Aplicació de la ciència de dades i dels sistemes d'informació geogràfica (SIG) a la demolingüística

Marc Belzunces
Dadística
dadistica.dades@gmail.com

Recepció: 21/03/2017, acceptació: 21/08/2017

Resum: Cada vegada més les diverses àrees de coneixement recorren a eines desenvolupades en altres disciplines. Aquesta interdisciplinarietat és un motor d'innovació i generació de nou coneixement, permetent afrontar problemes des d'altres perspectives i facilitar-ne la solució. Presentem dos camps que poden ser d'utilitat a la demolingüística. Els sistemes d'informació geogràfica (SIG), que s'apliquen des de fa dècades en altres disciplines, permeten tractar i analitzar dades amb un component geogràfic. La Ciència de Dades, una disciplina molt recent, ens proporciona eines per treballar amb un gran volum de dades, dotant-nos d'unes habilitats notablement superiors a les que ens proporcionen les eines tradicionals.

Mots claus: dades, sistemes d'informació geogràfica, demolingüística

Aplicación de la ciencia de datos y de los sistemas de información geográfica (SIG) en la demolingüística

Resumen: Cada vez más las diferentes áreas de conocimiento recurren a herramientas desarrolladas por otras disciplinas. Esta interdisciplinariedad es un motor de innovación i de generación de nuevo conocimiento que permite afrontar problemas des de otras perspectivas y facilitar así su solución. Presentamos dos campos que pueden ser de utilidad para la demolingüística. Los sistemas de información geográfica (SIG), que se aplican desde hace décadas en otras disciplinas, permiten tratar i analizar datos con componente geográfico. La Ciencia de Datos, una disciplina muy reciente, nos proporciona herramientas para trabajar con un gran volumen de datos, dotándonos de unas habilidades notablemente superiores a las que nos proporcionan las herramientas tradicionales.

Palabras clave: datos, sistemas de información geográfica, demolingüística

Application of Data Science and Geographic Information Systems (GIS) to demolinguistics

Abstract: Increasingly, several areas of knowledge are turning to tools developed in other disciplines. This interdisciplinarity is a driving force of innovation and generation of new knowledge, allowing to face problems from other perspectives and facilitate their solution. Here, two fields that can be useful for demolinguistics are introduced. Firstly, Geographic information systems (GIS), which have been applied for decades in other disciplines, allow the processing and analysis of data with a geographic component. Secondly, Data Science, a very recent discipline, give us tools to work with a large volume of data, providing us with skills that are significantly superior to those provided by traditional tolos.

Key words: data, geographic information system, demolinguistics

1. Introducció

La revolució tecnològica de finals del segle XX, amb l'aparició dels ordinadors personals, d'Internet i l'abaratiment i popularització de tecnologies fins fa poc només reservades a un minoria, ha augmentat notablement la interdisciplinarietat, tant en l'entorn científic com en l'entorn laboral. Fins al punt que investigadors que treballen en àmbits amb accés a tecnologies i tècniques molt sofisticades les han aplicat a altres àrees de coneixement completament fora de la seva expertesa, i han obtingut resultats innovadors d'àmbit mundial. N'és un exemple la culturòmica (Michel et al. 2010), que estudia les tendències culturals i el comportament humà a través d'aplicació de tècniques matemàtiques i computacionals sobre un gran volum de textos antics digitalitzats. Un altre exemple és l'establiment d'una taxonomia per a les llengües aplicant tècniques estadístiques de la biologia evolutiva, estudiant l'evolució d'uns mots determinats en textos que abasten un període llarg de temps, com si fossin gens biològics. O els estudis que s'han fet a partir de les xarxes socials, obtenint mapes d'ús de llengües, o l'estimació del turisme, amb un detall impensable fins fa molt poc.

Avui dia, per tant, l'aplicació a la nostra àrea de coneixement de recursos, habilitats i tècniques externes a ella és una font de coneixement nou arreu del món. I això mostra, també, que a banda dels coneixements de la nostra àrea, per avançar necessitem adquirir habilitats que fins ara no eren considerades necessàries o que no tenien res a veure amb la nostra àrea de coneixement. En aquest article comentem breument dues habilitats que poden resultat d'interès en la demolingüística.

