Twitter como corpus para la variación lingüística algunos ejemplos prácticos

ANTONIO RUIZ TINOCO, UNIVERSIDAD SOFÍA
UNIVERSITAT DE BARCELONA
9 DE MARZO, 2016

Objetivos

MOSTRAR

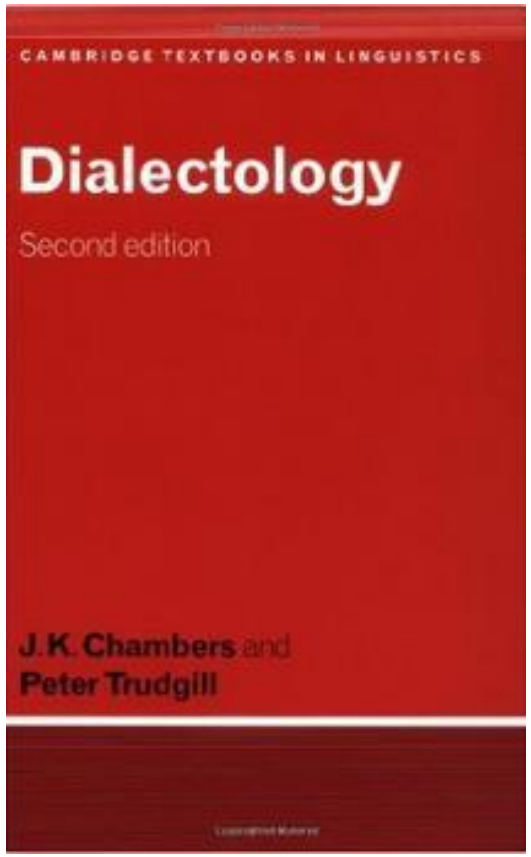
1. **Características básicas** de los datos geolocalizados de *Twitter*, qué son, cómo se obtienen y cómo se procesan.

2. **Ejemplos concretos** del uso de datos geocodificados obtenidos de ***Twitter***.

El **geocorpus** actual contiene más de 20 millones de tuits (más de **300 millones** de palabras) recogidos parcialmente en los años 2014-2016.

3. Preparación de varios ejemplos básicos de **mapas de distribución** de variación

Investigación tradicional de la dialectología



NORM

Non-mobile

Older

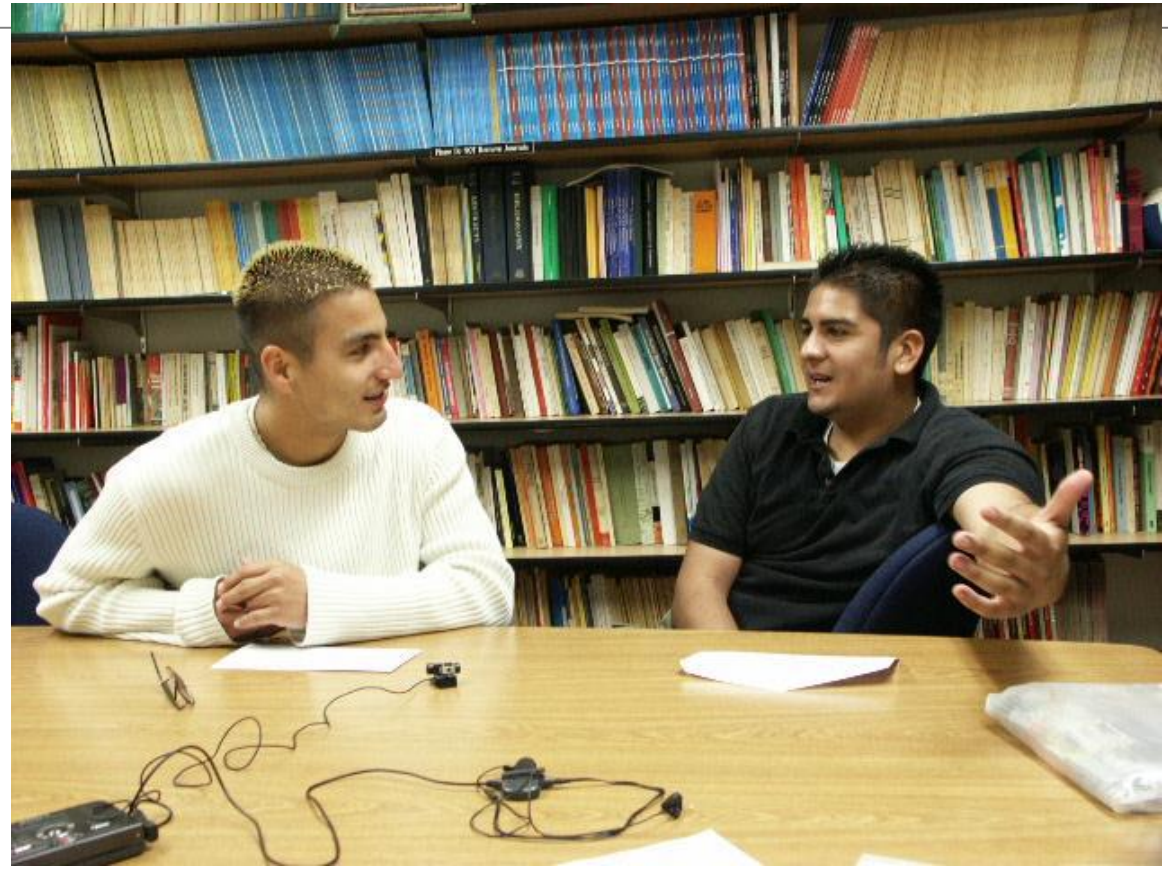
Rural

Male

Preguntar a los nativos

Encuesta en El Alto (Bolivia)

Encuesta en Los Angeles (USA)



Encuesta en Quito, Ecuador

Datos de *Twitter*

- En ***Twitter*** se escriben alrededor de **500 millones de *tuits* diarios.**
- Aproximadamente el **5% está en español**.
- Solamente alrededor del **1%** se puede obtener **gratis**.
- Según la zona geográfica, **solamente el 0.5% - 3% contiene información de las coordenadas**.
- Los **términos de uso** de *Twitter* limitan la libre distribución de los datos
- Hay un nivel muy alto de **ruido** (mensajes indescifrables, spam, citas, repeticiones, etc.)
- Casi el **80% del tiempo** dedicado al análisis hay que dedicarlo al **preprocesamiento de los datos** y en la **desambiguación manual** de los datos.

¿Qué información contiene un *tuit*?

- Un texto de **140 caracteres** como máximo
- Texto generalmente muy **espontáneo**, **sincrónico** y cercano al español hablado
- **Coordenadas** del lugar de procedencia (no todos) y hora exacta de su envío (1 seg)
- **Nombre** del usuario, aunque puede ser ficticio.
- **Perfil** del usuario
- Número de *tuits* escritos por el usuario
- Número de seguidores
- Número de usuarios a los que sigue
- etc.

¿Cómo se obtienen los datos de Twitter?

Hardware: **VPS**, Virtual Private Server (2G memory)

Entorno: **LAMP** (Linux, Apache, MySQL, PHP)
SIG (GIS) (cartografía): QGIS, SAGA, GDAL, etc.

Conexión a **Twitter**: https://twitter.com/signup + developper account
Oauth: Obtener claves de seguridad

Base de datos: MySQL, PostgreSQL+PostGIS, SpatiaLite, etc.

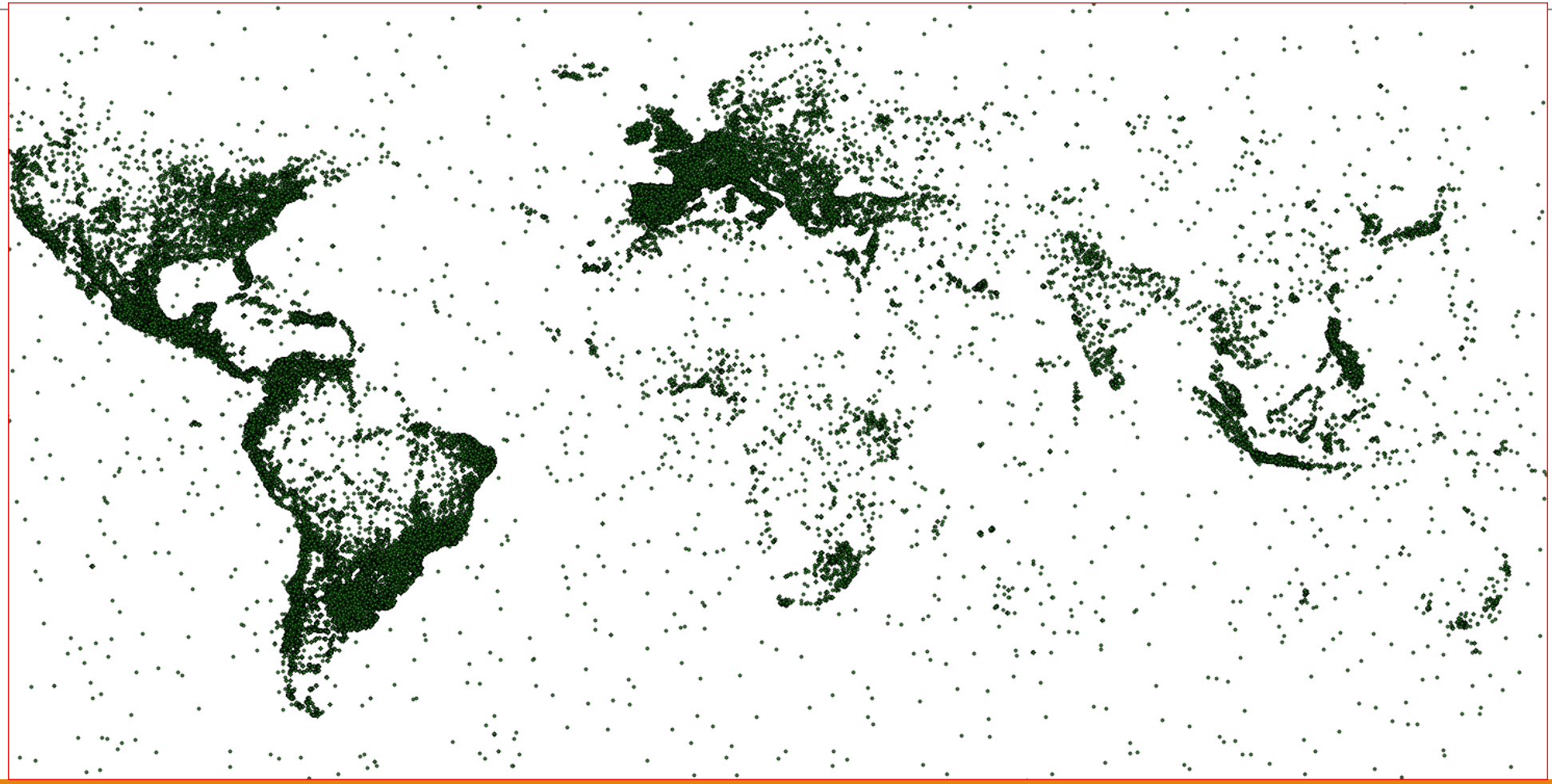
Varios métodos de obtención de datos

- Por palabras o secuencias de palabras
- Por coordenadas
- Pasados (10 días - 2 semanas)
- En tiempo real sin restricciones:

Streaming API (aprox. 60.000 tuits /día)

Distribución de 5.478.227 tuits

¿Dónde se usa el español?



Limpieza de los datos (ruido)

Detección de lengua:

tarda, verbo “tardar”o “tarda”(tarde) de catalán

Desambiguación

medias (prenda femenina, calcetines, parte de otra expression (“a medias”)

Citas de otras personas: RT @ (retweets)

Falta de coordenadas, spam, bots, publicidad, errores de hardware, etc.

Con frecuencia, comprobación final manual

Interfaz para la base de datos

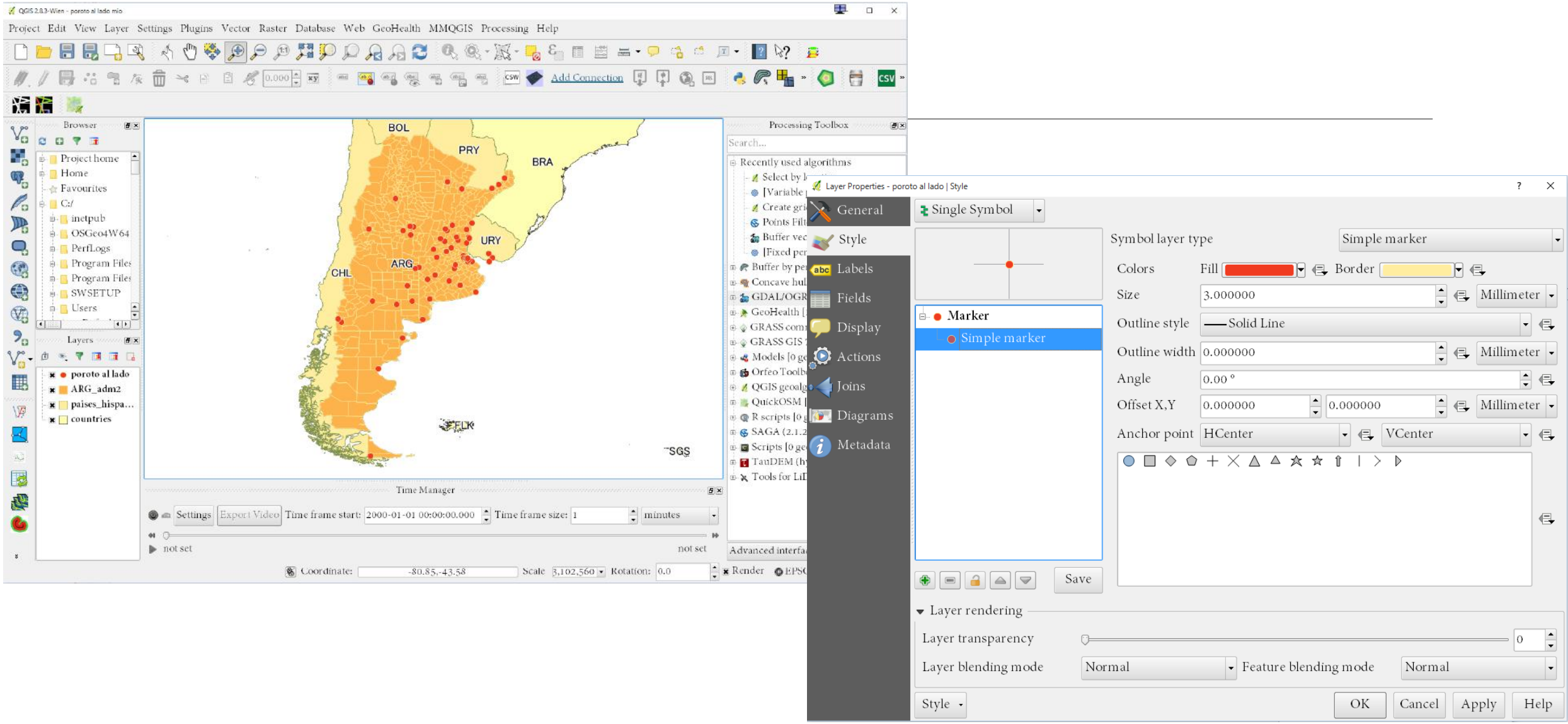
tweet_text	created_at	geo_lat	geo_long
El pato criollo un poroto al lado mio media pelia eri	2015-10-15 20:22:59	-34.41014	-58.61285
Cuando me habla me molesta , y cuando no me habla también me molesta La gata flora un poroto al la...	2015-10-15 21:13:00	-34.89978	-57.98931
Vivir sola & comer 3 huevos fritos con Pan. Master chef un poroto.	2015-10-16 13:32:00	-34.45580	-58.57694
#Picasso un poroto, by Agustín Monzón	2015-10-16 15:03:19	-38.00000	-57.55000
I'm at El Palacio Del Poroto Con Riendas in Estación Central, Santiago Metropolitan Region https://t...	2015-10-17 13:51:34	-33.46003	-70.69598
@MiluDhondt @GuidoRosano Luisiti de cdp. Los dos un poroto.. Se comenta q a su casamiento lo hacen e...	2015-10-17 14:04:36	-32.90422	-60.90742
@agustinadevivo un poroto ese	2015-10-17 15:13:28	-34.63931	-58.37492
Valu Ramallo un poroto. Y con los amargos de la Peco niiiiiii te… https://t.co/2TX3XOsB8P	2015-10-17 18:48:28	-32.95153	-60.64822
Con joni y poroto	2015-10-17 19:34:39	-33.01373	-58.55139
El coco basile un poroto al lado de mi prima jajajjaj	2015-10-18 17:34:08	-31.59192	-59.89131
@Camiilariadna como olvidar esa noche? Jajajaja los mineros un poroto,los quieroo	2015-10-18 19:14:52	-34.80018	-58.21428
Planté un poroto.. Y ahora se está convirtiendo en la primera planta de nuestro hogar..	2015-10-18 23:17:24	-33.43363	-70.65831
"@PrestaPrestico: El FBI un poroto al lado de una mina celosa." @MelinaFagliani nosotras somos mejor...	2015-10-19 03:30:54	-32.39281	-63.24969
Leche barbara un poroto jajaja	2015-10-19 04:05:22	-34.66517	-58.44575
"@NicolasssTm: El FBI un poroto al lado de una mina celosa." Jajajaja	2015-10-19 08:21:09	-32.97764	-60.65242
Bombón ! #Poroto @ Córdoba - Nueva Córdoba https://t.co/mzvyzp8rDm	2015-10-19 15:04:05	-31.42513	-64.18030
Y salió eso , #masterchef un poroto	2015-10-19 20:30:37	-34.65000	-58.58330
Me voy a morir de la tos que tengo, el pinguino de Toy Story un poroto al lado mio	2015-10-20 10:22:42	-41.14423	-71.26183
@AgustinDiaz88 la profe un poroto al lado tuyo	2015-10-20 12:20:59	-37.00225	-57.12177
Listo ya limpie todo Isaura la esclava un poroto al lado mío, ahora si , me fui :D	2015-10-20 13:30:29	-34.66793	-58.47340
Cupido un poroto	2015-10-20 15:29:33	-38.92254	-67.98016

SIG (Sistemas de Información Geográfica)

