partekatuen kopurua handitzea da, lortutako artxiboa sendoagoa izan dadin. Beste bat, da akats marjinalari eta fusioaren bidez lortutako datuei eman behar zaien konfiantza-mailari buruz gehiago ikertu behar da. Gai hau funtsezkoa eta

(EUSTAT y AIMC) y que ha sido resuelto de manera original en este proyecto.

Fruto de ese trabajo ha sido la obtención de una base de datos sintética de 2.010 registros y 260 preguntas, divididas en 6.198 variables.

Durante el proceso se han detectado limitaciones correspondientes a las bases utilizadas. Así, si bien la fusión en sí ha resultado exitosa, dichas limitaciones, ya expuestas en el correspondiente apartado, han supuesto que el alcance del análisis de la base sintética en relación a las prácticas comunicativas en euskera quedase también limitado. En ese sentido, más que ante un producto final, nos encontramos ante el comienzo de una línea que en el futuro inmediato nos permitirá, tras aplicar las correcciones necesarias en las bases a fusionar –algo que ya se está materializando en 2023–, realizar un seguimiento bastante preciso de cómo se está produciendo el tránsito de los consumos de medios tradicionales a las prácticas en red, y cómo ello puede afectar a la comunicación en euskera.

También se han identificado otros elementos a considerar de cara a futuras fusiones. Uno de ellos tiene que ver con la conveniencia de aumentar el número de variables compartidas en las encuestas cuyos datos se pretenda posteriormente fusionar, pues ello

sakon aztertu beharrekoa da, nahiz eta proiektu honetan ezin izan den behar bezala ebatzi.

Bide honen zuzeneko onuradun, batetik, euskarazko komunikabideen sektorea izango da, bere publiko erreal eta potentzialen ezagutza zehatzagoa izango baitu. Bestetik, Eusko Jaurlaritzako Hizkuntza Politikarako Sailburuordetza eta Euskal Herrian hizkuntza- eta komunikazio-politikez arduratzen diren erakunde guztiak. Horrela, hobeto orientatu ahal izango dituzte ildo estrategikoak eta jarduera-ildoak, ebidentzietan oinarrituta. CIESek ere etekina atera diezaioke ildo horri, emaitzak audientzien ikerkuntzan aplika baitaitezke.

Halaber, Euskal Hedabideen Behategiari mesede handia egin dio proiektuak, batez ere, azterketa matematikoa bere lan-ildoetan sartzeko funtsezkoa izan delako. Horri esker, fusioaren metodologian sakondu ez ezik, beste ildo batzuetan sartu ahal izango da, hala nola prospektiban, ereduak eta profilak prestatzean, eta abar.

Baina komunikazio-praktiken eremutik harago, gure inguruko ikerkuntza soziologikorako ekarpen metodologiko erabilgarria egiten dugulakoan gaude, eta hori zen Datu Integralak proiektuaren eta Eusko Jaurlaritzaren Prospekzio Soziologikoen Kabineteak zein Soziored redundará en una mayor solidez del archivo conseguido. Otro es la necesidad de indagar más sobre el error marginal y el nivel de confianza que es necesario atribuir a los datos obtenidos mediante la fusión. Esta es una cuestión clave que debe ser analizada en profundidad, si bien en este proyecto no se ha podido dilucidar.

Los beneficiarios directos de esta vía emprendida serán, pues, de un lado el sector de los medios de comunicación en euskera, que dispondrá de un conocimiento más preciso de sus públicos reales y potenciales. Y de otro lado la Viceconsejería de Política Lingüística del Gobierno Vasco y todas aquellas otras instituciones encargadas de las políticas tanto lingüísticas como de comunicación en Euskal Herria, que podrán así orientar mejor sus líneas estratégicas y de actuación en base a evidencias. También CIES podría resultar beneficiada de esta línea, en la medida en que los resultados puedan aplicarse a su oferta de estudios de audiencias.