2. Els Sistemes d'Informació Geogràfica (SIG)

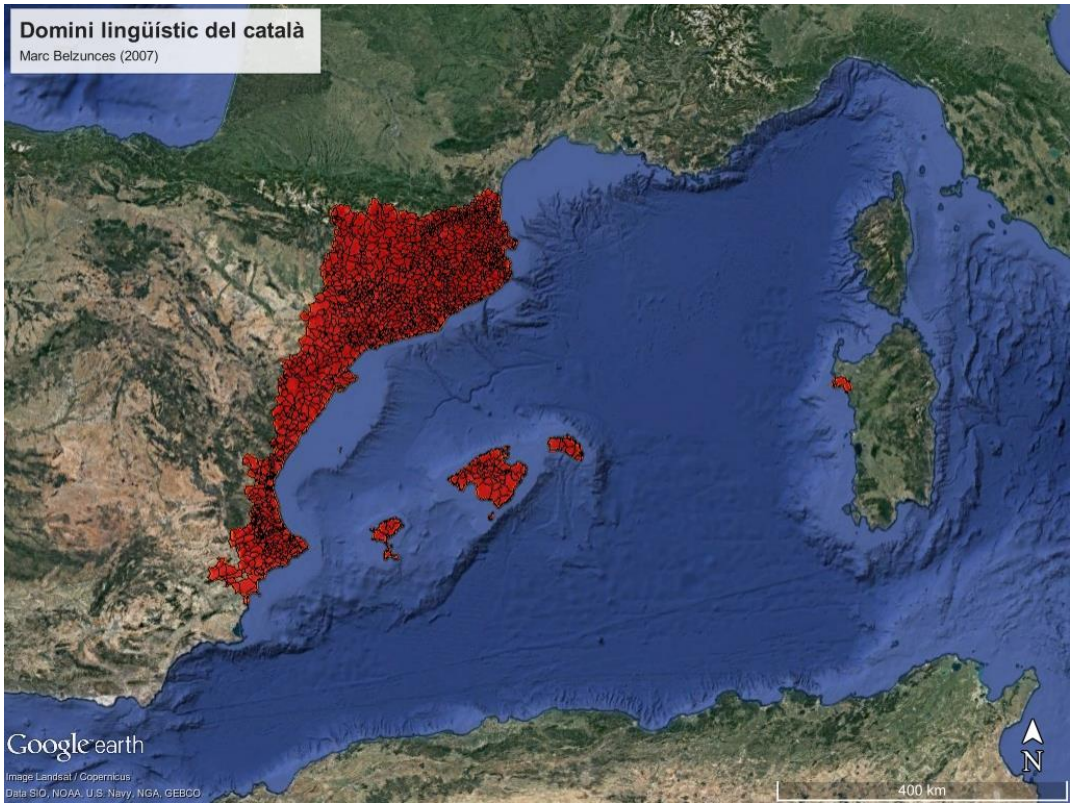
Els SIG són un sistema informàtic (generalment un programa d'escriptori) dissenyat per treballar amb dades amb un component geogràfic (típicament coordenades), analitzar-les i representar-les. Fa dècades que s'utilitzen en disciplines com la geografia, la geologia, l'edafologia o la gestió forestal. El seu resultat més visible són els mapes, però van molt més enllà. Un concepte associat als SIG és el de la cartogràfica temàtica. Si una persona comuna associa un mapa a un mapa polític, en el context del SIG un mapa polític és només un tipus de mapa temàtic. Els mapes poden representar qualsevol dada o tema, com per exemple l'edat geològica dels terrenys, els conreus que hi ha en una determinada zona o els riscos d'inundació. Són una eina més que ens permet treballar amb les nostres dades.