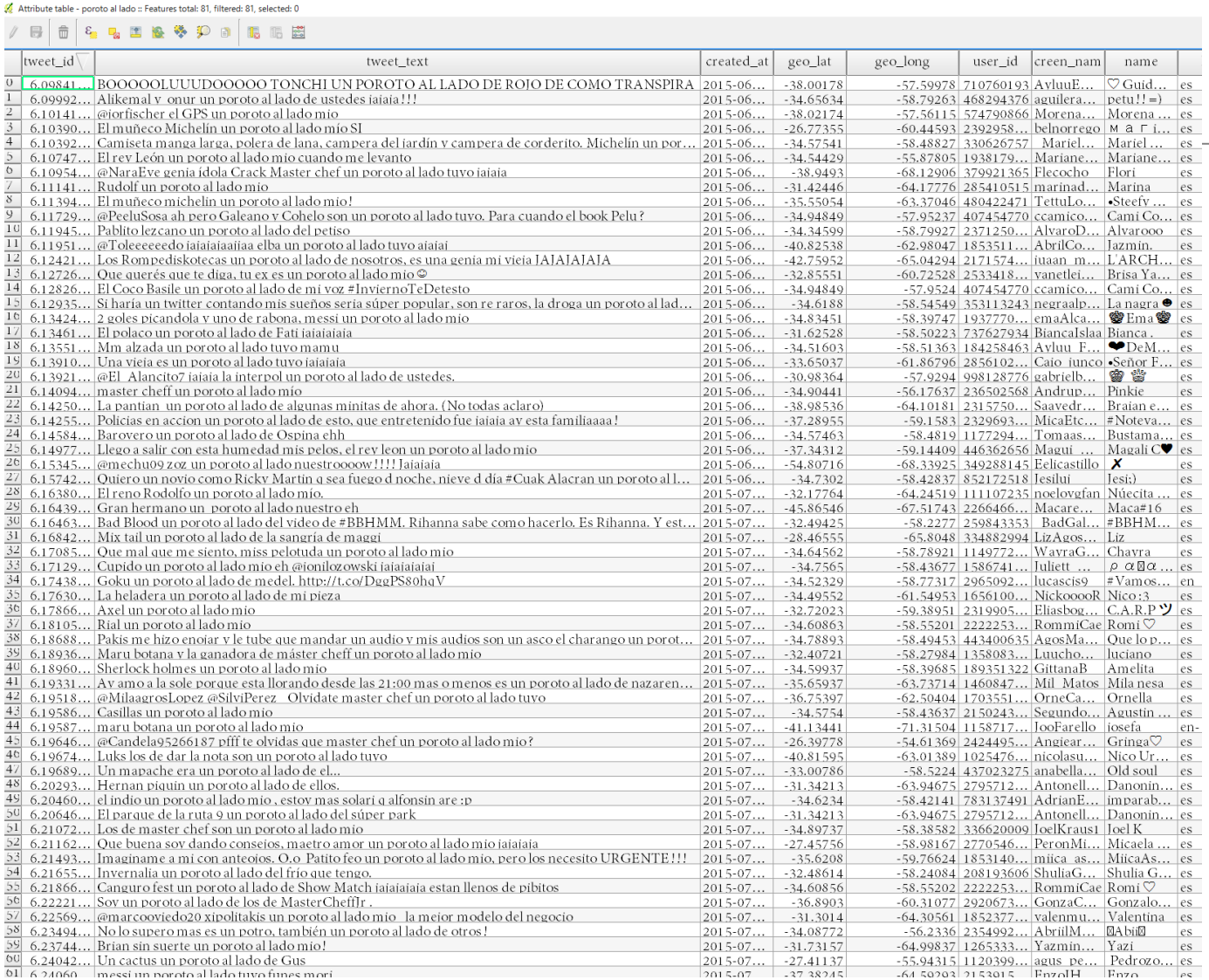
Los programas tipo SIG son herramientas capaces de integrar y almacenar información geocodificada.

Posibilitan consultas interactivas, analizar la información, editar datos, preparar mapas y presentar los resultados.

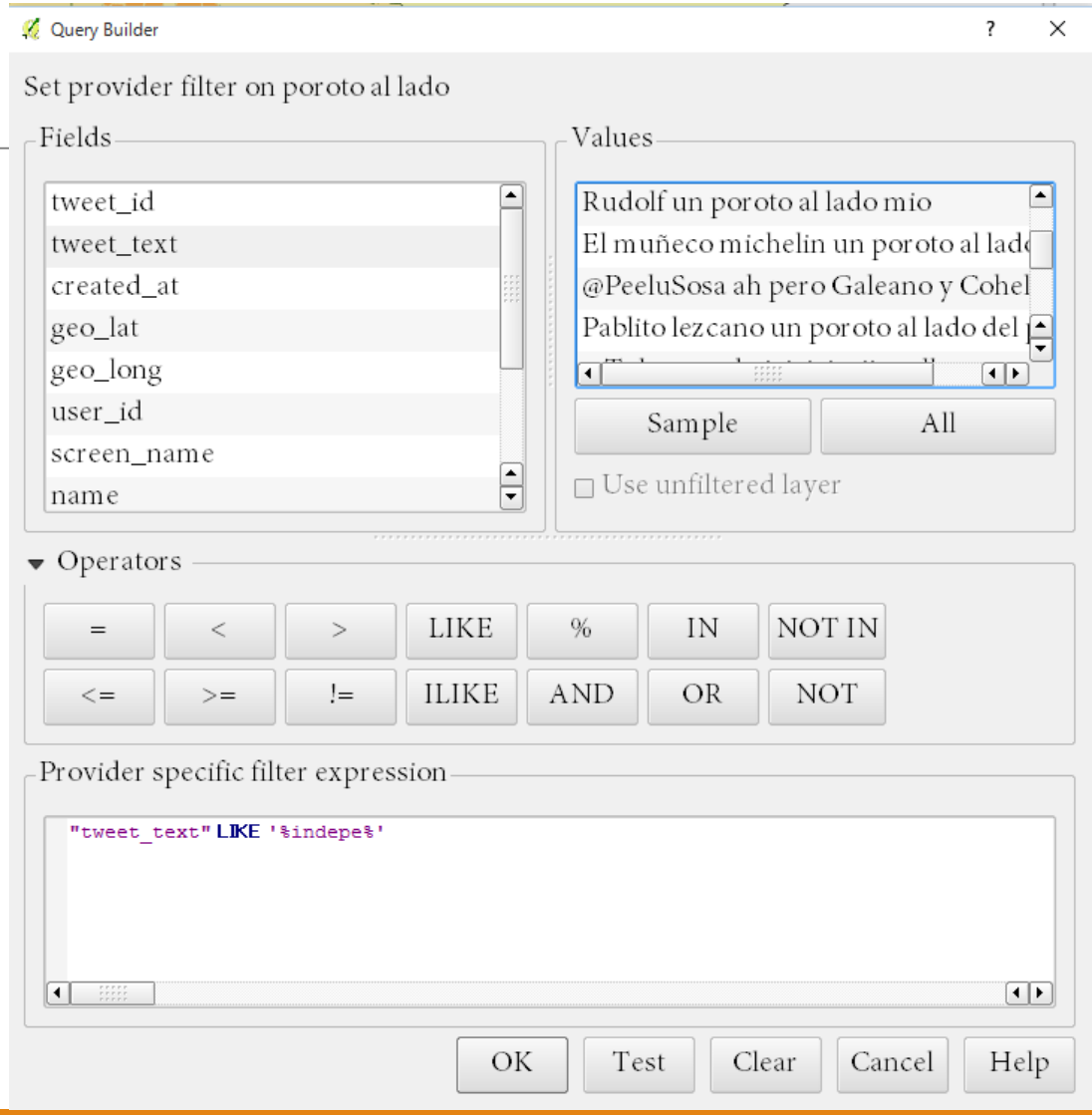
Interfaz de QGIS (1)



Interfaz de QGIS (2)



	tweet_id	tweet_text	created_at	geo_lat	geo_long	user_id	creen_nam	name	
0	6.09841…	BOOOOOLUUUDOOOOO TONCHI UN POROTO AL LADO DE ROJO DE COMO TRANSPIRA	2015-06…	-38.00178	-57.59978	710760193	AvluuE…	♡ Guid…	es
1	6.09992…	Alikemal y onur un poroto al lado de ustedes jajaja!!!	2015-06…	-34.65634	-58.79263	468294376	aguilera…	petu!!=)	es
2	6.10141…	@iorfischer el GPS un poroto al lado mio	2015-06…	-38.02174	-57.56115	574790866	Morena…	Morena …	es
3	6.10390…	El muñeco Michelín un poroto al lado mío SI	2015-06…	-26.77355	-60.44593	2392958…	belnorrego	ᴍ a ᴦ i…	es
4	6.10392…	Camiseta manga larga, polera de lana, campera del jardín y campera de corderito. Michelín un por…	2015-06…	-34.57541	-58.48827	330626757	Mariel…	Mariel …	es
5	6.10747…	El rey León un poroto al lado mio cuando me levanto	2015-06…	-34.54429	-55.87805	1938179…	Mariane…	Mariane…	es
6	6.10954…	@NaraEve genia ídola Crack Master chef un poroto al lado tuyo jajaja	2015-06…	-38.9493	-68.12906	379921365	Flecocho	Flori	es
7	6.11141…	Rudolf un poroto al lado mio	2015-06…	-31.42446	-64.17776	285410515	marinad…	Marina	es
8	6.11394…	El muñeco michelin un poroto al lado mio!	2015-06…	-35.55054	-63.37046	480422471	TettuLo…	•Steefv …	es
9	6.11729…	@PeeluSosa ah pero Galeano y Cohelo son un poroto al lado tuyo. Para cuando el book Pelu?	2015-06…	-34.94849	-57.95237	407454770	ccamico…	Cami Co…	es
10	6.11945…	Pablito lezcano un poroto al lado del petiso	2015-06…	-34.34599	-58.79927	2371250…	AlvaroD…	Alvarooo	es
11	6.11951…	@Toleeeeeedo jajajajaajaa elba un poroto al lado tuyo ajajai	2015-06…	-40.82538	-62.98047	1853511…	AbrilCo…	Jazmin.	es
12	6.12421…	Los Rompediskotecas un poroto al lado de nosotros, es una genia mi vieja JAJAJAJAJA	2015-06…	-42.75952	-65.04294	2171574…	juaan_m…	L'ARCH…	es
13	6.12726…	Que querés que te diga, tu ex es un poroto al lado mio ☺	2015-06…	-32.85551	-60.72528	2533418…	vanetlei…	Brisa Ya…	es
14	6.12826…	El Coco Basile un poroto al lado de mi voz #InviernoTeDetesto	2015-06…	-34.94849	-57.9524	407454770	ccamico…	Cami Co…	es
15	6.12935…	Si haría un twitter contando mis sueños sería súper popular, son re raros, la droga un poroto al lad…	2015-06…	-34.6188	-58.54549	353113243	negraalp…	La nagra ☻	es
16	6.13424…	2 goles picandola y uno de rabona, messi un poroto al lado mio	2015-06…	-34.83451	-58.39747	1937770…	emaAlca…	🕷Ema🕷	es
17	6.13461…	El polaco un poroto al lado de Fati jajaiajaia	2015-06…	-31.62528	-58.50223	737627934	BiancaIslaa	Bianca .	es
18	6.13551…	Mm alzada un poroto al lado tuyo mamu	2015-06…	-34.51603	-58.51363	184258463	Avluu_F…	❤DeM…	es
19	6.13910…	Una vieja es un poroto al lado tuyo jajajaja	2015-06…	-33.65037	-61.86796	2856102…	Caio_iunco	•Señor F…	es
20	6.13921…	@El_Alancito7 jajaja la interpol un poroto al lado de ustedes.	2015-06…	-30.98364	-57.9294	998128776	gabrielb…	🕷 🕷	es
21	6.14094…	master cheff un poroto al lado mio	2015-06…	-34.90441	-56.17637	236502568	Andrup…	Pinkie	es
22	6.14250…	La pantian un poroto al lado de algunas minitas de ahora. (No todas aclaro)	2015-06…	-38.98536	-64.10181	2315750…	Saavedr…	Braian e…	es
23	6.14255…	Policias en accion un poroto al lado de esto, que entretenido fue jajaja ay esta familiaaaa!	2015-06…	-37.28955	-59.1583	2329693…	MicaEtc…	#Noteva…	es
24	6.14584…	Barovero un poroto al lado de Ospina chh	2015-06…	-34.57463	-58.4819	1177294…	Tomaas…	Bustama…	es
25	6.14977…	Llego a salir con esta humedad mis pelos, el rey leon un poroto al lado mio	2015-06…	-37.34312	-59.14409	446362656	Magui_…	Magali C❤	es
26	6.15345…	@mechu09 zoz un poroto al lado nuestrooooy!!!! Jajajaja	2015-06…	-54.80716	-68.33925	349288145	Eelicastillo	✗	es
27	6.15742…	Quiero un novio como Ricky Martin q sea fuego d noche, nieve d dia #Cuak Alacran un poroto al l…	2015-06…	-34.7302	-58.42837	852172518	Jesilui	Jesi:)	es
28	6.16380…	El reno Rodolfo un poroto al lado mio.	2015-07…	-32.17764	-64.24519	111107235	noelovgfan	Núecita …	es
29	6.16439…	Gran hermano un poroto al lado nuestro eh	2015-07…	-45.86546	-67.51743	2266466…	Macare…	Maca#16	es
30	6.16463…	Bad Blood un poroto al lado del video de #BBHMM. Rihanna sabe como hacerlo. Es Rihanna. Y est…	2015-07…	-32.49425	-58.2277	259843353	_BadGal…	#BBHM…	es
31	6.16842…	Mix tail un poroto al lado de la sangria de maggi	2015-07…	-28.46555	-65.8048	334882994	LizAgos…	Liz	es
32	6.17085…	Que mal que me siento, miss pelotuda un poroto al lado mio	2015-07…	-34.64562	-58.78921	1149772…	WayraG…	Chayra	es
33	6.17129…	Cupido un poroto al lado mio eh @ionilozowski jajaiajaiai	2015-07…	-34.7565	-58.43677	1586741…	Juliett_…	ρ α▯α …	es
34	6.17438…	Goku un poroto al lado de medel. http://t.co/DggPS80hqV	2015-07…	-34.52329	-58.77317	2965092…	lucascis9	#Vamos…	en
35	6.17630…	La heladera un poroto al lado de mi pieza	2015-07…	-34.49552	-61.54953	1656100…	NickoooooR	Nico:3	es
36	6.17866…	Axel un poroto al lado mio	2015-07…	-32.72023	-59.38951	2319905…	Eliasbog…	C.A.R.P ツ	es
37	6.18105…	Rial un poroto al lado mio	2015-07…	-34.60863	-58.55201	2222253…	RommiCae	Romi ♡	es
38	6.18688…	Pakis me hizo enojar y le tube que mandar un audio y mis audios son un asco el charango un porot…	2015-07…	-34.78893	-58.49453	443400635	AgosMa…	Que lo p…	es
39	6.18936…	Maru botana y la ganadora de máster cheff un poroto al lado mio	2015-07…	-32.40721	-58.27984	1358083…	Luucho…	luciano	es
40	6.18960…	Sherlock holmes un poroto al lado mio	2015-07…	-34.59937	-58.39685	189351322	GittanaB	Amelita	es
41	6.19331…	Ay amo a la sole porque esta llorando desde las 21:00 mas o menos es un poroto al lado de nazaren…	2015-07…	-35.65937	-63.73714	1460847…	Mil_Matos	Mila nesa	es
42	6.19518…	@MilaagrosLopez @SilviPerez_ Olvidate master chef un poroto al lado tuyo	2015-07…	-36.75397	-62.50404	1703551…	OrneCa…	Ornella	es
43	6.19586…	Casillas un poroto al lado mio	2015-07…	-34.5754	-58.43637	2150243…	Segundo…	Agustin …	es
44	6.19587…	maru botana un poroto al lado mio	2015-07…	-41.13441	-71.31504	1158717…	JooFarello	josefa	en-
45	6.19646…	@Candela95266187 pfff te olvidas que master chef un poroto al lado mio?	2015-07…	-26.39778	-54.61369	2424495…	Angiear…	Gringa♡	es
46	6.19674…	Luks los de dar la nota son un poroto al lado tuyo	2015-07…	-40.81595	-63.01389	1025476…	nicolasu…	Nico Ur…	es
47	6.19689…	Un mapache era un poroto al lado de el...	2015-07…	-33.00786	-58.5224	437023275	anabella…	Old soul	es
48	6.20293…	Hernan piquin un poroto al lado de ellos.	2015-07…	-31.34213	-63.94675	2795712…	Antonell…	Danonin…	es
49	6.20460…	el indio un poroto al lado mio , estoy mas solari q alfonsin are :p	2015-07…	-34.6234	-58.42141	783137491	AdrianE…	imparab…	es
50	6.20646…	El parque de la ruta 9 un poroto al lado del súper park	2015-07…	-31.34213	-63.94675	2795712…	Antonell…	Danonin…	es
51	6.21072…	Los de master chef son un poroto al lado mio	2015-07…	-34.89737	-58.38582	336620009	JoelKraus1	Joel K	es
52	6.21162…	Que buena soy dando consejos, maetro amor un poroto al lado mio jajajaja	2015-07…	-27.45756	-58.98167	2770546…	PeronMi…	Micaela …	es
53	6.21493…	Imaginame a mi con anteojos. O.o Patito feo un poroto al lado mio, pero los necesito URGENTE!!!	2015-07…	-35.6208	-59.76624	1853140…	miica_as…	MiicaAs…	es
54	6.21655…	Invernalia un poroto al lado del frío que tengo.	2015-07…	-32.48614	-58.24084	208193606	ShuliaG…	Shulia G…	es
55	6.21866…	Canguro fest un poroto al lado de Show Match jajajaiaja estan llenos de pibitos	2015-07…	-34.60856	-58.55202	2222253…	RommiCae	Romi ♡	es
56	6.22221…	Soy un poroto al lado de los de MasterChefflr .	2015-07…	-36.8903	-60.31077	2920673…	GonzaC…	Gonzalo…	es
57	6.22569…	@marcooviedo20 xipolitakis un poroto al lado mio la mejor modelo del negocio	2015-07…	-31.3014	-64.30561	1852377…	valenmu…	Valentina	es
58	6.23494…	No lo supero mas es un potro, también un poroto al lado de otros !	2015-07…	-34.08772	-56.2336	2354992…	AbriilM…	▯Abii▯	es
59	6.23744…	Brian sin suerte un poroto al lado mio!	2015-07…	-31.73157	-64.99837	1265333…	Yazmin…	Yazi	es
60	6.24042…	Un cactus un poroto al lado de Gus	2015-07…	-27.41137	-55.94315	1120399…	agus_pe…	Pedrozo…	es
61	6.24060…	messi un poroto al lado tuyo fines mori	2015-07…	-37.38245	-64.59293	2153915…	EnzoLH…	Enzo	es