El Observatorio de Medios en Euskera Behategia ha resultado asimismo claramente favorecido con este proyecto, fundamentalmente por haber resultado clave para la introducción del análisis matemático en sus líneas de trabajo. Ello permitirá no solo profundizar en la metodología de la fusión, sino adentrarse en otras líneas como la prospectiva, la elaboración de modelos y perfiles, etcétera.

logia eta Zientzia Politikoaren Euskal Elkarteak deitutako *José Ignacio Ruiz Olabuenaga 2. ikerketa-bekaren* bigarren helburu nagusia. Aldaketa azkarren garaian bizi gara. Komunikazioaren ikuspegitik azaldu ditugun aldaketa horietako batzuk, baina gizarteko beste arlo batzuei ere eragiten diete. Eta aldaketa horiek begirada arin eta zorrotza eskatzen dute. Bestalde, Big Data garaian bizi gara, eta sekula baino informazio gehiago dugu portaera sozialei buruz. Beraz, iturri propio eta ad hoc diseinatuak, batetik, eta eskura dauden bestelako iturriak, bestetik –izan norberarenak edo izan besterenak– konbinatuko dituzten ikerkuntza-estrategiak ezartzea posible eta beharrezkoa dela dirudi, aprobetxamendua hobetzeko. Datu irekien politika publikoek (adibidez, mikrodatuen biltegiek) estrategia horiek erraztu ditzakete. Hala, inkestetako zein paneletako datuak eta zentsu-datuak fusionatzea, adibidez, oso erabilgarria izan daiteke soziologian. Horri esker, ikerketa-eremu zabalagoak hartu ahal izango dira, askotan bokazio publikoa duen gizarte ikerkuntzaren ahaletik oso urrun dauden inbertsio ekonomiko erraldoirik gabe.

Pero más allá del ámbito de las prácticas comunicativas, creemos hacer una aportación metodológica útil para la investigación sociológica en nuestro entorno, lo cual constituía el segundo gran objetivo del proyecto Datos Integrales y de la 2ª Beca de Investigación José Ignacio Ruiz Olabuenaga, convocada por el Gabinete de Prospección Sociológica del Gobierno Vasco y la Asociación Vasca de Sociología y Ciencia Política, gracias a la cual aquel se ha podido materializar. Vivimos en una época de cambios veloces, cambios que hemos ilustrado desde el punto de vista de la comunicación pero que afectan a muchas otras áreas de la sociedad. Y esos cambios precisan de una mirada ágil y atenta. Por otro lado, habitamos en la era del Big Data, con más información que la que nunca ha habido sobre los comportamientos sociales. Parece pues posible y necesario establecer estrategias de investigación que combinen la búsqueda de fuentes propias y diseñadas ad hoc con el aprovechamiento de otras fuentes disponibles, propias o ajenas. Las políticas públicas de datos abiertos (como por ejemplo los repositorios de microdatos) pueden facilitar esas estrategias. Así, la fusión de datos de encuestas y paneles con datos censales, por ejemplo, puede resultar de gran utilidad para la sociología. Ello permite abarcar ámbitos de investigación más amplios, sin precisar inversiones económicas muchas veces inalcanzables para la investigación social de vocación pública.

Erreferentziak / Bibliografía

Amezaga Albizu, J. (2022). Audientziak aztertzeko sistema: Burutzeke dugun erronka. In L. Mimenza Castillo (Arg.), *Euskal hedabideen urtekaria 2021* (or. 147–162). Hekimen. https://behategia.eus/wp-content/uploads/2021/07/urtekaria2020-web.pdf

Anand, D. (2020, ekainak 19). Gower's Distance. *Analytics Vidhya*. https://medium.com/analytics-vidhya/gowers-distance-899f9c4bd553

Anderson, B. R. O. G. (1983). *Imagined communities: Reflections on the origin and spread of nationalism*. Verso.

Arana Arrieta, E. (2011). *Estrategias de programación televisiva*. Síntesis.

Asociación para la Investigación de Medios de Comunicación AIMC. (2022). *Muestra teórica EGM 2022*. https://www.aimc.es/egm/que-es-el-egm/universo-y-muestra/

Bárcena, M. J., & Tusell, F. (1999). Enlace de encuestas: Una propuesta metodológica y aplicación a la Encuesta de Presupuestos de Tiempo. *Documentos de Trabajo BILTOKI, Nº. 7, 1998*.

Beck, J. (2015). SKO, Video Integration Model: Building the Factory. *EMRO Conference 2015*. EMRO Conference 2015, Stockholm.

Bello, A. L. (1993). Choosing among imputation techniques for incomplete multivariate data: A simulation study. *Communications in Statistics - Theory and Methods*, *22*(3), Article 3. https://doi.org/10.1080/03610929308831061

Billig, M. (1995). *Banal Nationalism* (1. arg.). SAGE Publications Ltd.

Calhoun, C. (1991). Indirect Relationships and Imagined Communities: Large Scale Integration and the Transformation of Every Day Life. In P. Bourdieu & J. Coleman (Arg.), *Social Theory for a Changing Society* (or. 95–121). Westview Press.