Un SIG es compon, en general, de dos elements: una base de dades amb dades que contenen de manera directa o indirecta una informació geogràfica, i una cartografia digital. De manera directa entenem dades que contenen coordenades geogràfiques, i de manera indirecta dades com ara el nom d'una comarca, una població, una adreça o un codi postal. Per cartografia digital, que també conté una base de dades pròpia, entenem mapes com ara els de límits municipals, comarcals, estatals, de tipus de sols, etc. Mapes que generalment obtenim gratuïtament d'institucions, però que també podem generar nosaltres amb el SIG a partir de les nostres dades o adquirint-los a tercers.

A diferència dels mapes creats amb programes de dibuix, que difícilment es poden reaprofitar, a un SIG els mapes estan sempre situats espacialment damunt del globus terraqüi. Un cop hem creat una capa cartogràfica amb les nostres dades, podem superposar altres capes georeferenciades, de manera que podrem anar combinant successivament aquestes capes. A més de poder reaprofitar altres dades o capes cartogràfiques, els SIG ens permeten millorar la qualitat de la nostra cartografia, ja que podem superposar imatgeria satel·lital, on podem veure els accidents geogràfics, o altres capes amb límits administratius de més qualitat. Un exemple de tot això el mostrem a la Figura 1.1. Vam recopilar en un SIG la llista dels municipis del domini lingüístic del català, es va enllaçar amb la cartografia digital dels municipis, i es va acabar editant per generar els límits del Carxe (inexistents en

format digital i poc precisos en els mapes tradicionals) i seleccionar aquelles parts catalanoparlants de municipis oficialment castellanoparlants. Atès que tota la informació estava georeferenciada, vam exportar la capa del domini al Google Earth (un visualitzador més que no pas un SIG) i automàticament tenim el domini lingüístic sobreimposat amb altres capes: relleu, imatgeria satèl·lit, carreteres, límits polítics, etc.

FIGURA 1. El domini lingüístic del català per municipis al Google Earth



Font: Elaboració pròpia

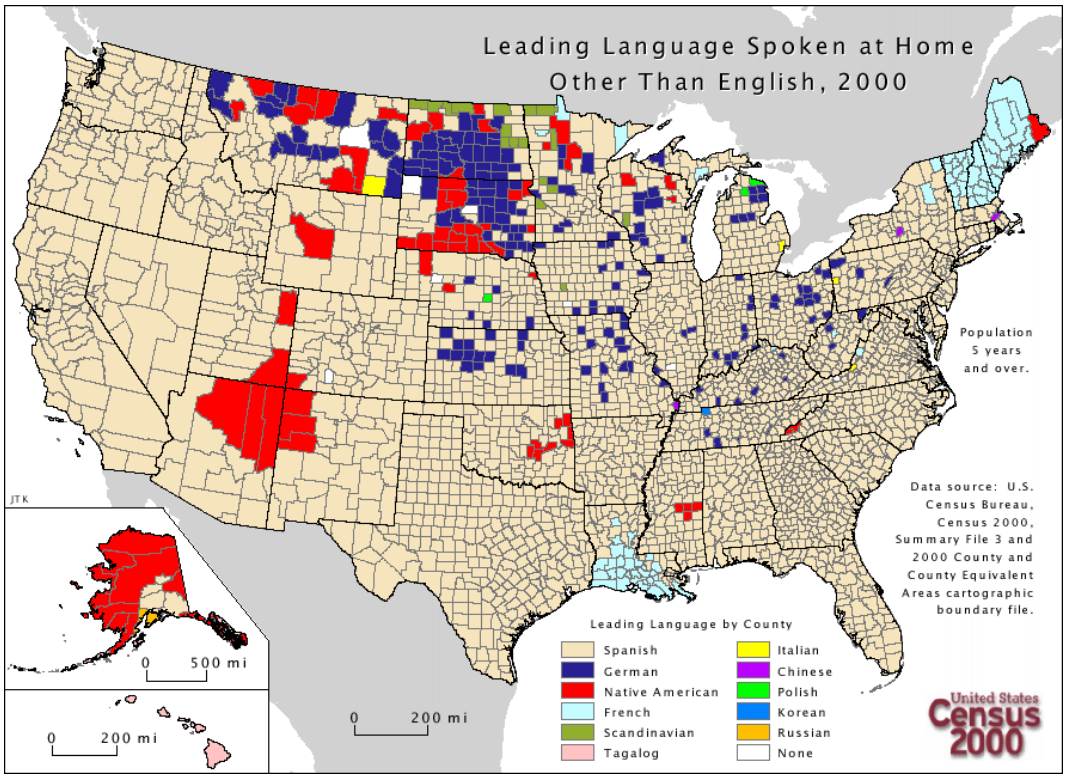
Els SIG, però, permeten no només crear representacions formals de mapes. L'aspecte fonamental és el d'anàlisi de les dades, on generem multitud de mapes en brut que responen a determinades qüestions. Per exemple, a la base de dades de la Figura 1.1 podríem afegir les dades municipalitzades de població, de lloc de naixement, de PIB o de turisme. A partir d'aquí, els SIG permeten fer consultes i obtenir una resposta. Així, podríem preguntar: *Quin són els municipis del domini lingüístic català amb menys de 25.000 habitants, amb una taxa d'atur inferior al 10% i on el 80% dels habitants hagin nascut al domini?* Obtindríem automàticament un mapa amb aquests municipis. A més, ens permet mostrar dades d'una manera més entenedora, o ampliar-ne la utilitat. Per exemple, podríem tenir una taula que digués que als municipis de menys de 10.000 habitants més del 60% dels habitants tenen com a llengua materna el català, sense especificar-ne el nom. Quins són aquests municipis? Només ens caldria fer la pregunta corresponent al SIG, i atès que hem afegit les dades d'habitants a la cartografia dels municipis, obtindríem ràpidament la resposta. Passem d'una taula que ens dona una idea poc precisa, a un mapa navegable i consultable que ens permet obtenir una millor idea. Això resulta de gran utilitat a l'hora d'aprofundir en el coneixement de les nostres dades i la validació o rebuig d'hipòtesis que podem plantejar.