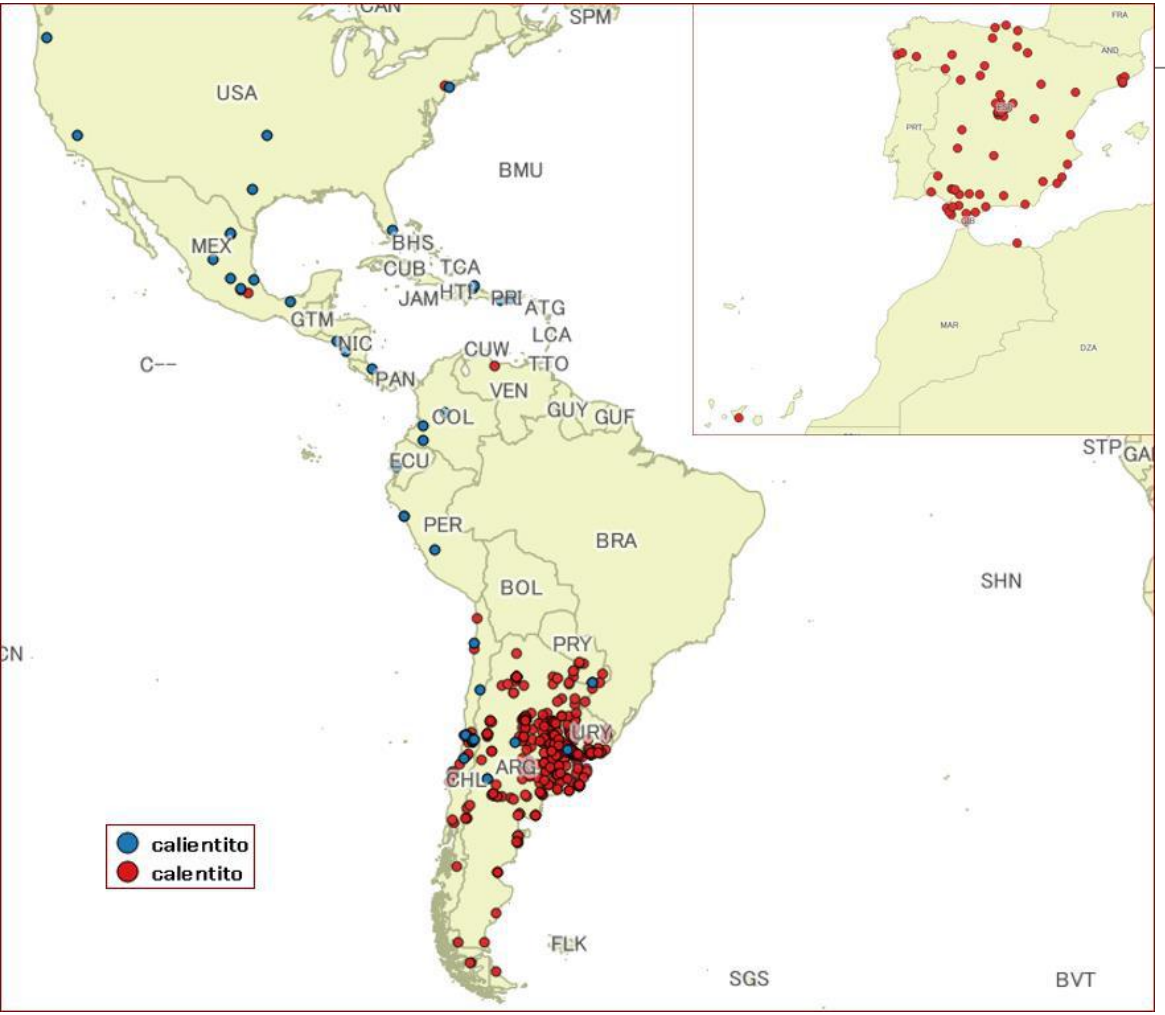


Morfología

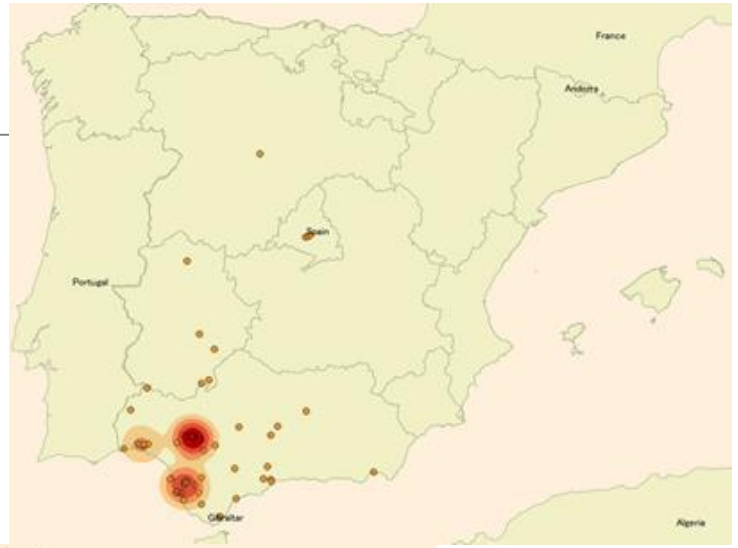
Argentinismos
quilombo & un poroto al lado mío

Morfología

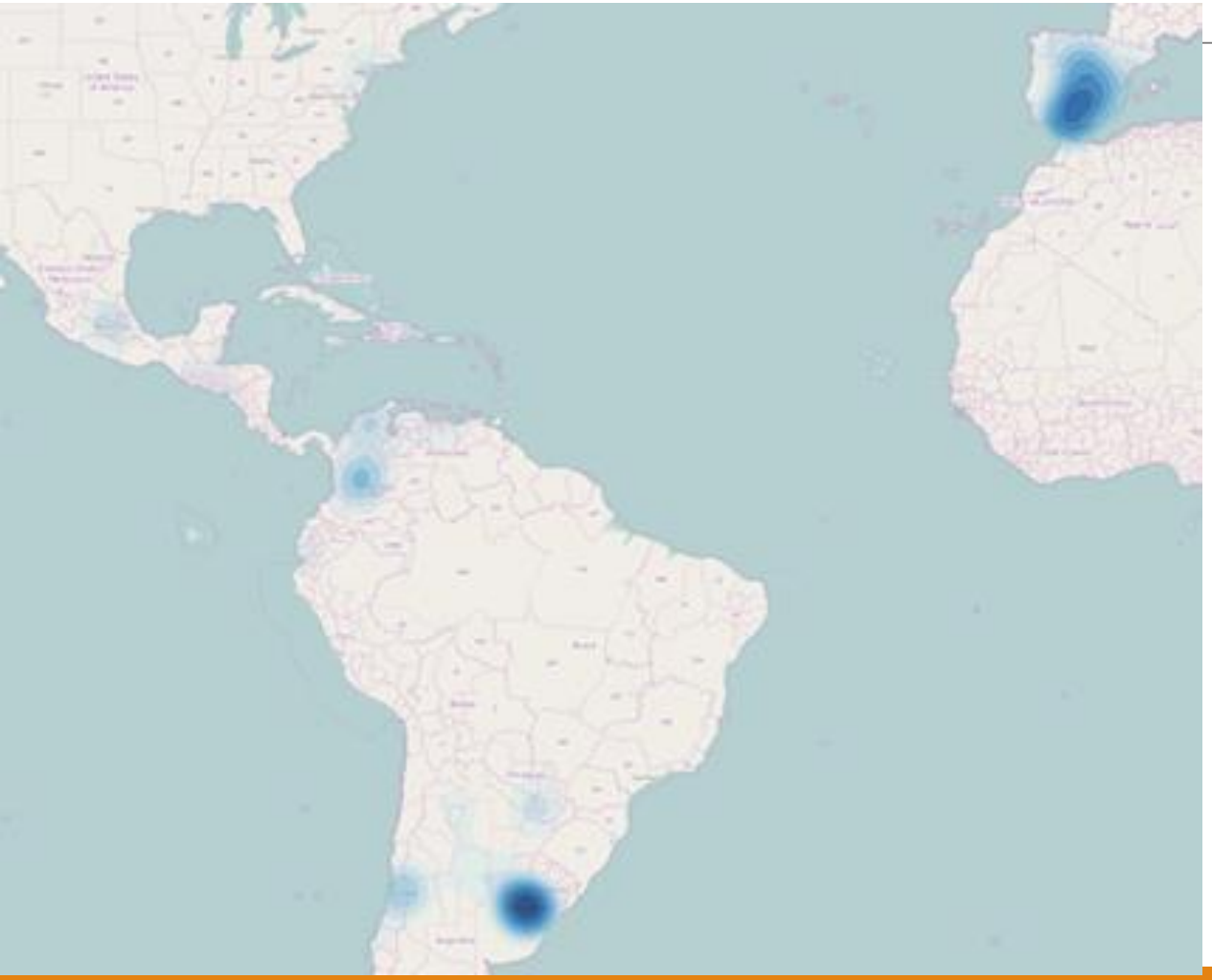
calentito vs calientito



la calor (femenino)



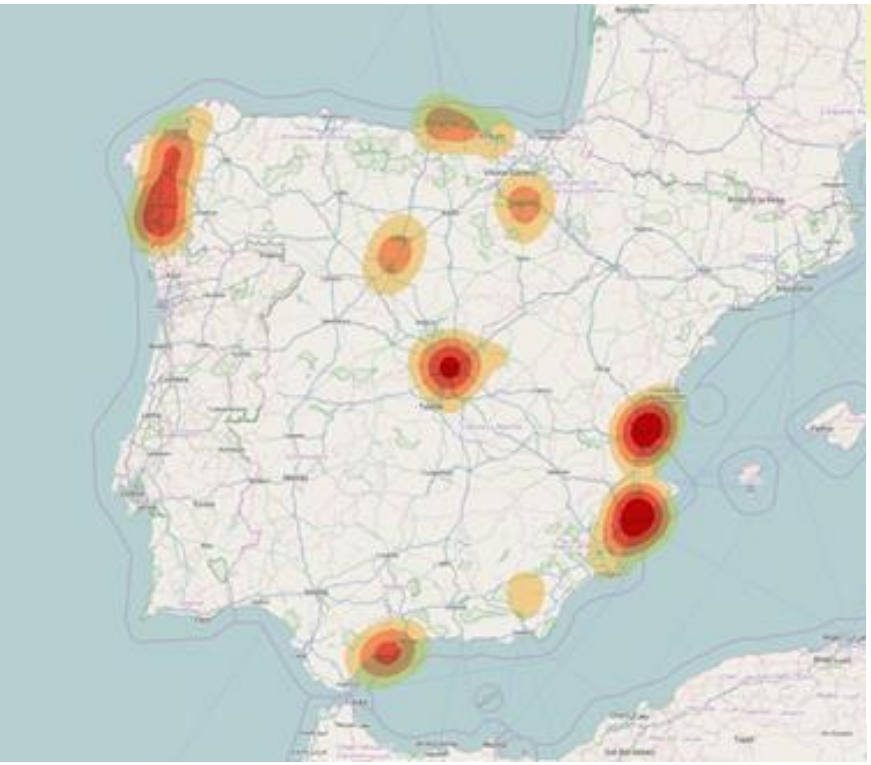
Morfología quizá vs quizás



Cada punto del mapa puede mostrar la información contenida en la base de datos.

tweet_id	514372603302404000
tweet_text	Una desgracia tener que entregar la maqueta con esta lluvia.. La llene de bolsas con tal de que no se moje ah
created_at	23/09/14 07:15
geo_lat	-38.00485
geo_long	-57.54692
user_id	150783054
screen_name	Matiaspintoss
name	Mati PintosLarrondo
location	Mar del Plata
description	El éxito esta en demostrar que se puede Coordinador de Extensión de los Clubes Leo del Sur Argentino.
followers_count	276
friends_count	292
statuses_count	29647
street	Corrientes
locality	Mar del Plata
postal_town	null
admin_level_3	null
admin_level_2	General Pueyrredón
admin_level_1	Buenos Aires
postcode	null
country	Argentina
preposicion	con tal
que	de que