CIES. (2021). *Estudio de audiencia de medios*.

CIES. (2022). *Estudio de audiencia de medios. Acumulado 2022*.

Comscore. (2019, maiatzak 28). Acuerdo Marco Comscore (2019-2020). *Asociación para la Investigación de Medios de Comunicación*. https://www.aimc.es/blog/acuerdo-marco-comscore-2019-2020/

Cormack, M. (1998). Minority language media in Western Europe: Preliminary considerations. *European Journal of Communication*, *13*(1), 33–52.

Curran, J. (1991). Rethinking the media as a public sphere. In P. Dahlgren & C. Sparks (Arg.), *Communication and citizenship* (or. 27–57). Routledge.

De Waal, T. (2015). *Statistical matching: Experimental results and future research questions*. CBS. https://doi.org/10.13140/RG.2.1.1969.4161

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation*. Wiley.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Deutsch, K. W. (1953). *Nationalism and Social Communication* (1966. arg.). MIT Press.

D'Orazio, M. (2013). *Statistical matching: Metodological issues and practice with R-StatMatch*.

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

EMRO European Media Research Organisation. (2022). *Emro Audience Survey Inventory (EASI) 2022*. https://www.emro.org/easi/easi2022.html

European Commission. (2023). *EUROSTAT*. http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

EUROSTAT. (2013). *Statistical matching: A model based approach for data integration : 2013 edition*. Publications Office. https://data.europa.eu/doi/10.2785/44822

Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Jour-*

nal of the American Statistical Association, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Fishman, J. A. (1991). *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*. Multilingual Matters.

Garmendia Navarro, I. (2014). *Enlace de encuestas*. Euskal Estatistika Erakundea. https://www.eustat.eus/documentos/datos/Enlace_de_encuestas_c.pdf

Gerbner, G. (1986). *Living with Television: The Dynamics of the Cultivation Process*. Taylor & Francis.

Gifreu, J. (1989). *Comunicació i reconstrucció nacional*. Pórtic.

Habermas, J. (1962). *Historia y crítica de la opinión pública. La transformación estructural de la vida pública* (3. arg., Libk. 1986). Gustavo Gili.

Hilmarsson-dunn, A. M. (2006). Protectionist Language Policies in the Face of the Forces of English. The Case of Iceland. *Language Policy*, *5*(3), 295–314. https://doi.org/10.1007/s10993-006-9027-2

Jerabek, H. (2001). Paul Lazarsfeld--The Founder of Modern Empirical Sociology: A Research Biography. *International Journal of Public Opinion Research*, *13*(3), 229–244. https://doi.org/10.1093/ijpor/13.3.229

Jones, E. H. G. (2013). Minority Language Media, convergence culture and the indices of linguistic vitality. In E. H. G. Jones & E. Uribe-Jongbloed (Arg.), *Social Media and Minority Languages: Convergence and the Creative Industries* (or. 58–72). Multilingual Matters.

Kadane, J. B. (2001). Some Statistical Problems in Merging Data Files. *Journal of Official Statistics*, *17*(3), 423–433.

Kaye, B. K., & Johnson, T. J. (2003). From here to obscurity?: Media substitution theory and traditional media in an on-line world. *Journal of the American Society for Information Science and Technology*, *54*(3), 260–273. https://doi.org/10.1002/asi.10212

Morán Alaez, E., Martínez Rollón, P., & Vázquez Sancho, P. (2008). Nuevos métodos en la elaboración de estadísticas censales de pobla-

ción. EPV06. *JECAS: Jornadas de Estadística de las Comunidades Autónomas*. JECAS: Jornadas de Estadística de las Comunidades Autónomas, Santander.

Moriarity, C., & Scheuren, F. (2001). Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure. *Journal of Official Statistics*, *17*(3), Article 3.

Moring, T., Husband, C., Lojander-Visapää, C., Vincze, L., Fomina, J., & Mänty, N. N. (2011). Media use and Ethnolinguistic Vitality in bilingual communities. *Journal of Multilingual and Multicultural Development*, *32*(2), 169–186. https://doi.org/10.1080/01434632.2010.541918

Muñoz, A., & Villagarcía, T. (1997). Imputación de datos censurados mediante redes neuronales: Una aplicación a la EPA. *Cuadernos económicos de ICE*, *63*, 193–204.

Nafría, E., & Vásques, P. (2016). TV y Total Vídeo: La nueva medición Cross Media. *32o AEDEMO TV*. 32o AEDEMO TV, Girona.