Així doncs, és cabdal que quan recopilem informació pensem com aprofitar un Sistema d'Informació Geogràfica. Primer de tot hem de mirar de recollir la posició geogràfica amb el màxim detall. Si treballem amb dades recopilades per altres, podem mirar d'obtenir

aquesta posició de manera indirecta, com ara obtenir unes coordenades a partir d'una adreça postal. Després, hem de pensar quines altres capes d'informació podrem aprofitar, i quines haurem de crear expressament. Finalment, hem de mirar com representar gràficament de la millor manera les nostres dades al mapa.

Els SIG són programes complexos que tenen un corba d'aprenentatge difícil. Fins i tot poden requerir programació (típicament s'hi utilitza SQL, Python i Javascript) per desenvolupar nous tipus d'anàlisi o integrar-los en un entorn web. Tanmateix, en la majoria de casos podem construir progressivament un SIG amb les dades d'un full de càlcul i fer-ne anàlisis i visualitzacions senzilles que podem dur a terme directament amb els menús del programa. La Figura 1.2 mostra un d'aquests casos, el mapa de la llengua més parlada a la llar als EUA per comtat, a partir de les dades recopilades pel cens dels EUA.

FIGURA 2. Llengua més parlada a la llar als EUA, per comtat



Font: Google

Els SIG tenen una gran potencialitat en les ciències socials i les humanitats. En aquestes disciplines encara avui es fan servir majoritàriament programes de dibuix per dissenyar mapes. Tanmateix, hi ha diversos projectes per generar mapes històrics amb qualitat SIG, com ara la iniciativa comercial Euratlas, que cartografia l'evolució dels estats a Europa cada segle entre els anys 1 i 2000. Aquests tipus de projectes no només permeten una cartografia de més detall, sinó que comporten una recopilació de dades, la seva estructuració i l'establiment d'una jerarquia, adopció de criteris i la possibilitat d'anar incorporant-hi progressivament més dades (si està ben estructurat), corregir errors que vagin apareixent i adoptar nous criteris, generant automàticament nous mapes sense haver de refer tota la feina.