Otros

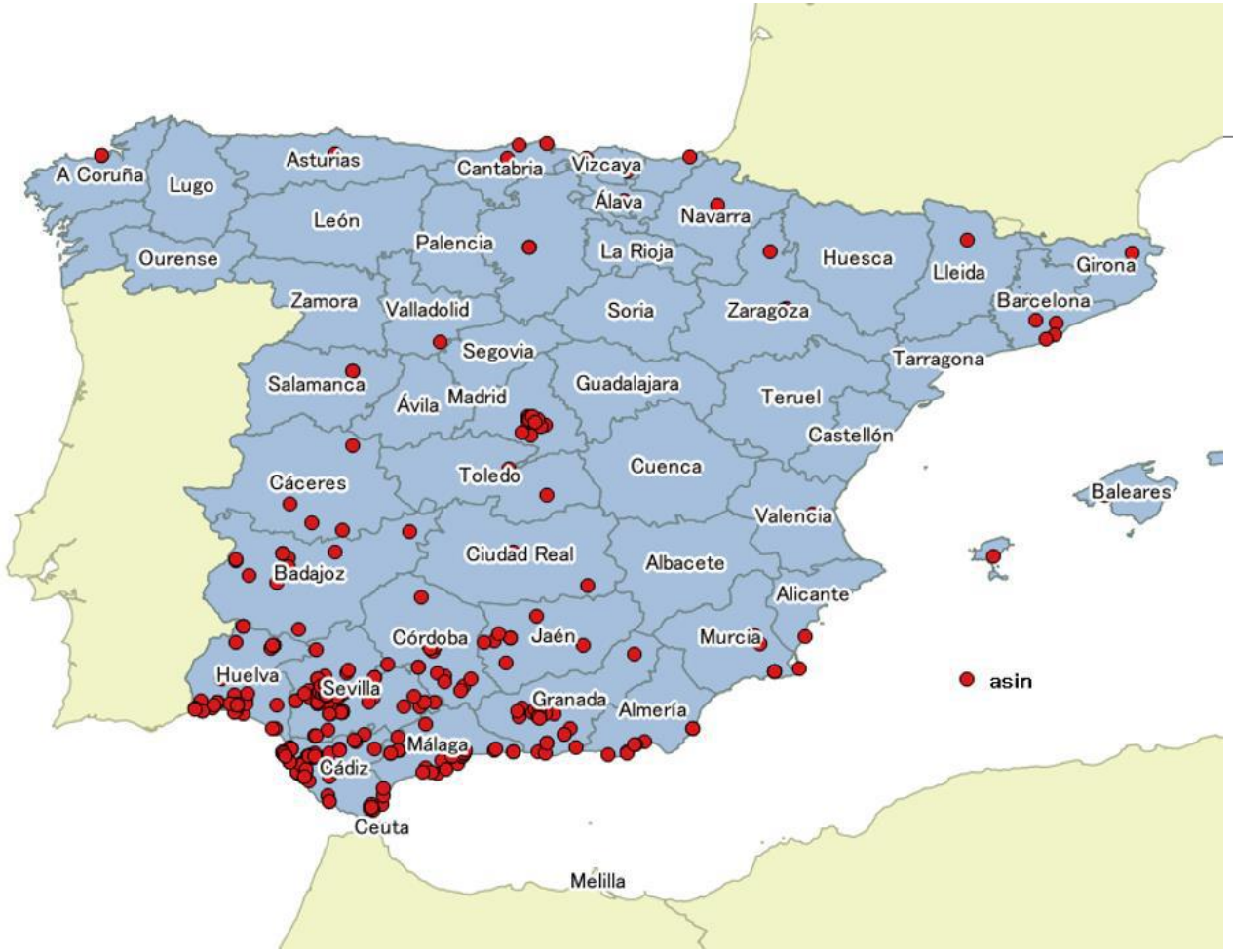


de la hostia

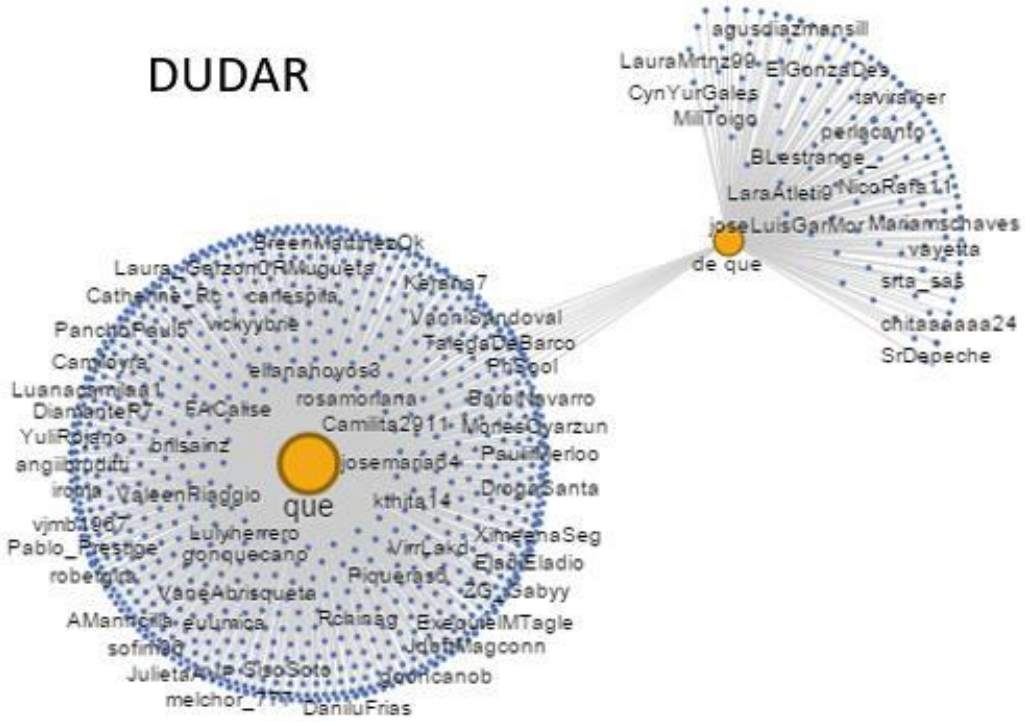
sushi

un poco bastante / demasiado

Otros

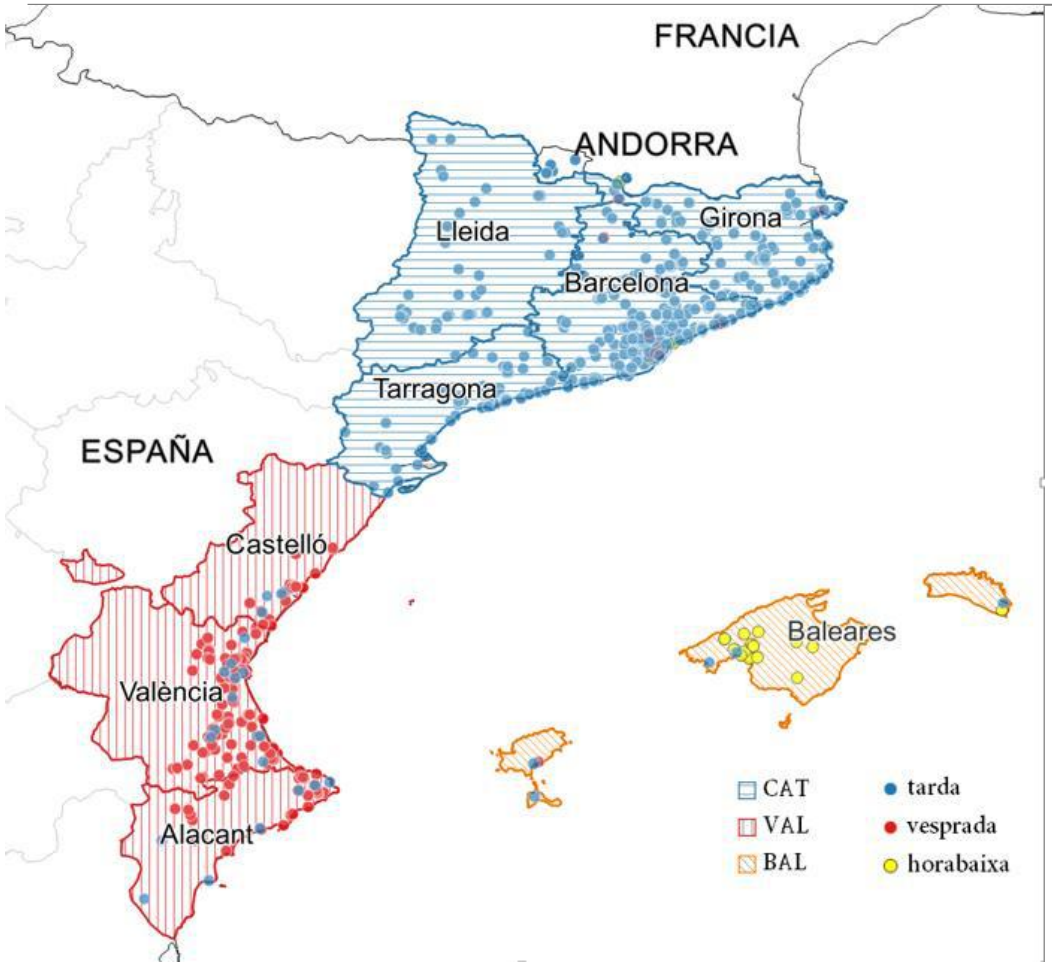


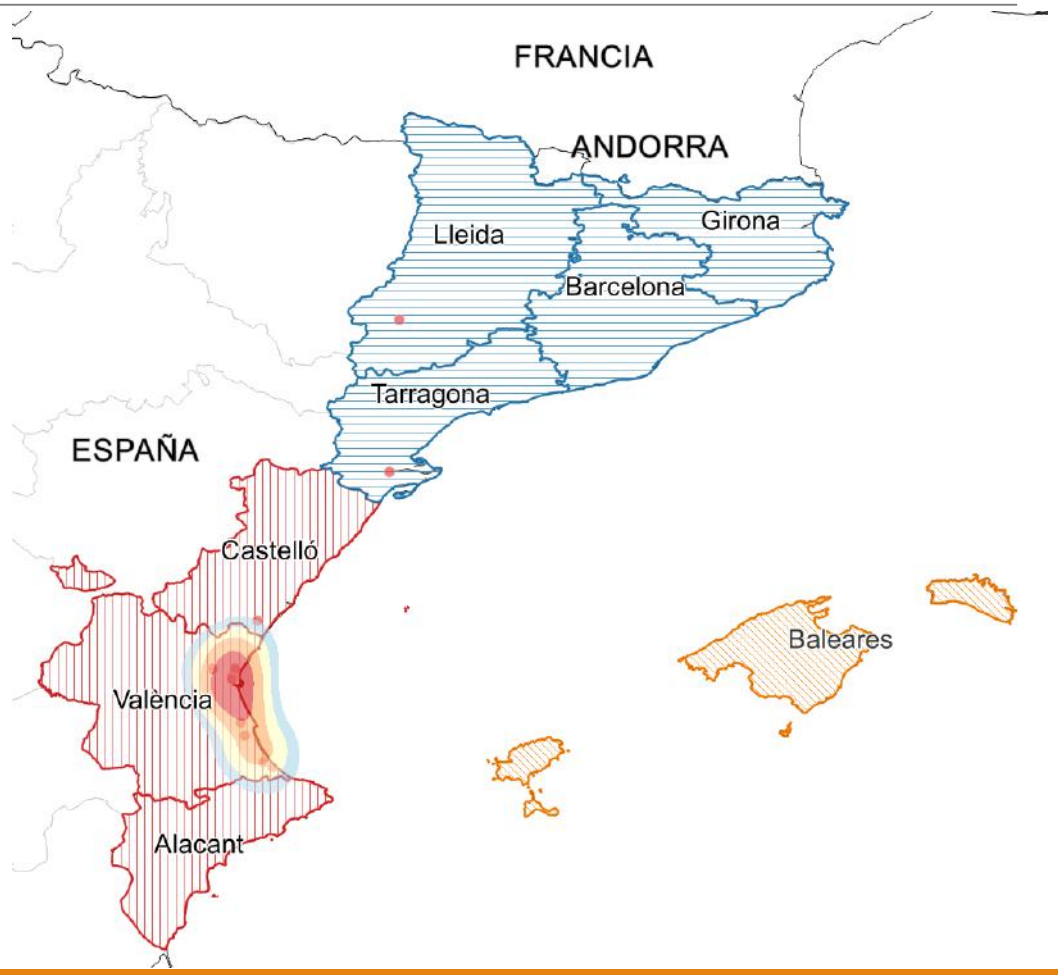
asín



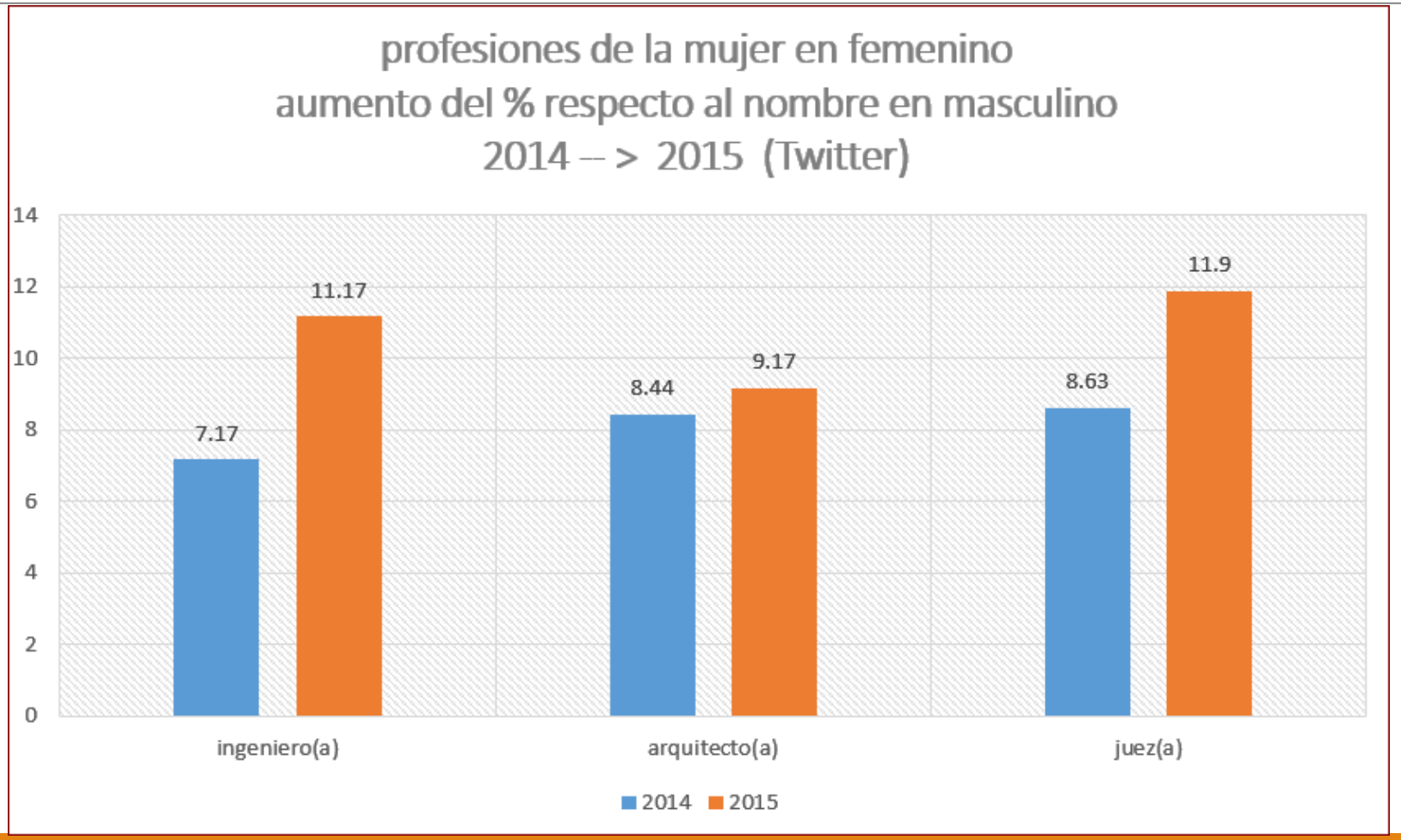
dequeísmo

Shapefiles: Puntos, líneas y polígonos



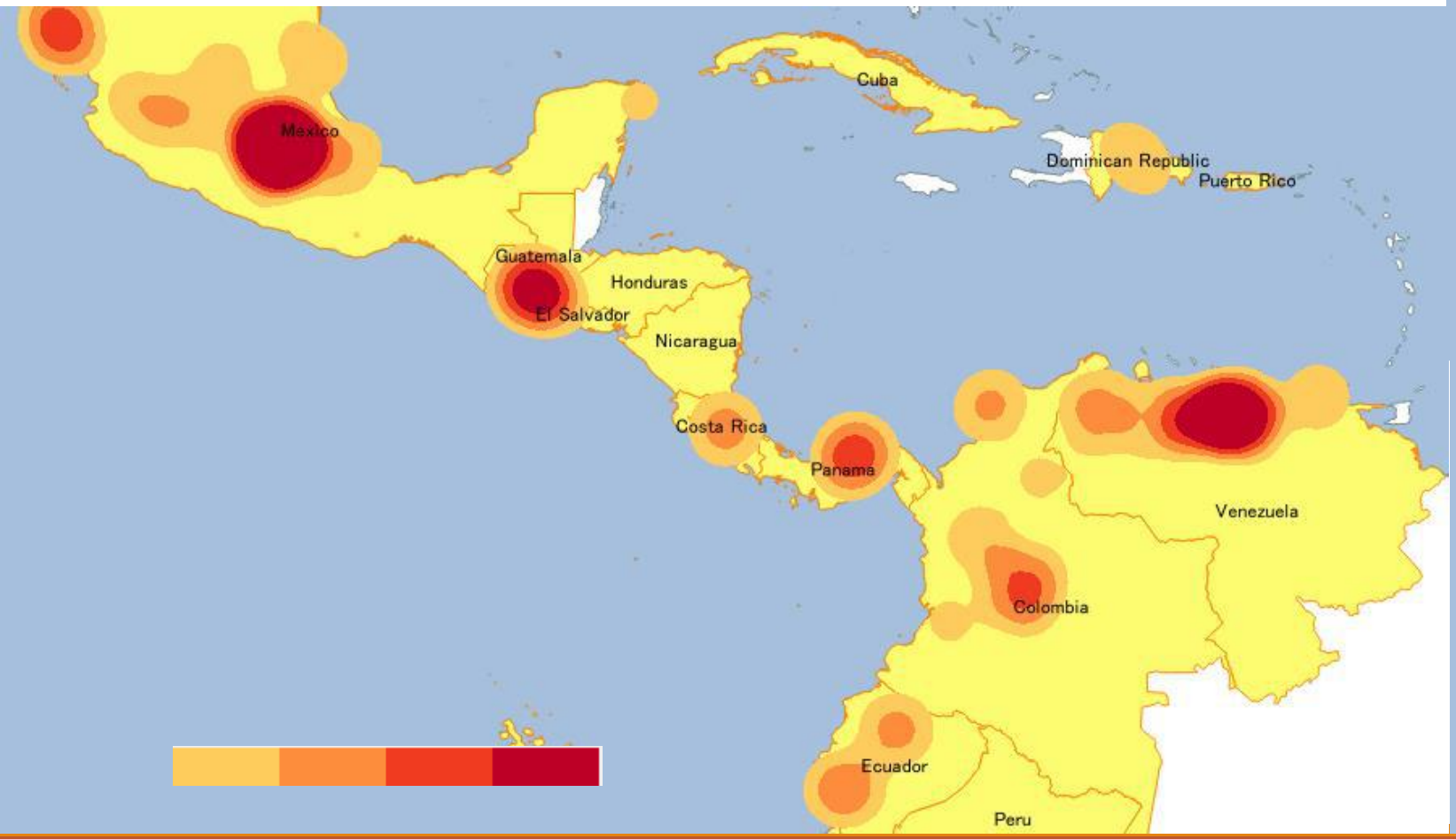


Variación en un período de tiempo



Habemos

(215 ocurrencias)



Ejemplos de uso de *habemos*

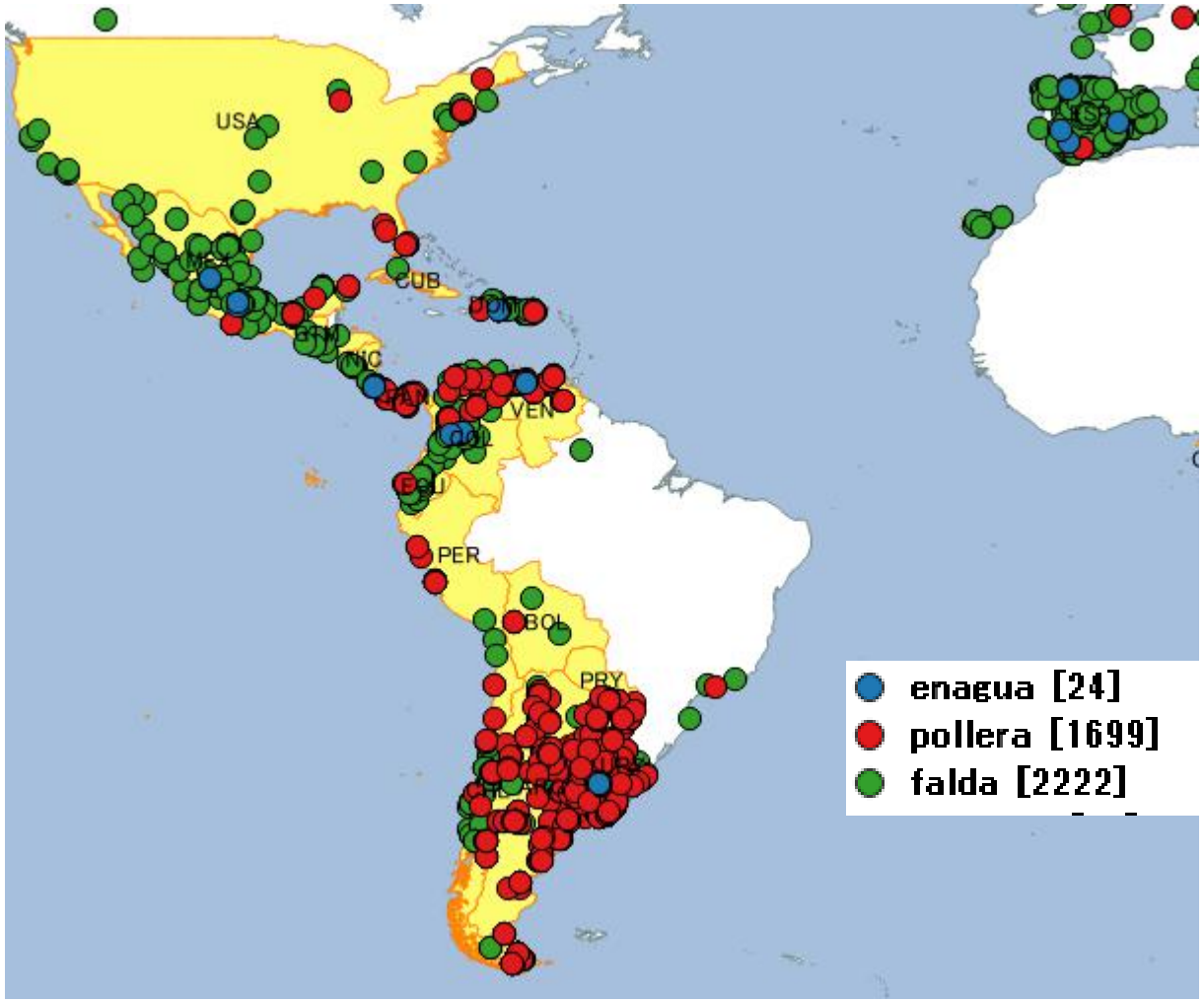
@KIkautsky "habemos sido" o no "habemos sido" esa es la incuestion ;)	Spain	Cataluna	Barcelona	n.a. (29)
@ManuDfs mas respeto Manolito habemos mujeres que nos instalamos a ver los partidos tb eh jajajajajaa	Ecuador	Guayas	Guayaquil	Guayaquil
@DatitosRivera nose vale el balon para ella porque habemos otros q respondimos mas rapidos y con todo y ella por equivocacion se lo lleva	Panama	Panama	Panama	Juan Diaz
@JohnTown_ pues yo no se que tiene tu amigo. Pero la verdad es que NO habemos suficientes especialistas en el pais. Y por eso IESS esta asi.	Ecuador	Guayas	Guayaquil	Guayaquil
@elmetrodepanama son las 5:30am y todavia no abren las puertas (habemos muchos esperando para transportarnos) http://t.co/Exw2a2YzTB	Panama	Panama	San Miguelito	Victoriano Lorenzo
Chicos entiendan: si habemos chica que disfrutamos de un alucinante partido, cuando por casi meten gol o cuando el portero la ataja...	Panama	Colon	Chagres	Nuevo Chagres
O cuando el equipo a quien le vas anota y obtiene la victoria, habemos chicas que si nos apasiona el deporte tal y como es #FootballIsMyLife	Panama	Colon	Chagres	Nuevo Chagres
Habemos unos cuantos, los tipicos que hemos visto esta pelicula, y siempre solteros, que penica damos xD	Spain	Region de Murcia	Murcia	Huera de Murcia
Habemos seis mas la profe en clase xD	Chile	Araucania	Cautin	Nueva Imperial
Insolito, 15 min esperando comprar el bono en @ClinicaServet mismos min q la doc lleva sin atender pacientes, y habemos 3 al menos afuera!	Chile	Region Metropolitana (	Santiago	Santiago
@FEFecuador vamos tri! Todavia habemos quienes confiamos en ti..!	Ecuador	Pichincha	Quito	Cumbaya
A los del metro se les olvida que en Verano habemos quienes curramos. #putometro	Spain	Comunidad de Madrid	Madrid	n.a. (177)
Solo habemos 2 en el salon, JAJA:(	Panama	Panama	Panama	Ancon
@jocanavarro ahora culpa es de economia mundial...pensaran que no habemos exportadores con acceso al mundo que sabemos que eso no es cierto?	Chile	Region Metropolitana (	Cordillera	San Jode de Maipo