Nordbotten, S. (1996). Neural Network Imputation Applied to the Norwegian 1990 Population Census Data. *Journal of Official Statistics*, *12*(4). https://www.proquest.com/central/docview/1266843391/abstract/70894107AE3A4988PQ/2

ODEC, & QUINAO. (2009). EGM 2008. Un año de fusión. *EGM 40, AIMC 20 ;1968-1988*. EGM 40, AIMC 20 ;1968-1988, Madrid.

Okner, B. (1972). Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File. In *Annals of Economic and Social Measurement, Volume 1, number 3* (or. 325–362). NBER. https://www.nber.org/books-and-chapters/annals-economic-and-social-measurement-volume-1-number-3/constructing-new-data-base-existing-microdata-sets-1966-merge-file

Otero Franco, L. (2010). *Unitate ekonomikoen erregistro administratiboak batu probabilitate-teknikak erabilita*. EUSTAT.

Pavlik, J. V. (2017). Austria's Legacy in Early Radio Broadcasting: Lessons for Audio Media in the 21st Century. *Athens Journal of Mass Media and Communications*, *3*(4), 273–296. https://doi.org/10.30958/ajmmc/3.4.1

Publicom AG. (2017). *Audience and Media Use Research—An International Perspective.pdf*. Swiss Federal Office of Communications.

Quintas-Froufe, N., González-Neira, A., & Ollero, M. (2021). *Los estudios de la audiencia: De la tradición a la innovación*. Gedisa.

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches* (Libk. 1–1 online resource (xviii, 264 pages)). Springer New York. https://doi.org/10.1007/978-1-4613-0053-3

Rius, R. (1994). *Insercion de datos de encuesta mediante analisis de componentes principales*. Congreso Nacional de Estadística e Investigación Operativa.

Rius, R., Nonell, R., & Aluja-Banet, T. (1996). File Grafting: A Data Sets Communication Tool. In A. Prat (Arg.), *COMPSTAT* (or. 417–422). Physica-Verlag HD. https://doi.org/10.1007/978-3-642-46992-3_56

Roda Fernández, R. (1989). *Medios de comunicación de masa.: Su influencia en la sociedad y en la cultura contemporánea*. CIS/Siglo XXI.

Rodgers, W. L. (1984). An Evaluation of Statistical Matching. *Journal of Business & Economic Statistics*, *2*(1), Article 1.

Rubin, D. B. (1986). Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics*, *4*(1), Article 1. https://doi.org/10.2307/1391390

Ruggles, N. D. (1974). *The Role of the Computer in Economic and Social Research in Latin America*. NBER. https://www.nber.org/books-and-chapters/role-computer-economic-and-social-research-latin-america

Sabaté, J. (2011). Els estudis d'audiència i la seva incidència en el sistema de comunicació. In *Informe de la Comunicació a Catalunya 2009- 2010* (or. 31–36). Institut de la Comunicació. Universitat Autónoma de Barcelona. https://incom.uab.cat/informe/index.html

Sabaté, J. (2020). The impact of audience measurement institutions on local media. Study of the Catalonian case. *Tripodos*, *46*, 137–156. Scopus.

Santiago, F., & AIMC. (2017). Internet: Un paso hacia la convergencia en la medición de audiencias. *AEDEMO TV*. AEDEMO TV, León.

Sigurjónsdóttir, S., & Nowenstein, I. (2021). Language acquisition in the digital age: L2 English input effects on children's L1 Icelandic. *Second Language Research*, *37*(4), 697–723. https://doi.org/10.1177/02676583211005505

The Ultimate Google Analytics Glossary. (2022). Loves Data. https://www.lovesdata.com/blog/google-analytics-glossary

UPV/EHU, EITB, Tabakalera, & Kulturaren Euskal Behatokia. (2023). *Ikusiker – Ikus-entzunezko produktuen kontsumo ikerketa*. Ikusiker. http://ikusiker.eus/

Yancey, W. E. (2007). Record Linkage: Theory and Practice. *U.S. Census Bureau*.

Zabaleta, I., Gutierrez, A., Ferré-Pavia, C., Fernandez, I., & Xamardo, N. (2018). Facts and transformations in European minority language media systems amid digitalization and economic crisis. *International Communication Gazette*, 174804851875474. https://doi.org/10.1177/1748048518754749

Eusko Jaurlaritzaren Argitalpen Zerbitzu Nagusia
Servicio Central de Publicaciones del Gobierno Vasco

ISBN: 978-84-457-3719-4