Pel que fa als programes de SIG, l'ArcGIS d'ESRI és la referència a escala mundial. Al nostre país tenim el Miramon, desenvolupat pel CREAF (UAB), o el gvSIG, impulsat per la Generalitat Valenciana. En programari lliure, el QGIS és possiblement el programa més utilitzat. Pel que fa a la formació, a banda de titulacions universitàries específiques, a Internet hi ha diversos cursos d'introducció (Coursera: 2), gran quantitat d'informació

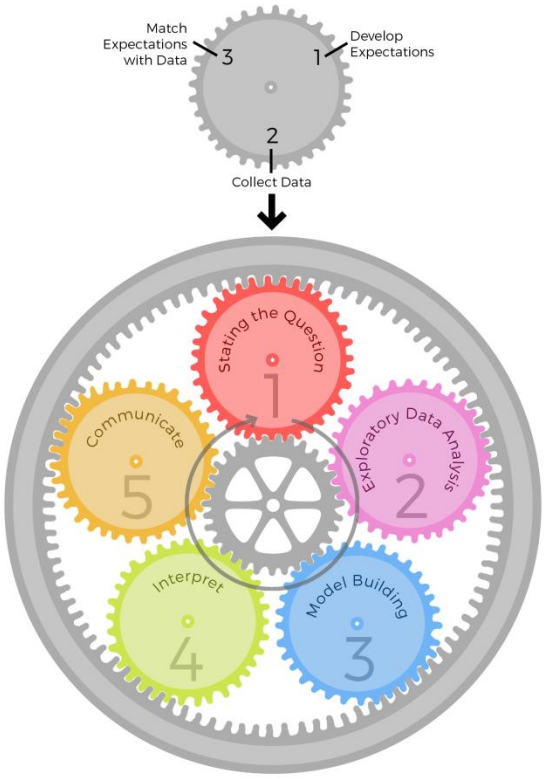
tècnica i nombrosos exemples que ens permetran anar millorant els nostres coneixements un cop adquirits els conceptes bàsics.

3. Ciència de Dades

Amb la ubiqüitat d'Internet i l'extraordinari abaratiment dels sensors, avui dia estem inundats de dades. Una persona pot amb facilitat recopilar centenars de milions de dades en pocs dies. Tant és així que es comença a parlar sobre el fet que la ciència ha entrat en un nou paradigma: l'anàlisi massiva de dades. És en aquest context que recentment s'està desenvolupant la ciència de dades, entès com el conjunt de coneixements i procediments que s'han de portar a terme per captar, tractar i analitzar grans quantitats d'informació, i extreure'n conclusions. Conceptes com el *big data* (dades massives), *data mining* (mineria de dades) o *machine learning* (aprenentatge automatitzat) no són més que aspectes parcials de la ciència de dades.

En ciència de dades convergeixen diversos perfils. Per una banda, cal saber un llenguatge de programació (típicament R o Python) per manipular les dades, i dotar-les de l'estructura necessària. Un full de càlcul va molt bé quan tenim poques dades, però a partir d'un nombre important no ens serveix. El llenguatge de programació permet addicionalment un aspecte important, especialment en el món acadèmic: la reproducibilitat. Tots els canvis que fem a les nostres dades queden recollits en el codi que hem programat, de manera que si volem modificar l'anàlisi, o detectem un error, podem tornar enrere, cosa prou més difícil, o impossible, en un full de càlcul.

Figura 3. Metodologia en ciència de dades



Font: *The Art of Data Science*, Roger D. Peng i Elizabeth Matsui

El segon perfil és el de l'estadística, que utilitzem per explorar les dades, interpretar-les i establir-ne un model matemàtic. Tant en el perfil anterior com en aquest, convé aclarir que no cal ser un especialista en aquestes àrees, però sí estar familiaritzat amb els

fonaments i la resolució pràctica. Igual que amb els SIG, trobem molts exemples i documentació a través d'Internet que ens permetrà anar resolent els problemes als quals ens enfrontem. Moltes de les grans aportacions en ciència de dades són fetes per persones curioses amb afany de resoldre els seus dubtes, més que no pas per especialistes de la programació o l'estadística.