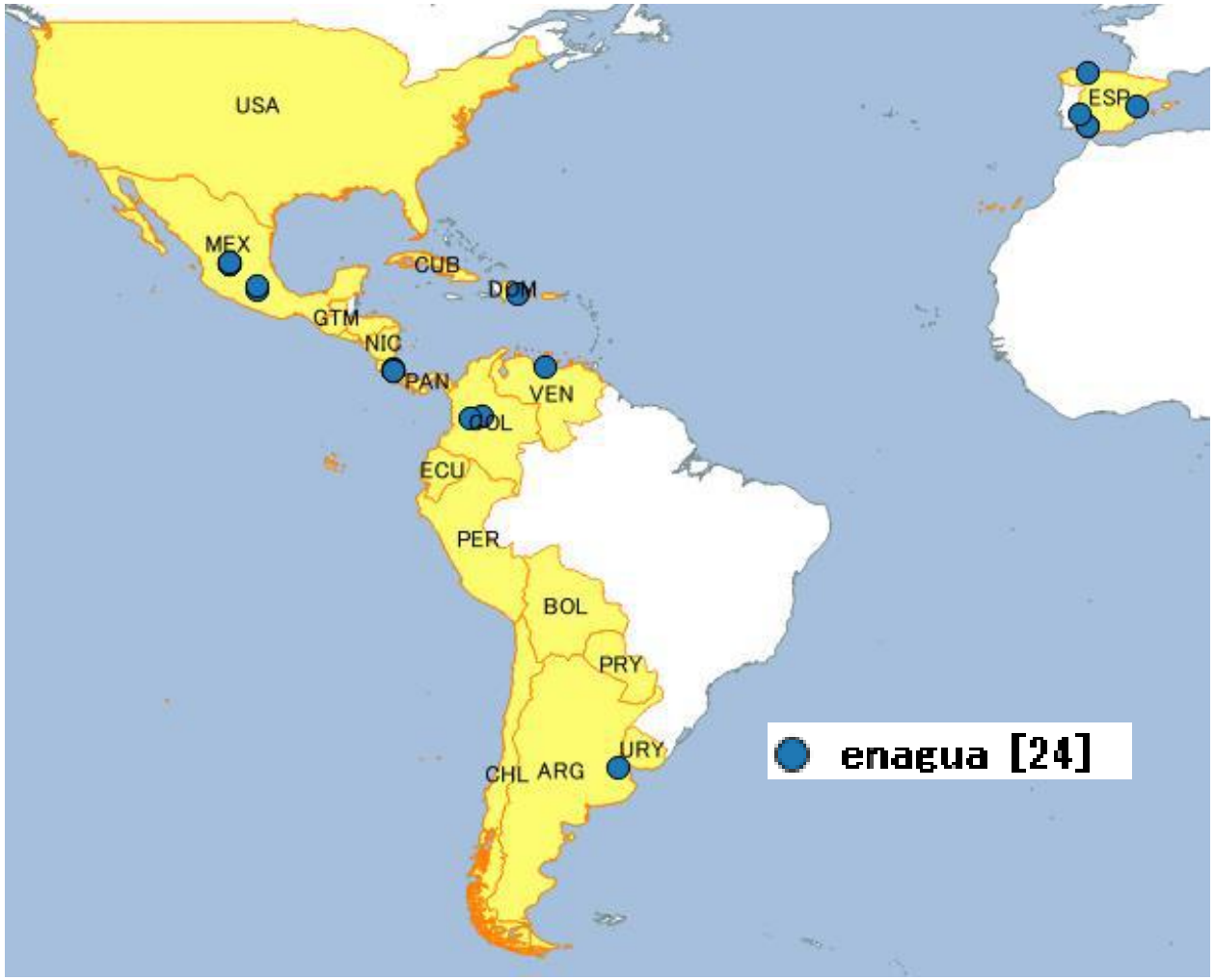
Solo habemos 2 en el salon, JAJA:((Panamá)

@Daniel61sneros habemos personas que estamos en zonas de riesgo que necesitamos ser reubicados o por lo menos saber si son o no son solares (Ecuador)

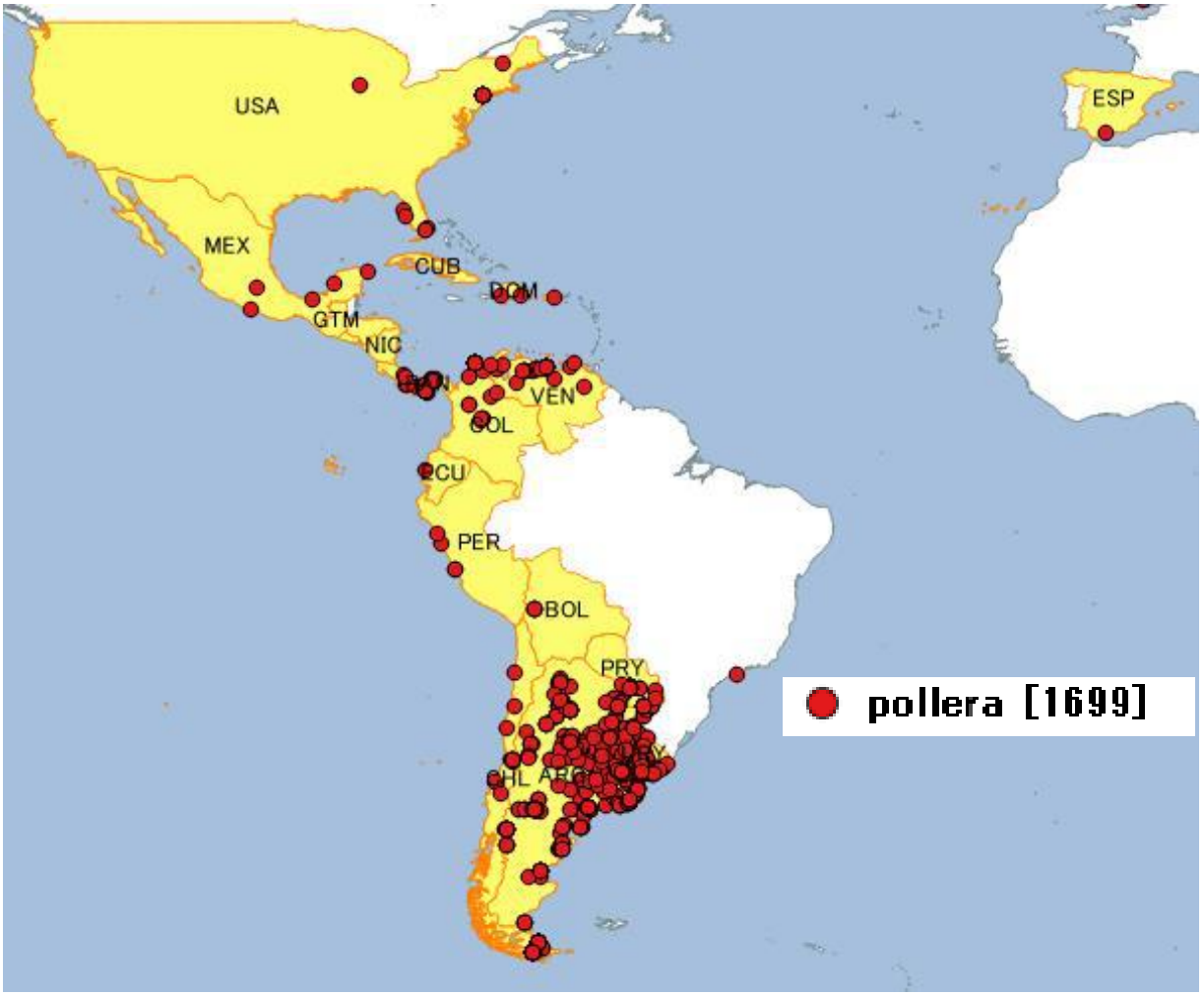
La salud mental es pesima en Chile, habemos muchos enfermos silenciosos (Chile)

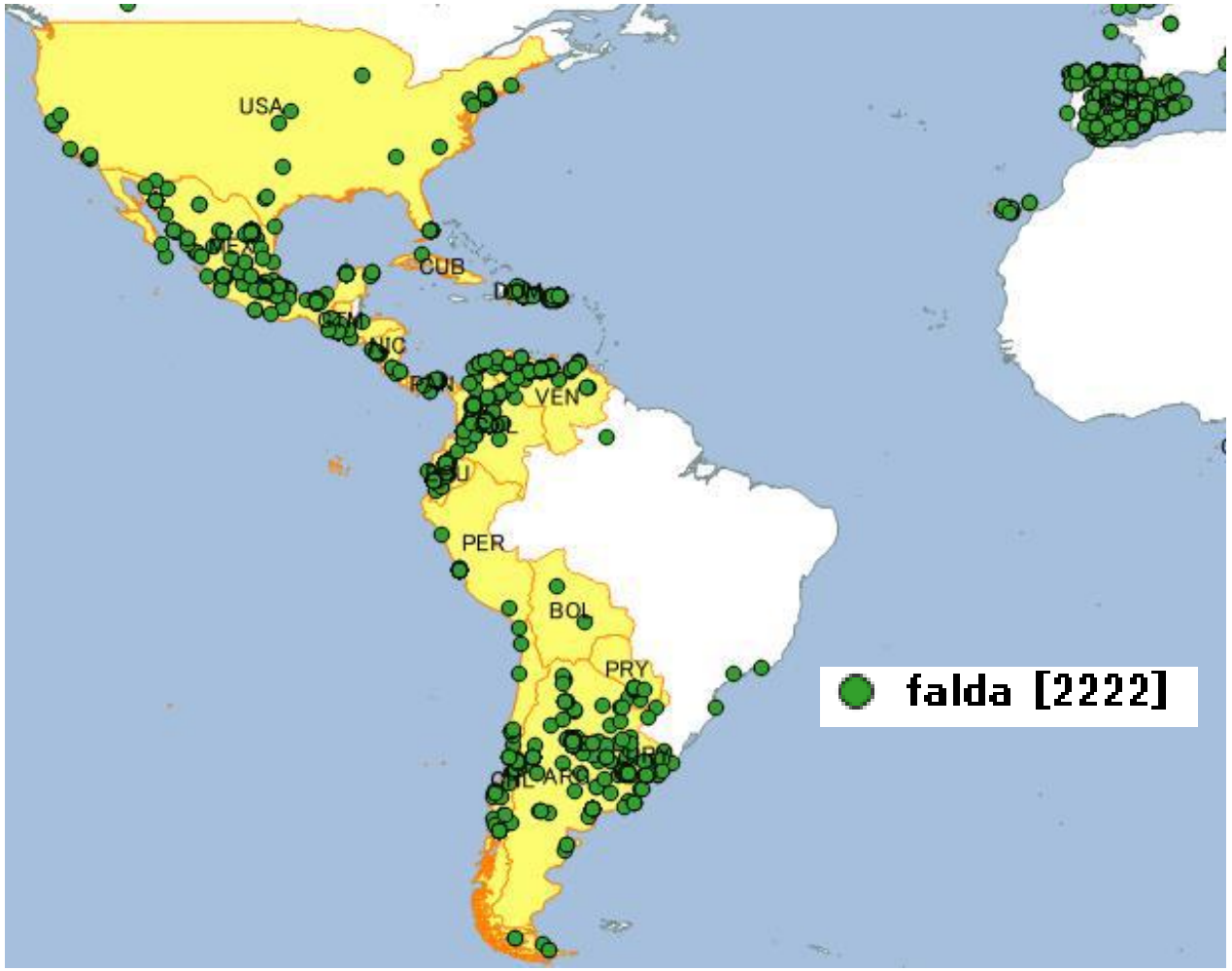
falda --- pollera -- enagua



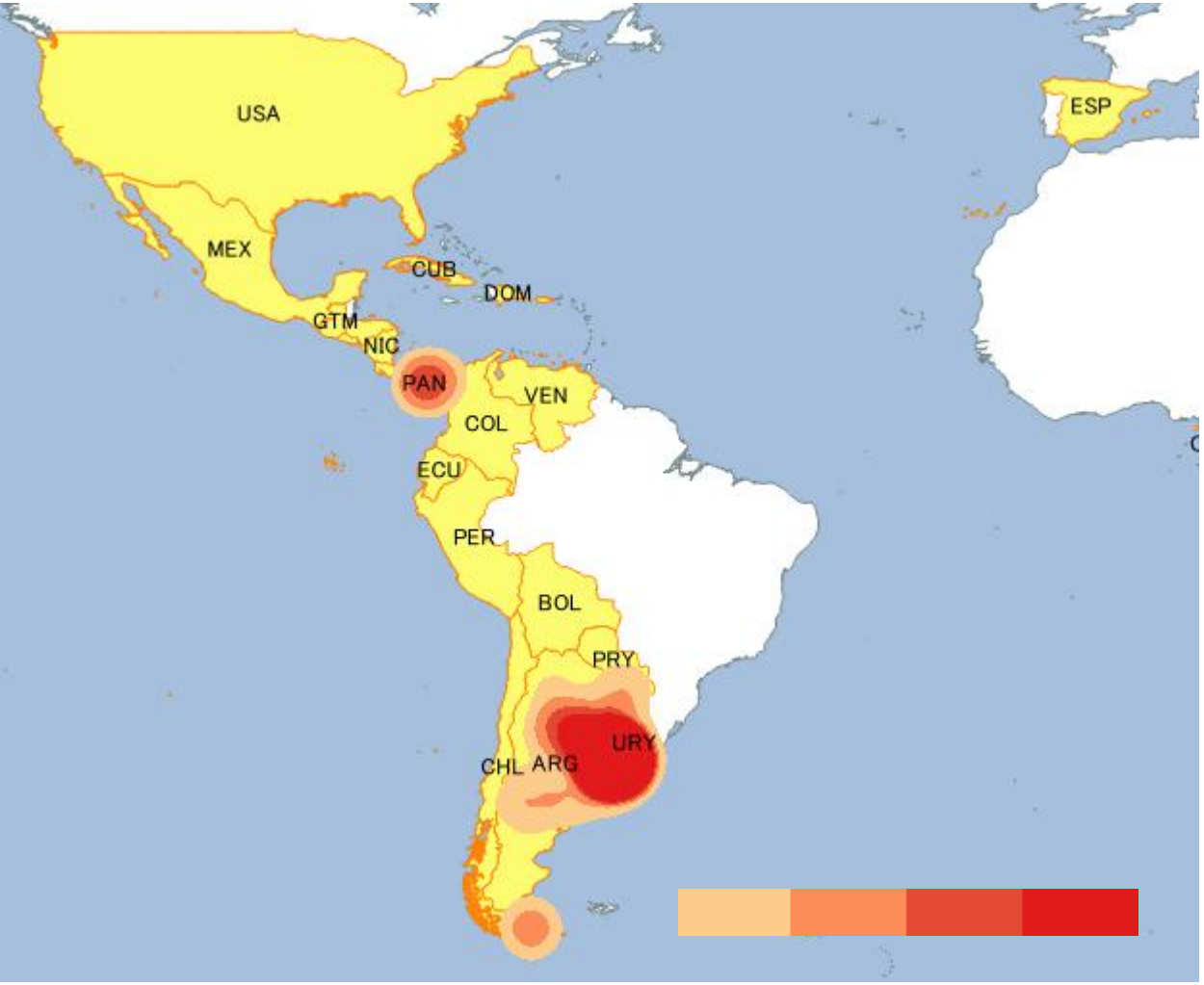


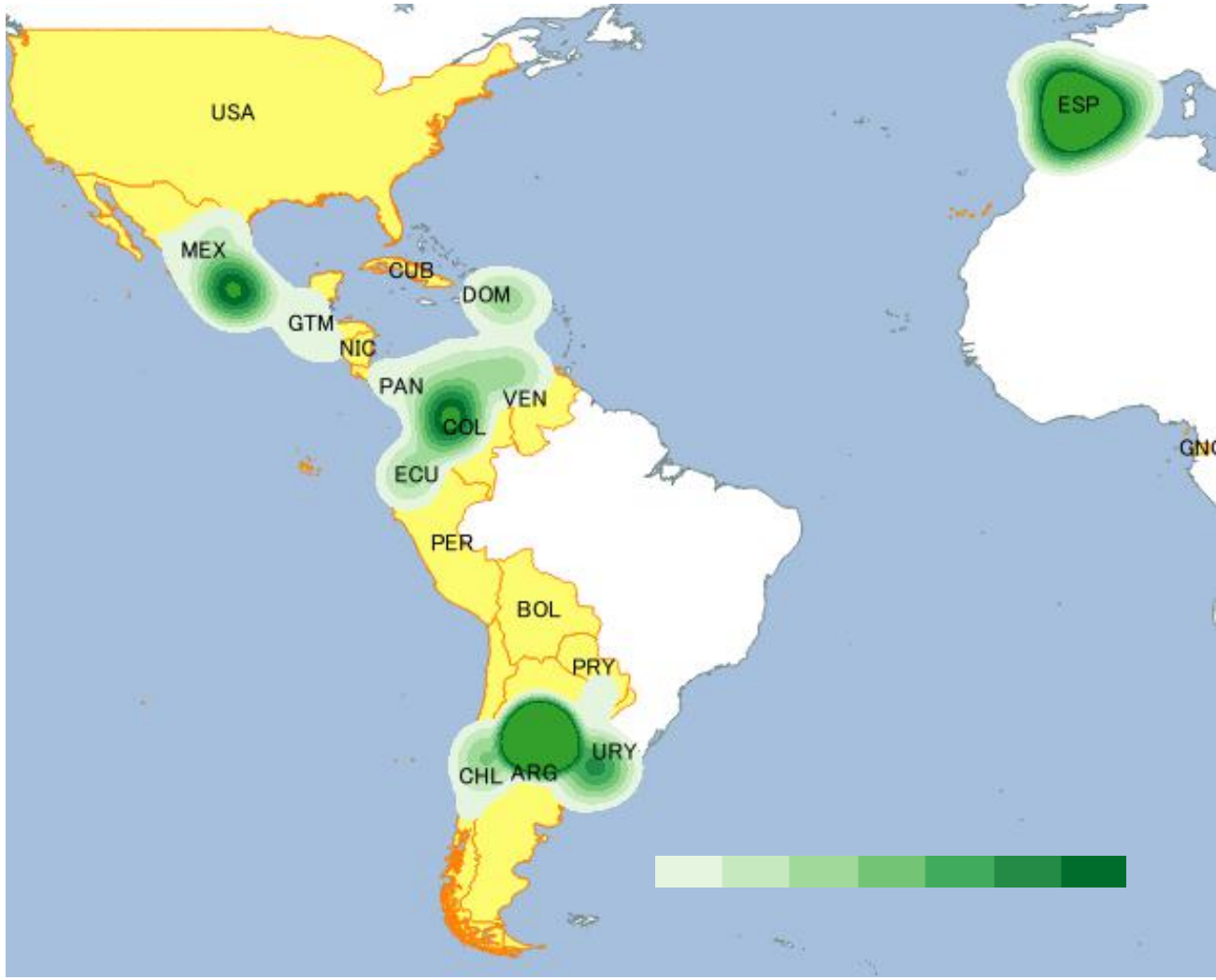
pollera --- falda



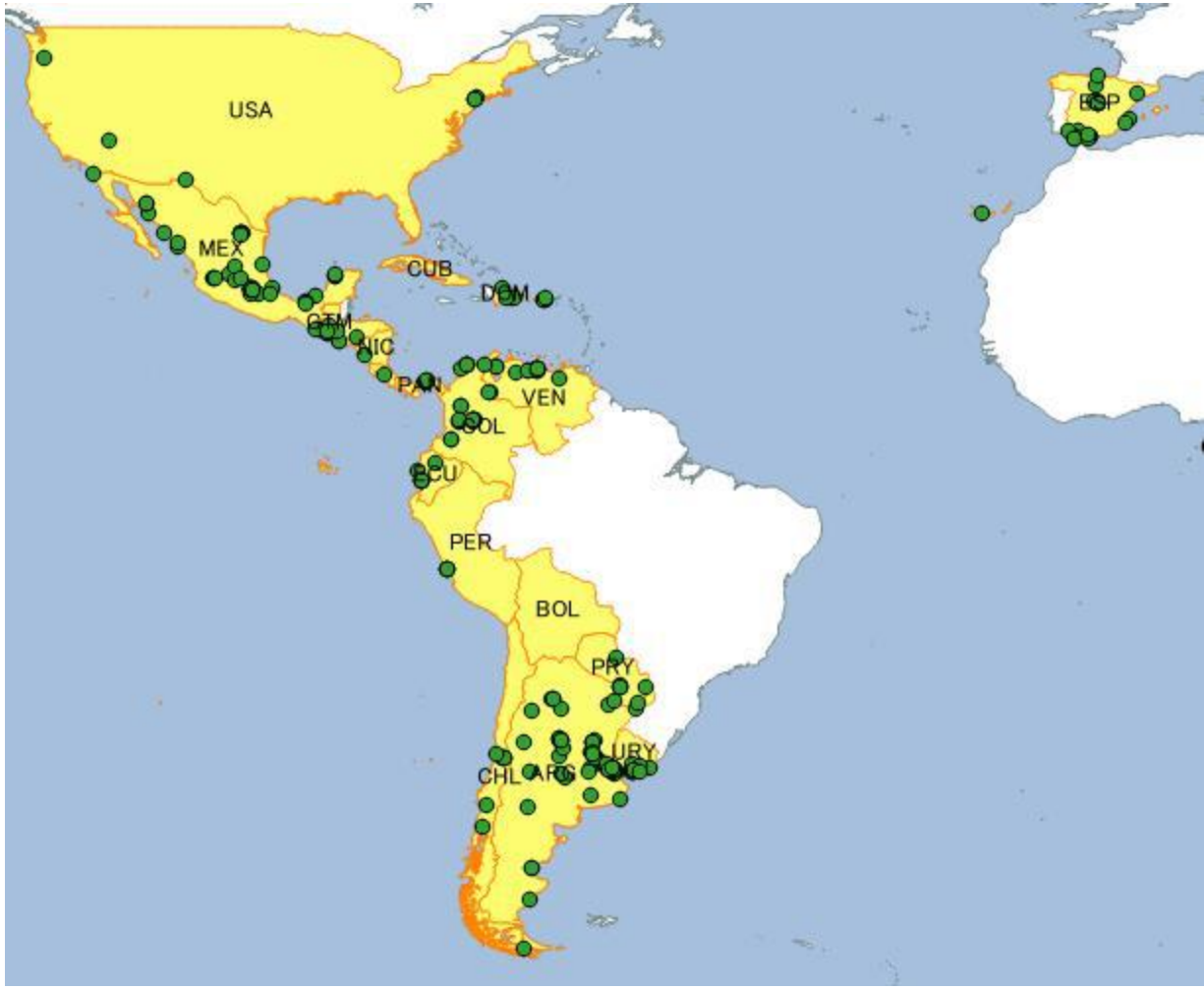


pollera --- falda

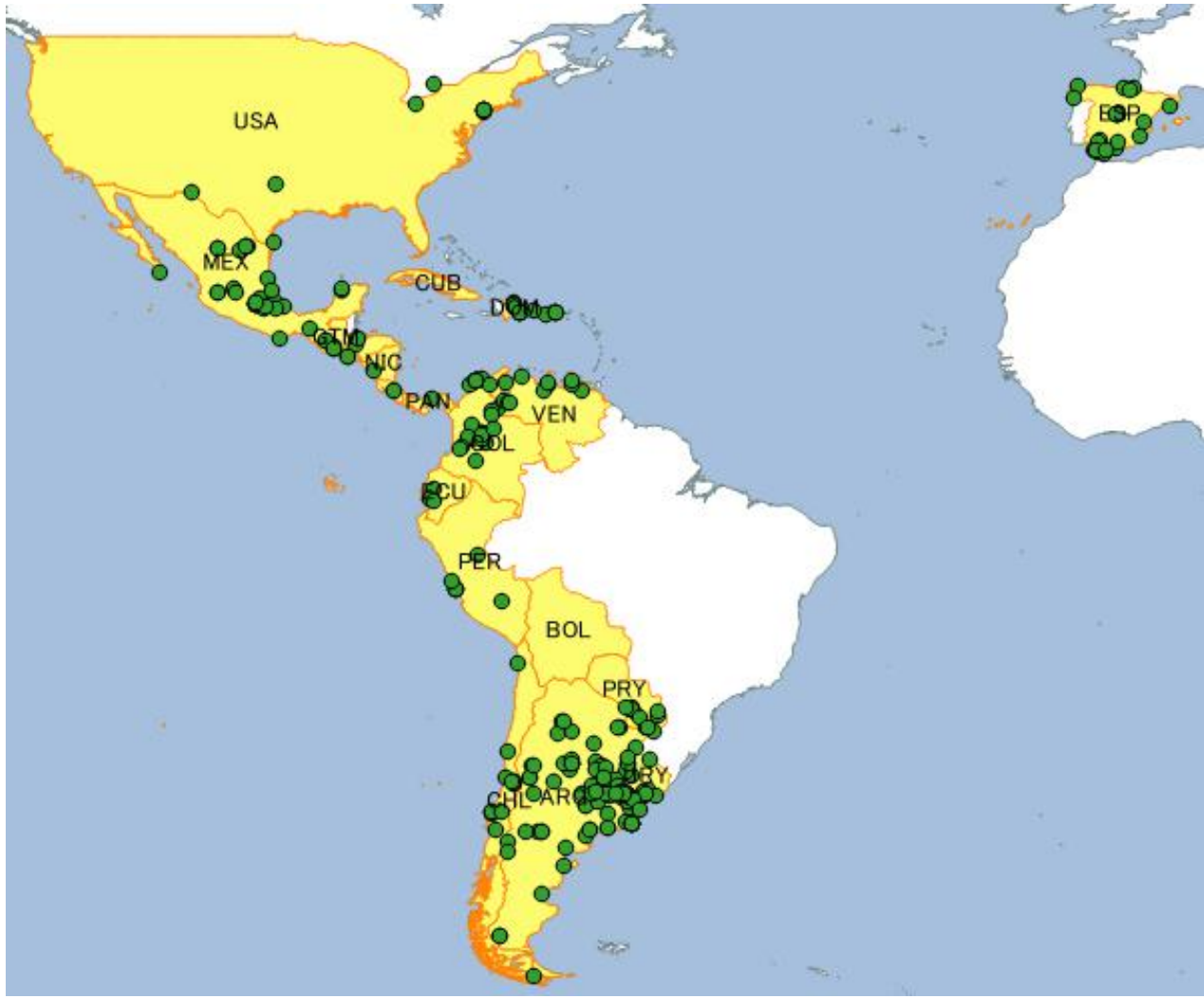




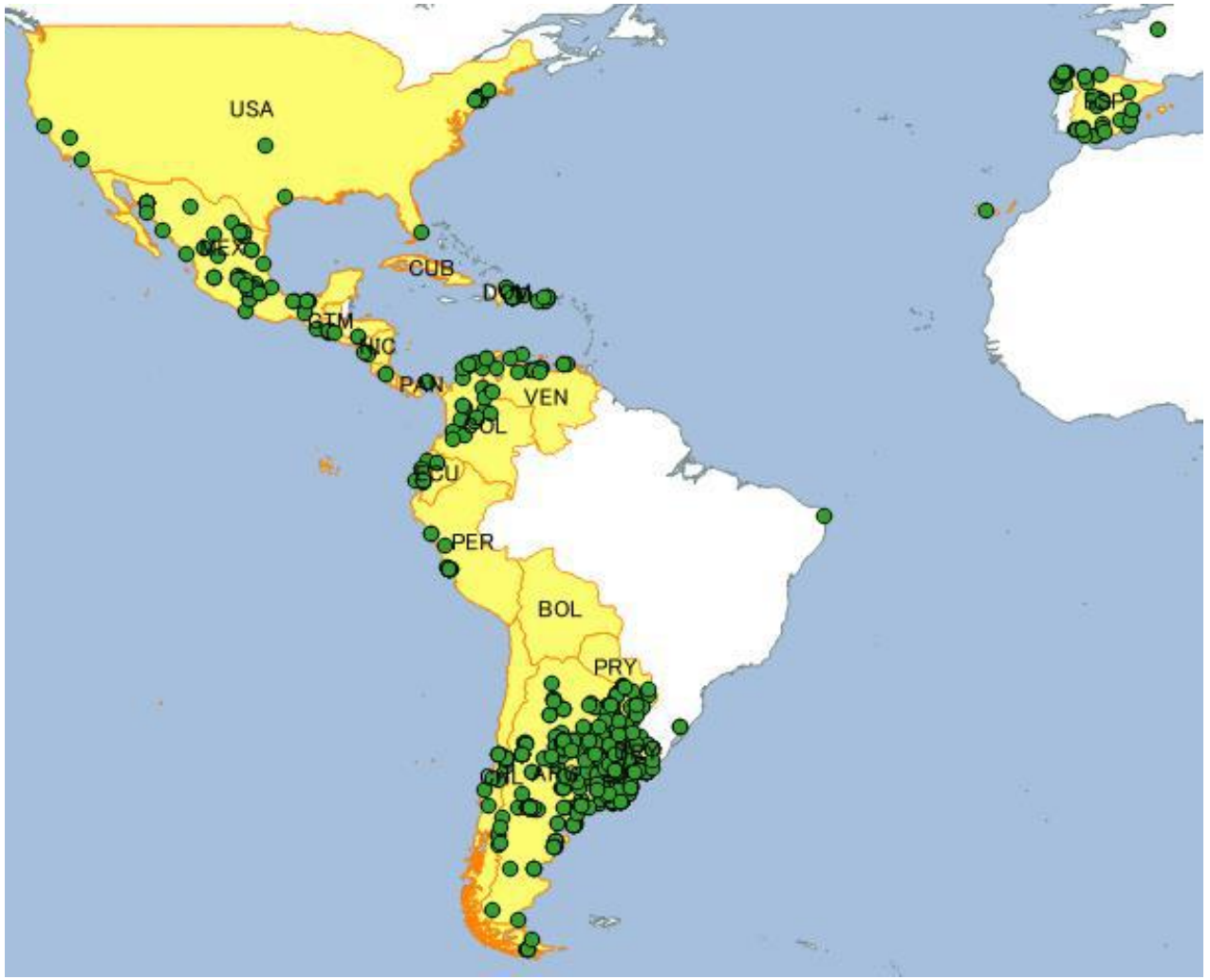