El tercer perfil que s'utilitza és el de la visualització de les dades. Els gràfics no són aquell resultat formal per presentar un resultat, sinó que esdevenen una eina fonamental per familiaritzar-se i entendre les dades. Gràcies a la programació, podem elaborar fàcilment desenes i centenars de gràfics en brut. Finalment, podem elaborar gràfics formals per presentar públicament les nostres conclusions.

La Figura 2.1 mostra el cicle típic en ciència de dades. El primer aspecte fonamental és elaborar una pregunta a respondre el més concreta possible. Quan disposem de milions de dades apareixen multituds de preguntes, de manera que si no en tenim definida una clarament, ens perdem. Tot seguit hem de recopilar les dades necessàries per respondre aquesta pregunta i dotar-les de l'estructura necessària. Aquest pas generalment és el que consumeix més temps i esforç. A continuació explorem totes les nostres dades, descrivint cada variable i la relació entre aquestes. Un cop fet això, estem en disposició d'establir un model estadístic de les nostres dades que ens permeti explicar-les i fins i tot fer-ne prediccions. A continuació interpretem els nostres resultats, i finalment el científic de dades ha de comunicar mitjançant gràfics, dades numèriques o eines web els seus resultats. Les rodetes dentades en cada un d'aquests passos a la Figura 2.1 indiquen que en qualsevol d'aquests passos, i en funció dels resultats que obtenim, hem de tornar a iniciar aquell pas o el conjunt de passos que hàgim fet fins aleshores, per anar refinant-los. Començar de nou, ja sigui completament o en un dels passos, és un fet freqüent que pot fins i tot dur a la modificació de la pregunta inicial.

Per a totes aquelles persones que vulguin iniciar-se en ciència de dades i aplicar-la al seu camp de coneixement, la referència internacional és l'especialització de ciència de dades de la Universitat John Hopkins (Coursera:1), que imparteix a distància a través de Coursera i que ja té més de tres milions d'alumnes. Presencialment, universitats com la UB comencen a impartir màsters en ciència de dades (*Master Foundations of Data Science*). Pel que fa a les eines, les persones que venen del camp de les ciències o el món acadèmic, se sentiran més còmodes amb R i el seu entorn de desenvolupament RStudio. Aquelles persones que provinguin del món de l'enginyeria, probablement se sentin més còmodes amb Python i entorns de desenvolupament com Rodeo.