vivistes --- viviste



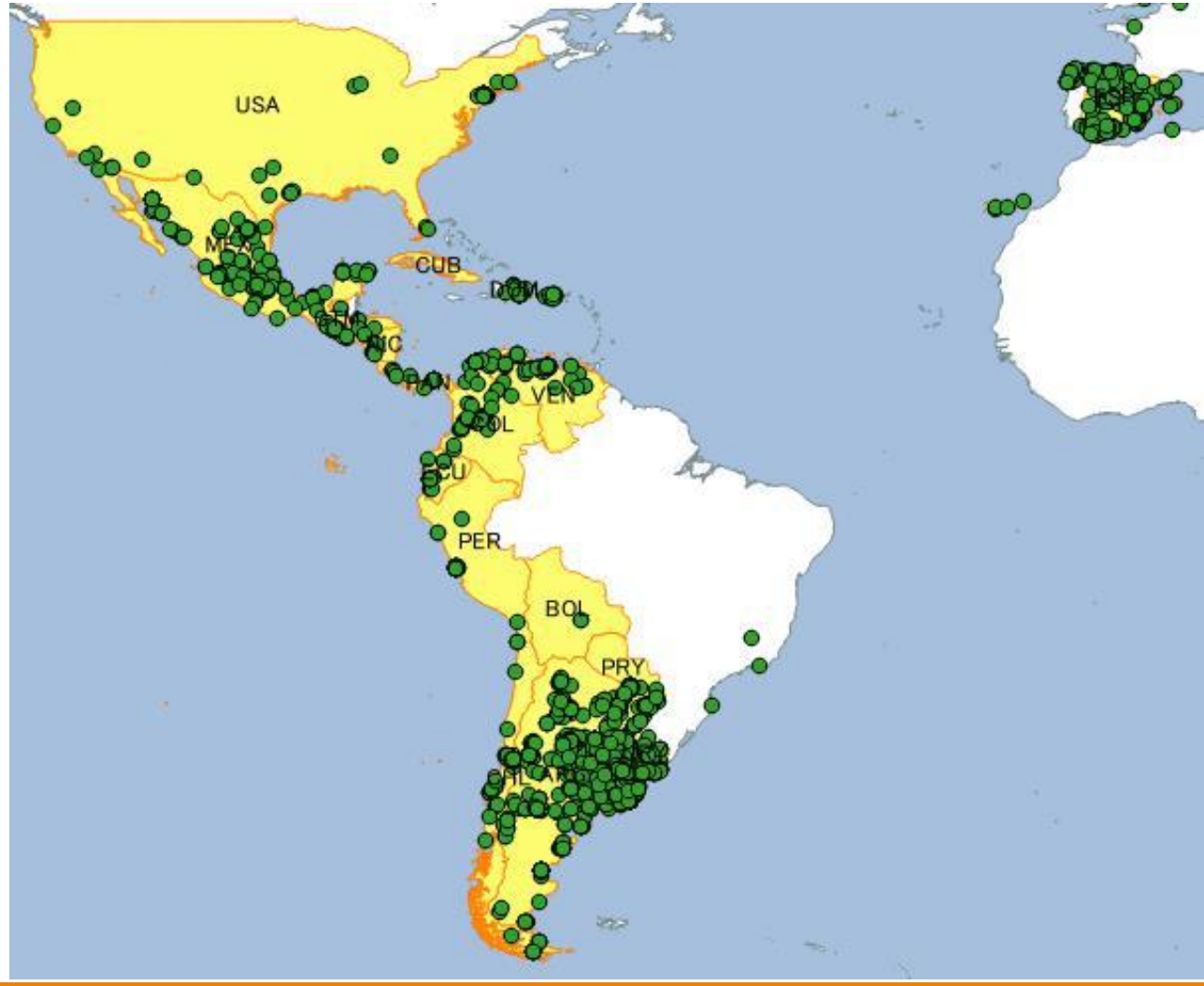
llamastes --- llamaste



hablastes --- hablaste

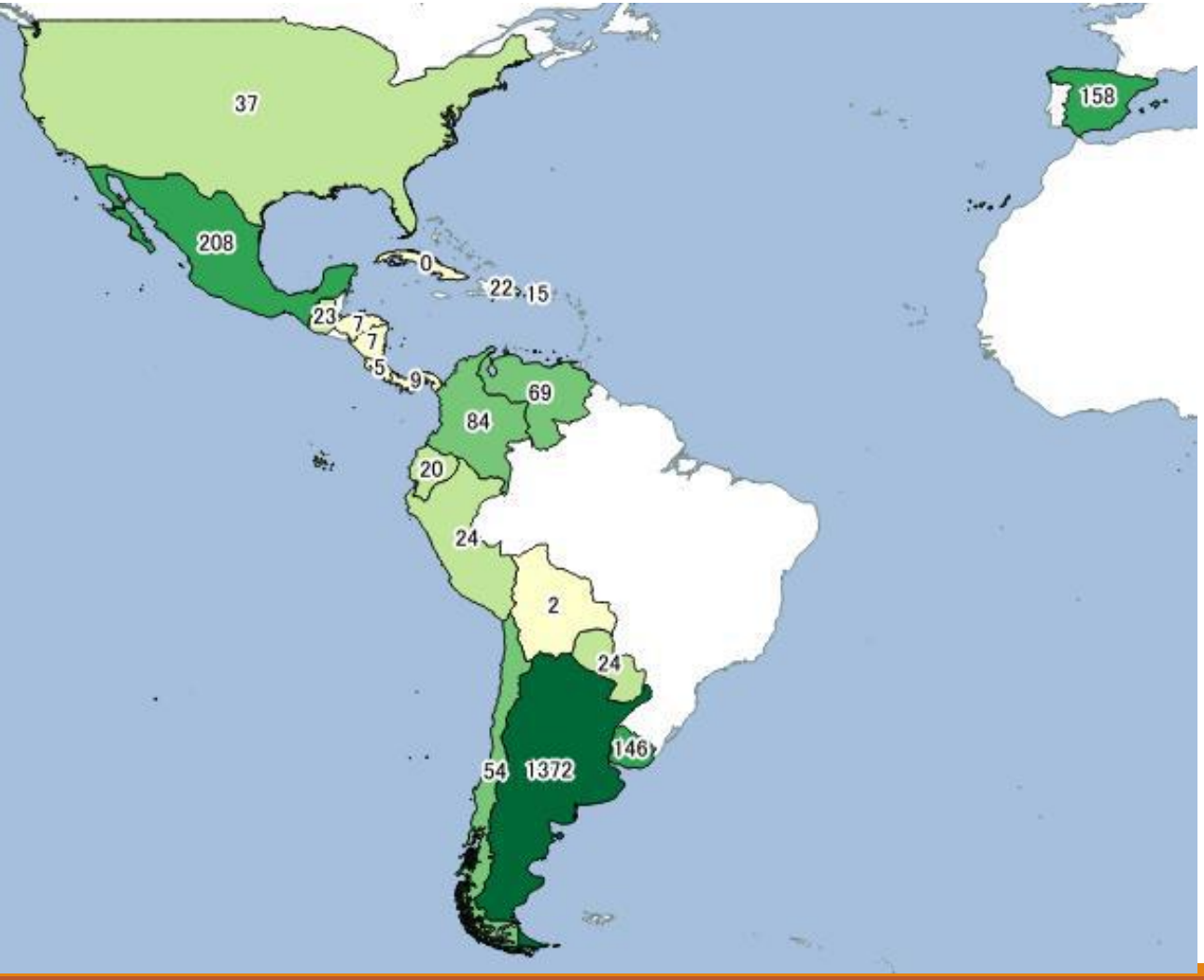


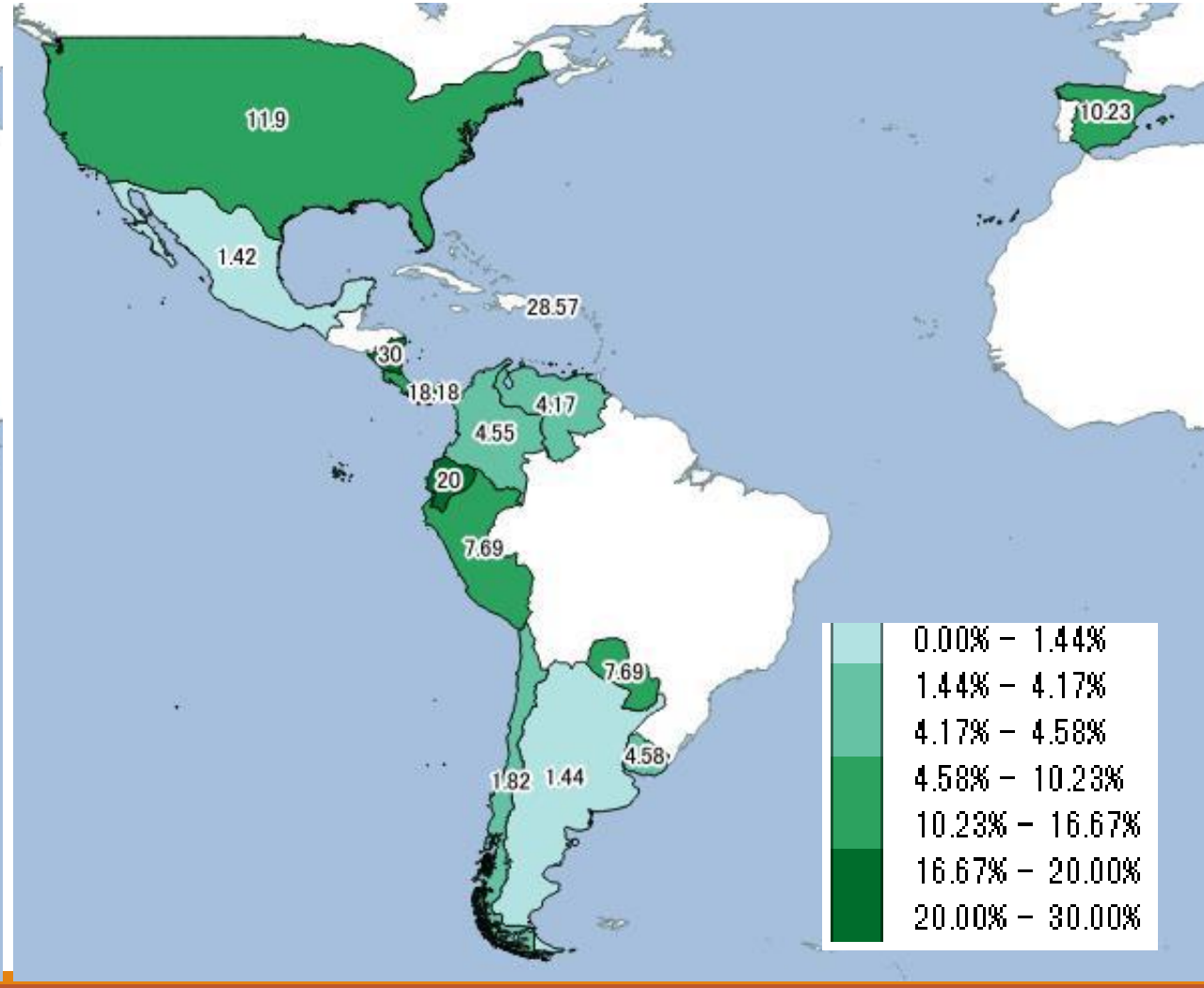
pusistes --- pusiste



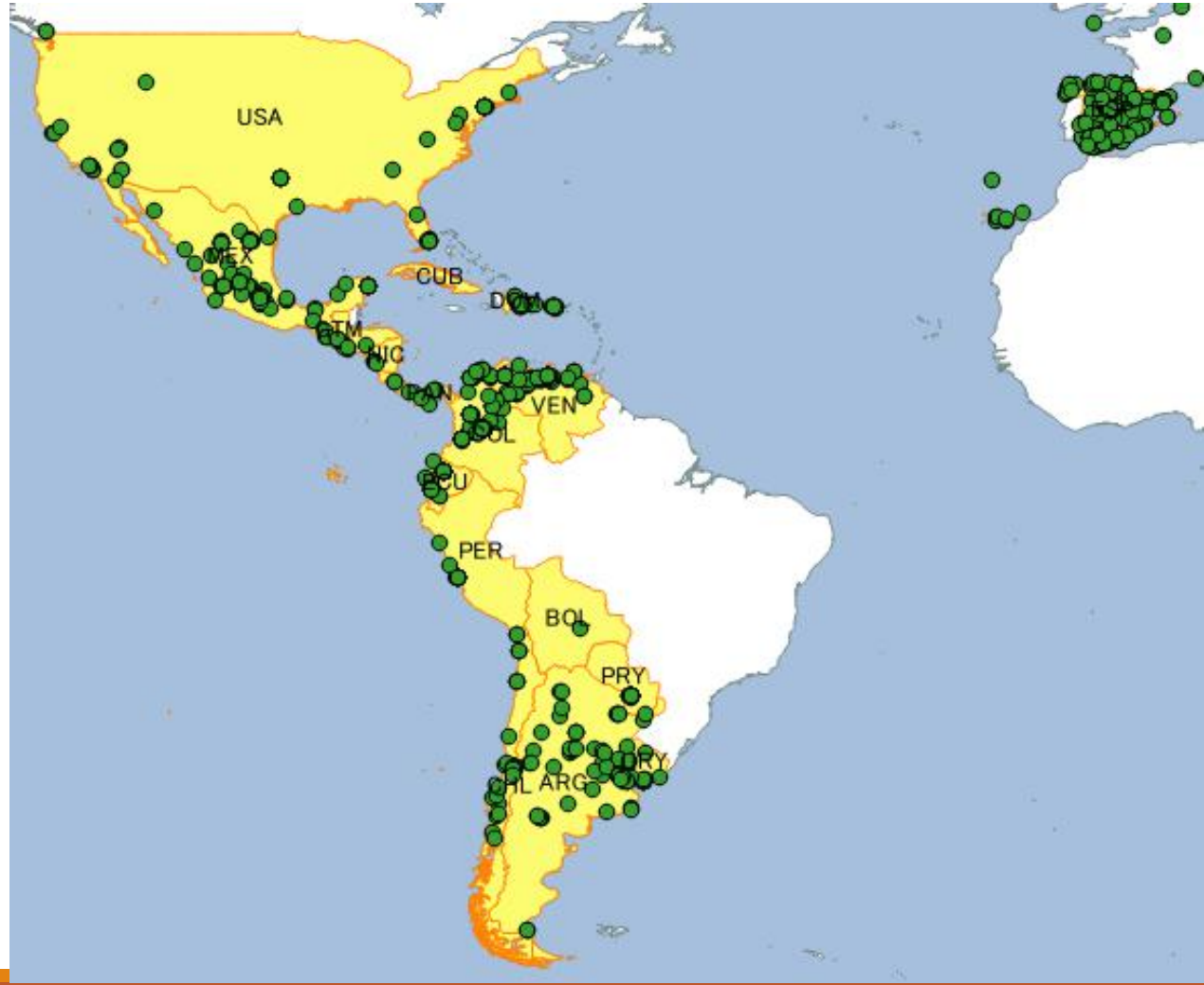
pusistes

pusistes (frecuencia absoluta vs. relativa)