TAULA 1. Fidelitat lingüística a les Illes Balears

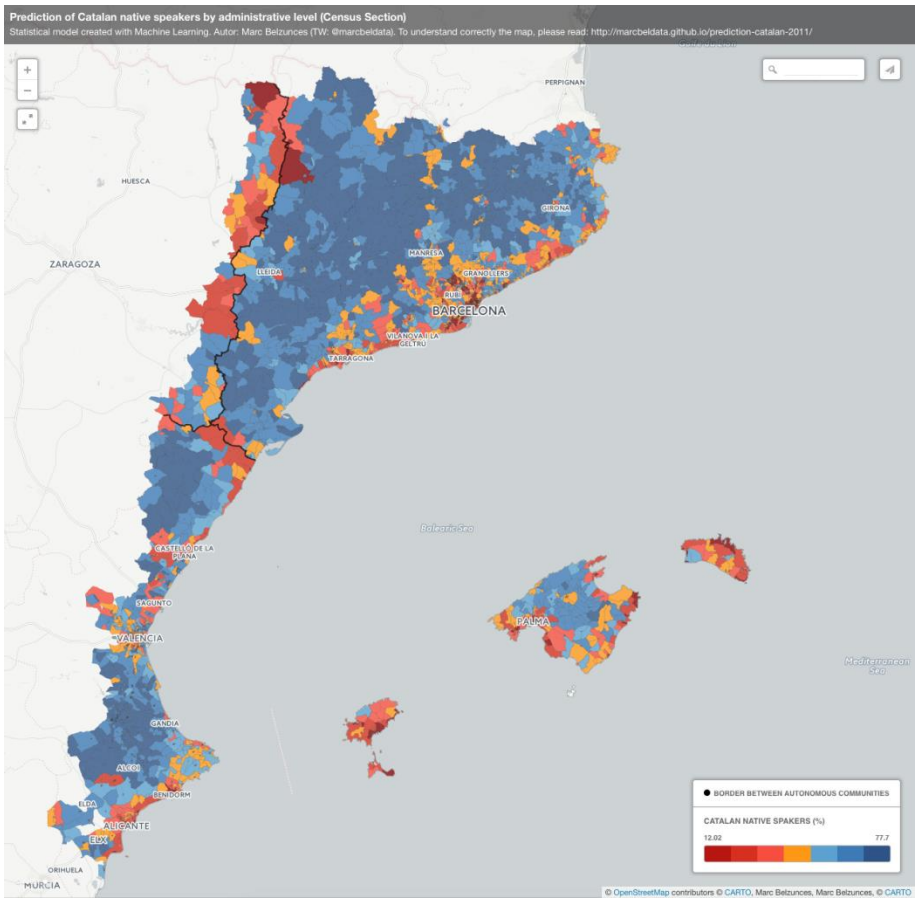
Grup lingüístic	**Percentatge**
Catalanoparlants: transmissió català	91,15 %
Castellanoparlants: transmissió castellà	86,98 %
Altres llengües: transmissió altres llengües	67,32 %
Catalanoparlants: abandonament català	3,44 %
Castellanoparlants: abandonament castellà	6,29 %
Altres llengües: transmissió català	19,29 %
Altres llengües: transmissió castellà	12,6 %
Bilingües: transmissió català	58,31 %
Bilingües: transmissió espanyol	26,78 %

Font: Belzunces (2016: 1)

3.1. Exemple d'aplicació: fidelitat lingüística a les Illes Balears

Fins ara hem parlat del cicle complet que es desenvolupa en ciència de dades, com mostra la Figura 2.1. Tanmateix, freqüentment la ciència de dades ens resulta útil quan, per desenvolupar un projecte, només ens cal dur a terme determinades tasques, sense completar tot el cicle. Com ara agafar dades en brut i dotar-les de l'estructura necessària, típicament una taula. Resulta especialment útil quan tenim un gran volum de dades i només ens interessa un o diversos subconjunts. O quan només volem explorar un conjunt de dades que són noves per a nosaltres. Un exemple d'aquesta aplicació parcial és el petit exercici que vam fer analitzant els resultats de *l'Enquesta modular d'hàbits socials 2010*, duta a terme per l'Institut d'Estadística de les Illes Balears (AUTOR 2016: 1). A partir d'aquesta enquesta vam analitzar amb relativa facilitat la fidelitat lingüística dels diversos grups lingüístics a les Illes Balears. A partir d'una pregunta inicial anem explorant les dades, la qual cosa genera noves preguntes que podem anar responent seguidament, un cop tenim l'estructura necessària de les dades i les eines (programació) per fer-ho. Addicionalment, proporcionant el codi i les dades originals, permetem que qualsevol persona pugui reproduir l'anàlisi, millorar-lo o detectar-ne errors (reproducibilitat). La Taula 2.1 mostra els resultats totals. Per a dades segmentades per edat, illes i comarques, així com el codi, consulteu Belzunces (2016: 1).

FIGURA 4. Mapa de parlants nadius per secció censal



Font: Belzunces (2016: 2). Consultable a Internet

3.2. Exemple d'aplicació: predicció de parlants nadius de català

Un exemple més elaborat i que mostra tota la potencialitat de les ciència de dades, integrada amb SIG a més, és l'estudi que vam fer per predir els parlants nadius de català (de llengua materna o inicial) en l'àmbit municipal i secció censal (AUTOR 2016: 2). A Catalunya no hi ha pràcticament dades sobre aquesta temàtica. A partir de *l'Enquesta demogràfica del 2007* feta per l'Institut d'Estadística de Catalunya vam voler explorar les potencialitats de la ciència de dades en un cas certament extrem, atès el nombre petit de la mostra. Aquesta enquesta és una de les poques que conté informació sobre parlants nadius de català, però només amb 52 observacions (per a cada comarca i per a les onze ciutats més grans de Catalunya). A partir d'aquí hi incorporem més dades (de l'Institut Nacional d'Estadística espanyol), establim un model estadístic i hi apliquem *Machine Learning* per obtenir un algoritme que ens permeti generar dades per secció censal, la unitat administrativa mínima a l'estat espanyol. La Figura 2.2 mostra el mapa dels resultats fet amb SIG, interfície que permet consultar-ne totes les dades (més informació i mapa consultable a AUTOR 2016: 2).

Tot això ens permet passar de 52 observacions a 8.780. A més, ampliem els resultats d'un territori (Catalunya) al conjunt del domini lingüístic sota administració espanyola. La metodologia, tanmateix, té problemes. Per una banda, partim d'un nombre de dades extremadament baix, quan normalment calen desenes de milers d'observacions. A més, no hem pogut accedir a través d'Internet a dades que ens consta que existeixen, i que permetrien millorar el model. I per altra banda, no hem anat més enllà atès que l'objectiu de l'exercici era estudiar-ne la potencialitat d'aplicació d'ús. En qualsevol cas, creiem que el resultat és notable, i des d'un punt de vista qualitatiu és correcte. Quantitativament els resultats cal agafar-los amb moltes precaucions, especialment pel que fa a les limitacions provocades per un nombre tan baix de dades.

4. Conclusions

Amb l'augment de la interdisciplinarietat i especialment amb la revolució tecnològica que ha dut a l'explosió de la disponibilitat de dades, eines com els sistemes d'informació geogràfica o la ciència de dades resulten de gran utilitat per a moltes àrees de coneixement, incloent-hi la demolingüística i les ciències socials i humanitats en general. Avui dia, que estem entrant en un nou paradigma científic, el de les dades, la capacitat de programar o de representar les dades de formes creatives són habilitats que comencen a ser indispensables, ja sigui des d'un punt de vista individual o dins del grup de treball. També plantegen nous reptes: necessitem augmentar la disponibilitat de dades lingüístiques, ja sigui fent-les públiques, noves enquestes o incorporant noves preguntes al padró o censos. Aquestes dades, però, han de ser de qualitat, i això ens obliga a replantejar-nos l'adquisició de dades, assegurant-nos que continguin tota aquella informació mínima necessària per després poder fer servir aquestes noves eines amb tota la seva potencialitat.

5. Referències Bibliogràfiques

Belzunces, Marc (2016). «*Language shift between Catalan and Spanish linguistic groups in the Balearic Islands» (2010)* [en línia]. Bloc personal.

<http://marcbeldata.github.io/language-shift-balearic-islands-2010/> [Consulta: 20 febrer 2016].

–– (2016). «*Extreme Machine Learning: Prediction of Catalan native speakers by administrative level*» [en línia]. Bloc personal.

<http://marcbeldata.github.io/prediction-catalan-2011/> [Consulta: 20 febrer 2016].

Coursera, John Hopkins University. Data Science Specialization [en línia].

<https://www.coursera.org/specializations/jhu-data-science> [Consulta: 20 febrer 2016].

Coursera, University of California. Geographic Information Systems (GIS) Specialization [en línia].

<https://www.coursera.org/specializations/gis> [Consulta: 20 febrer 2016].

Universitat de Barcelona. Master Foundations of Data Science [en línia].

<http://www.ub.edu/datascience/master/> [Consulta: 20 febrer 2016].

MICHEL, JEAN-BAPTISTE; LIEBERMAN AIDEN, EREZ; et al. (2010). «Quantitative Analysis of Culture Using Millions of Digitized Books». Revista *Science*, 16 de desembre de 2010. <http://science.sciencemag.org/content/early/2010/12/15/science.1199644> [Consulta: 20 febrer 2016].