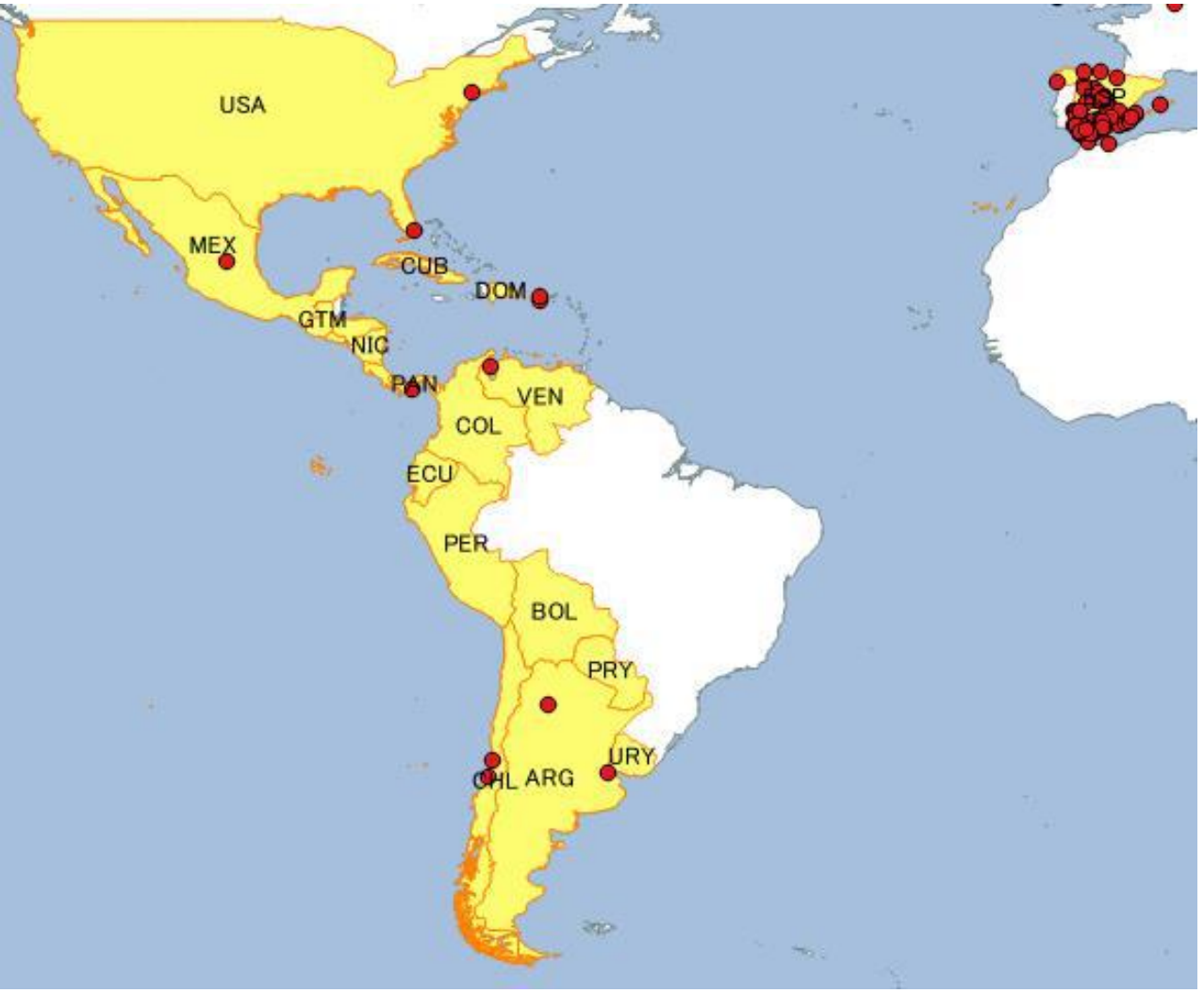


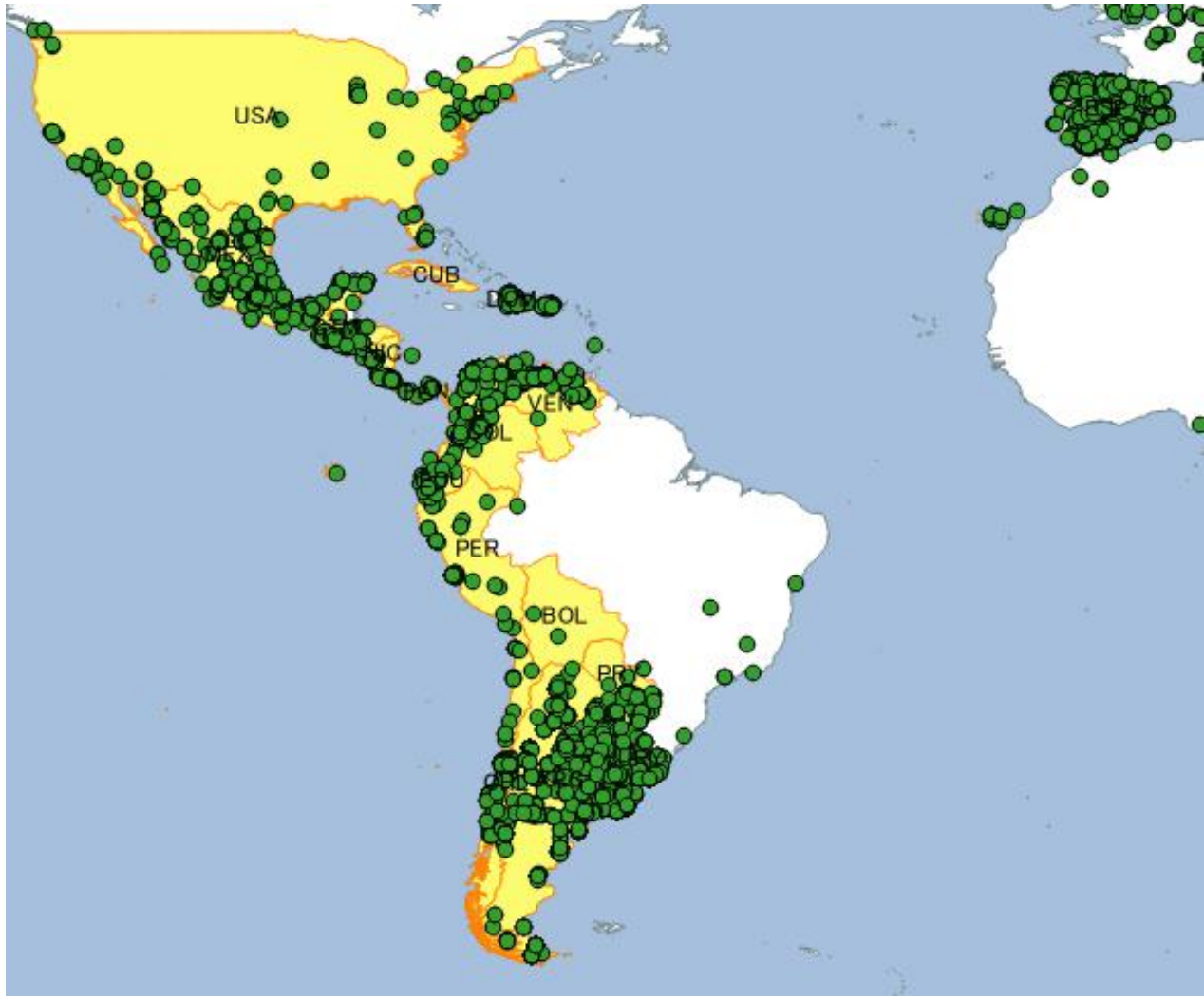


presentao --- presentado

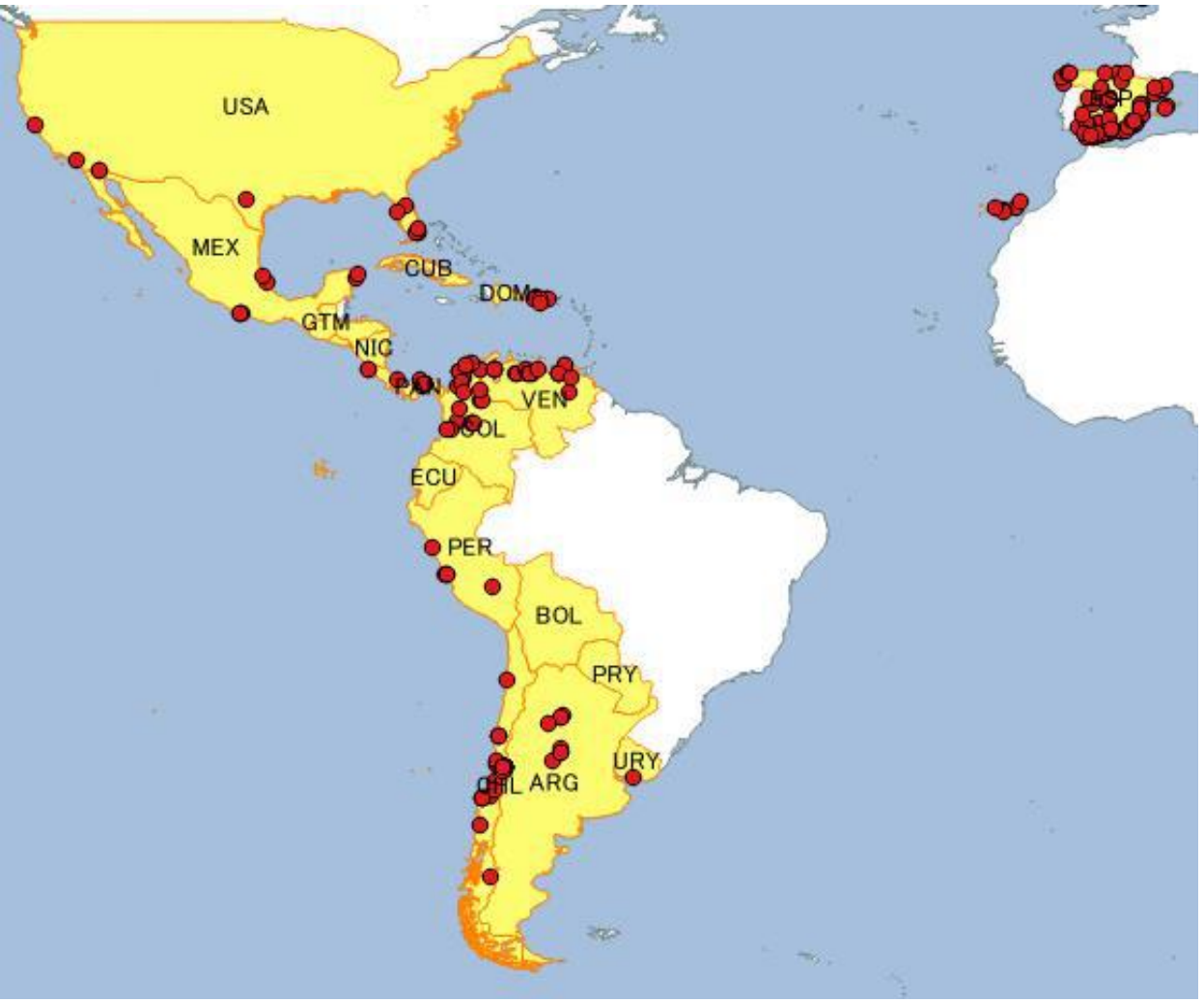


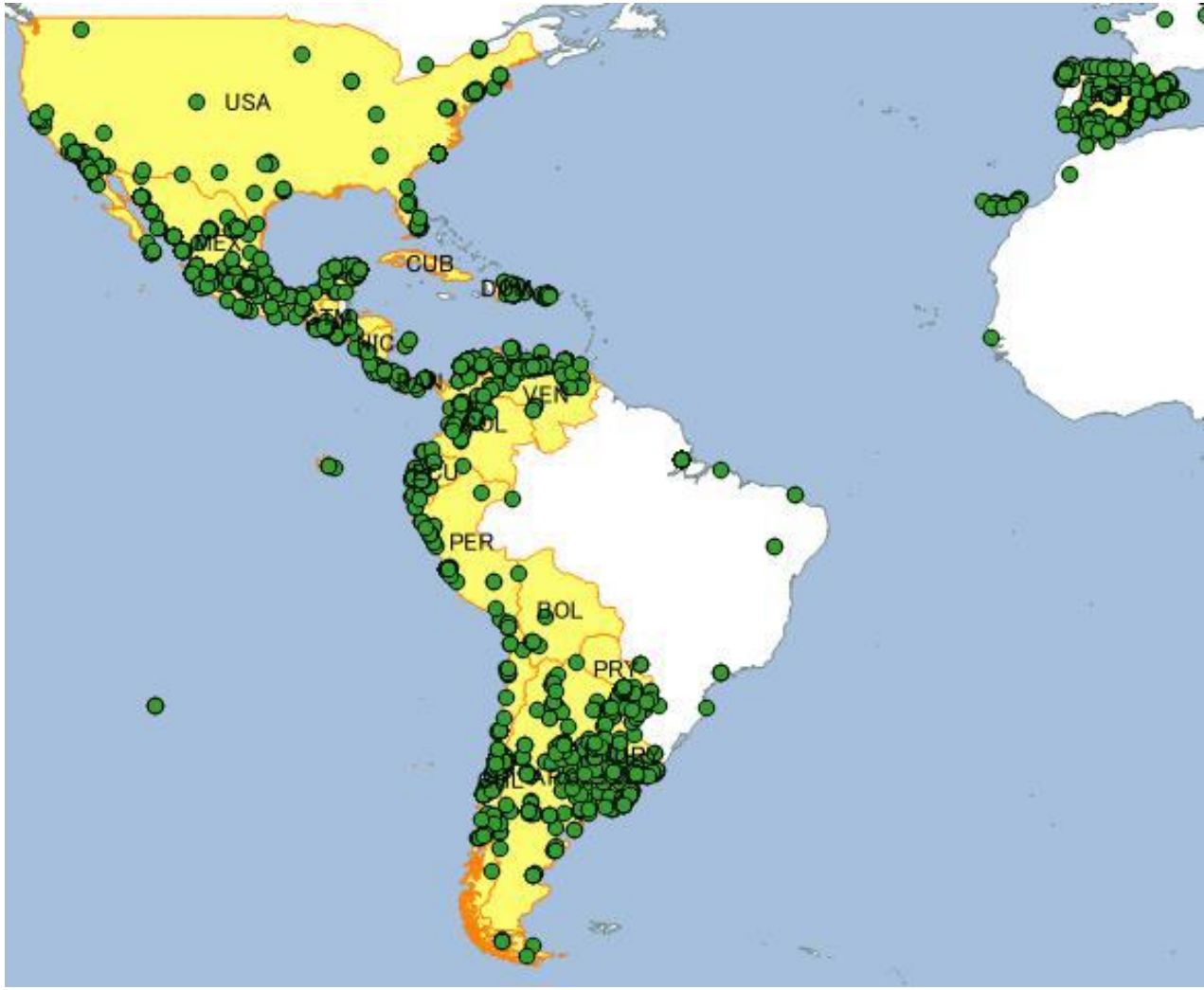
olvidao --- olvidado



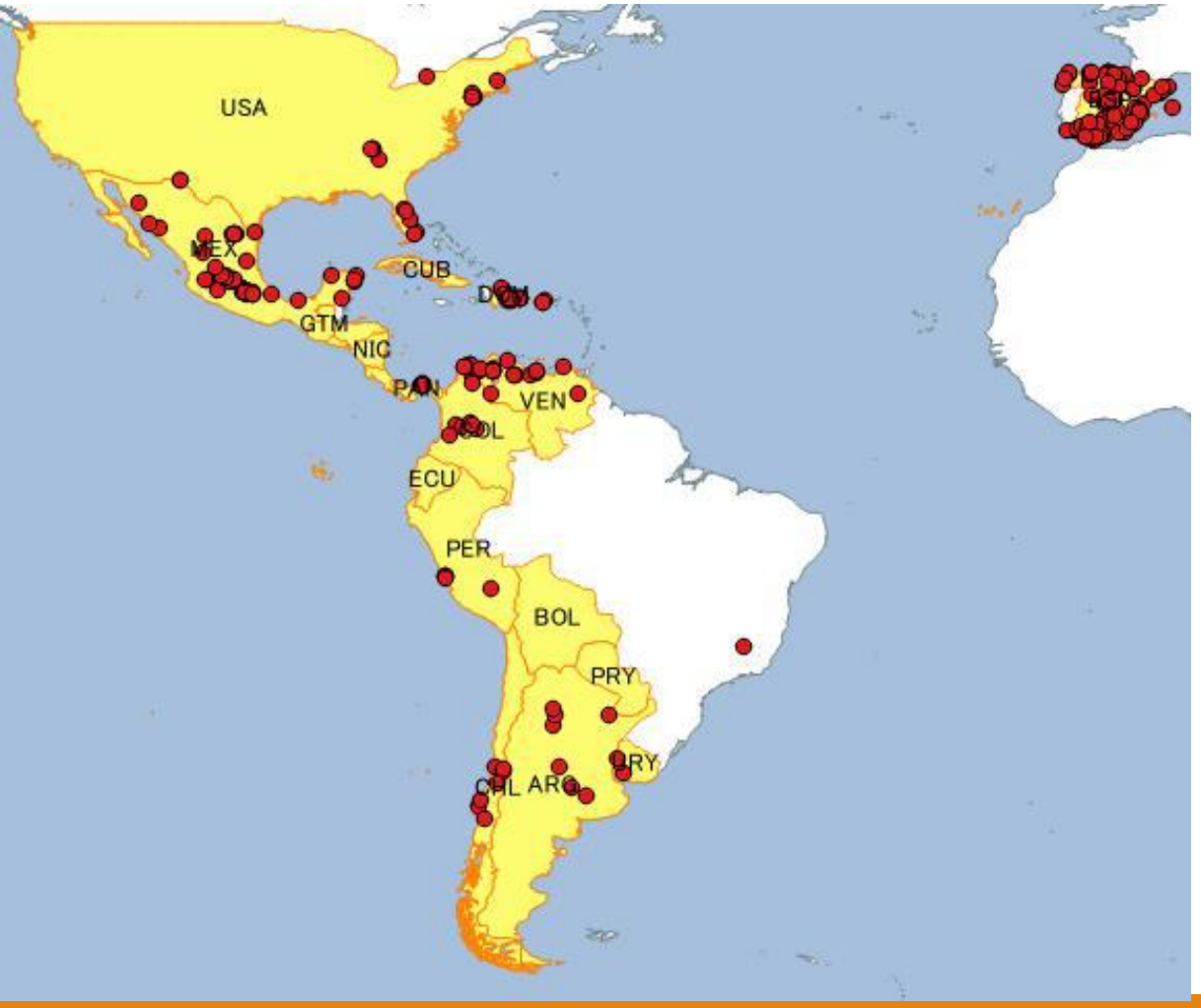


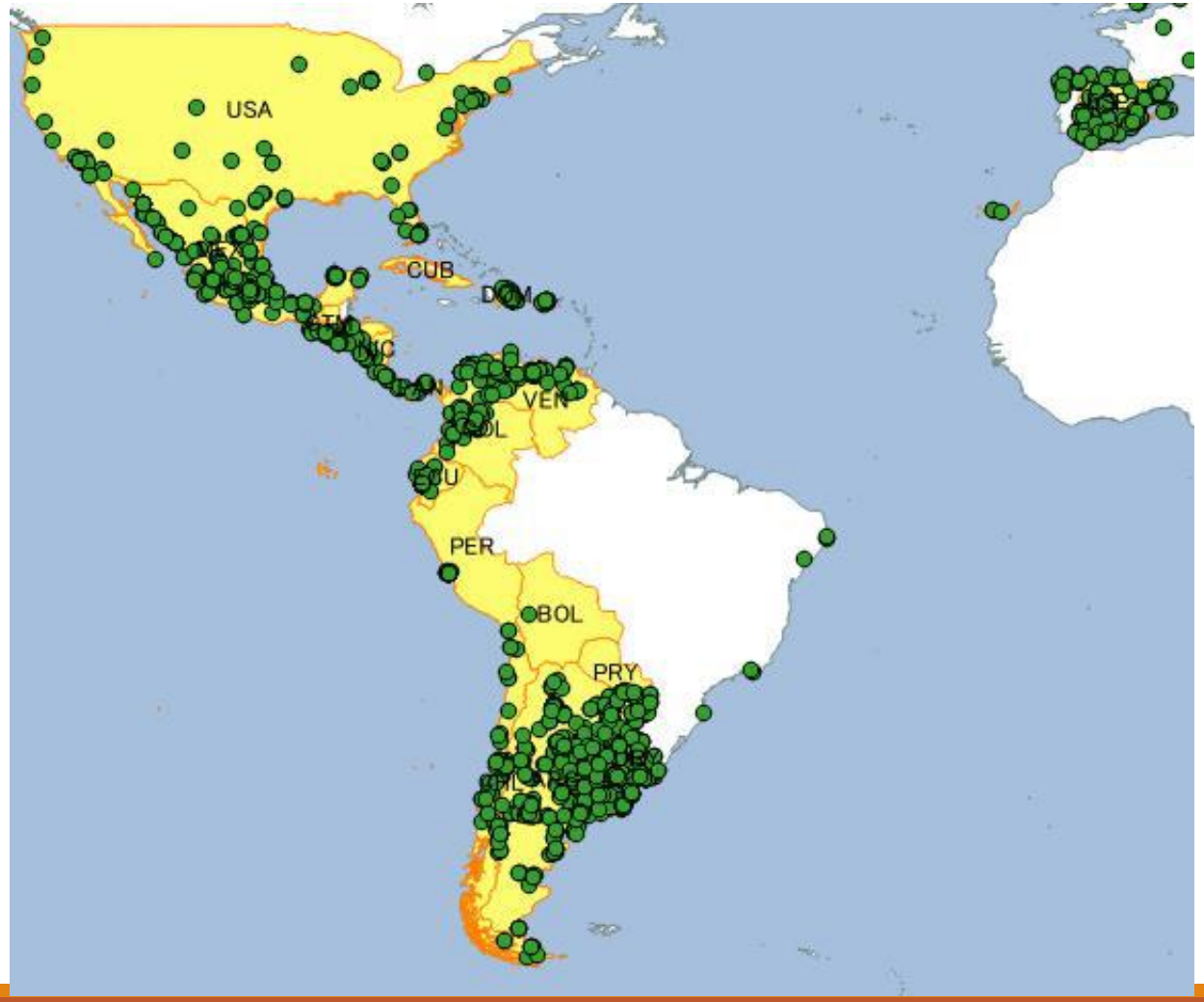
pescao --- pescado



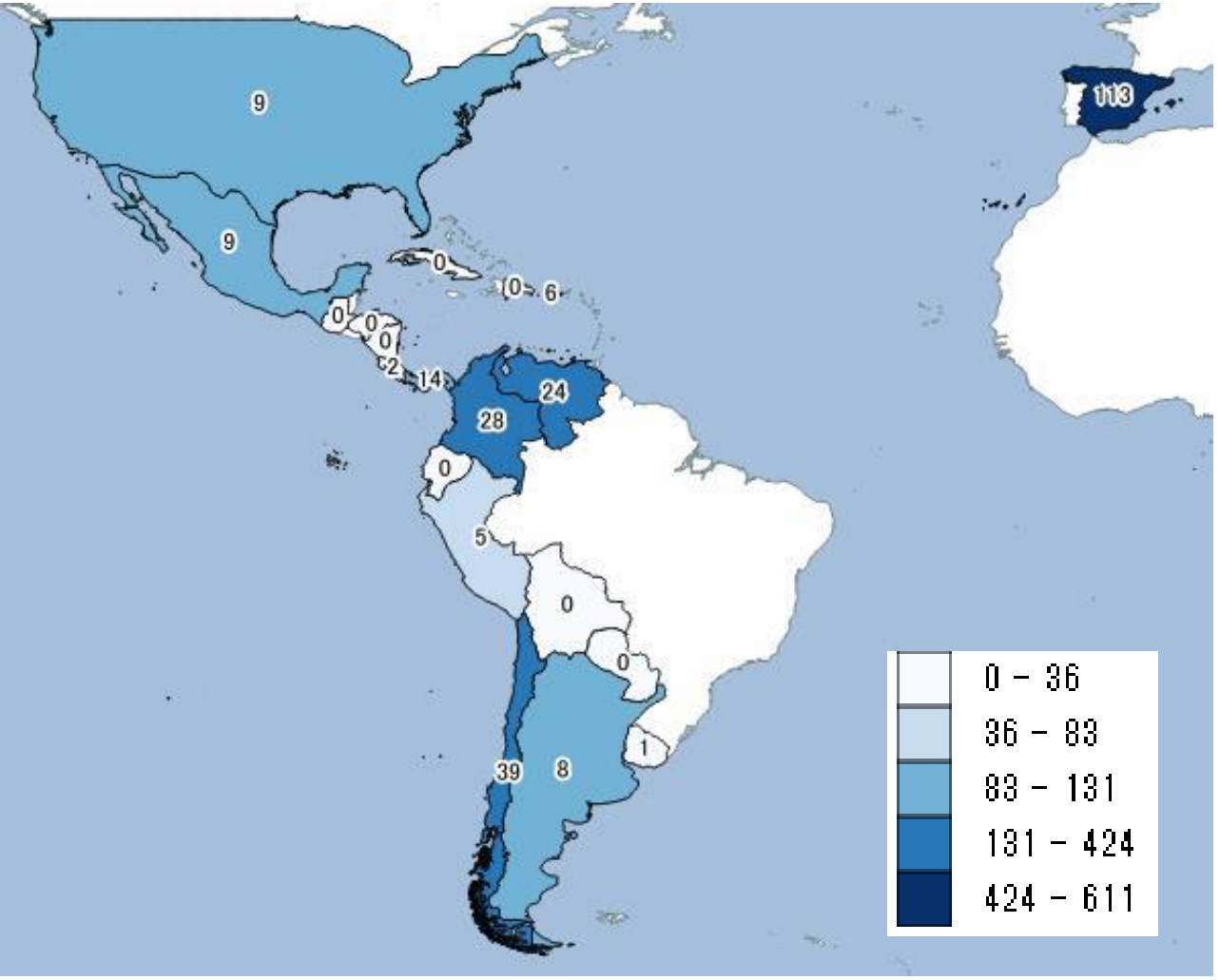


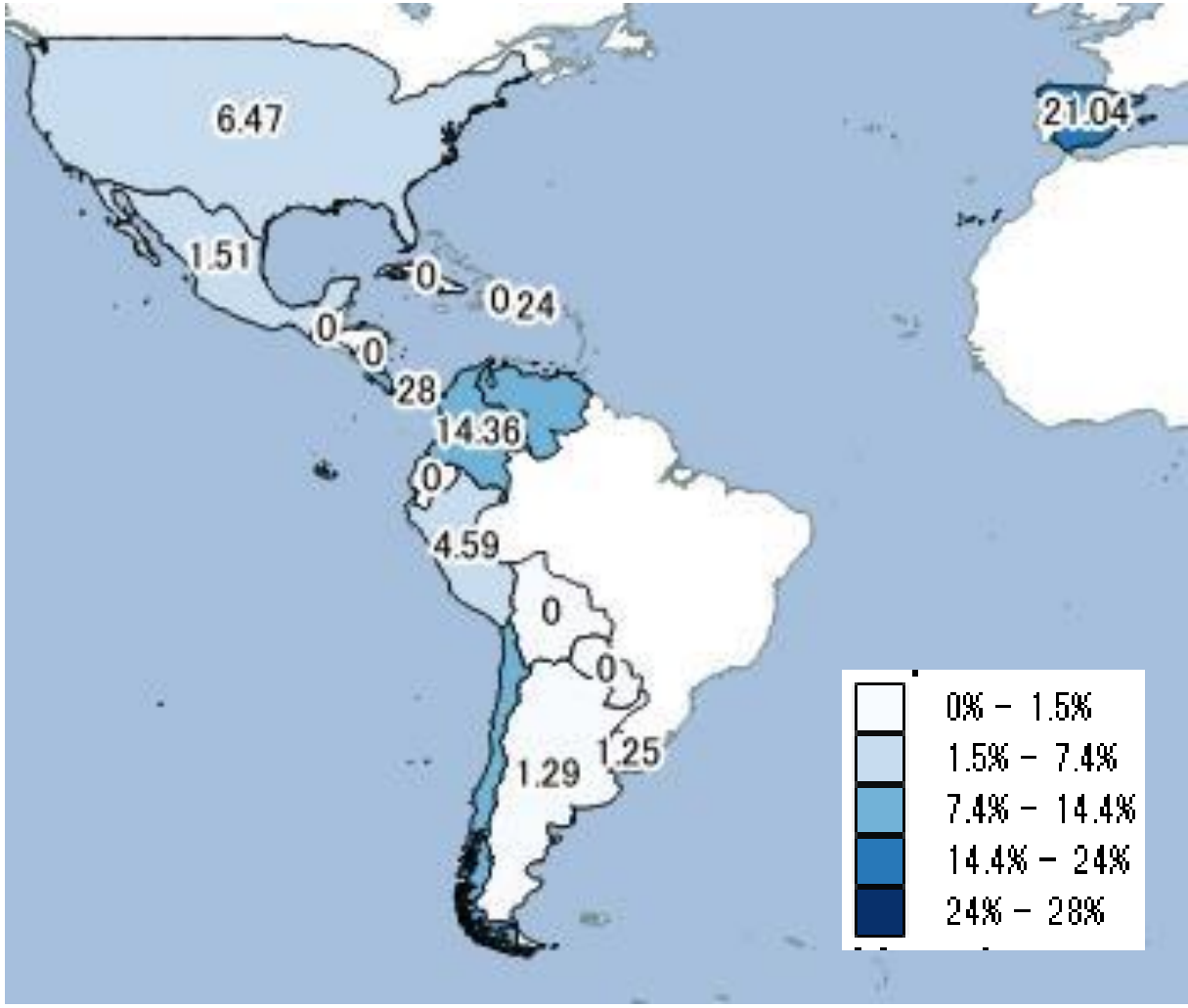
cuñao --- cuñado



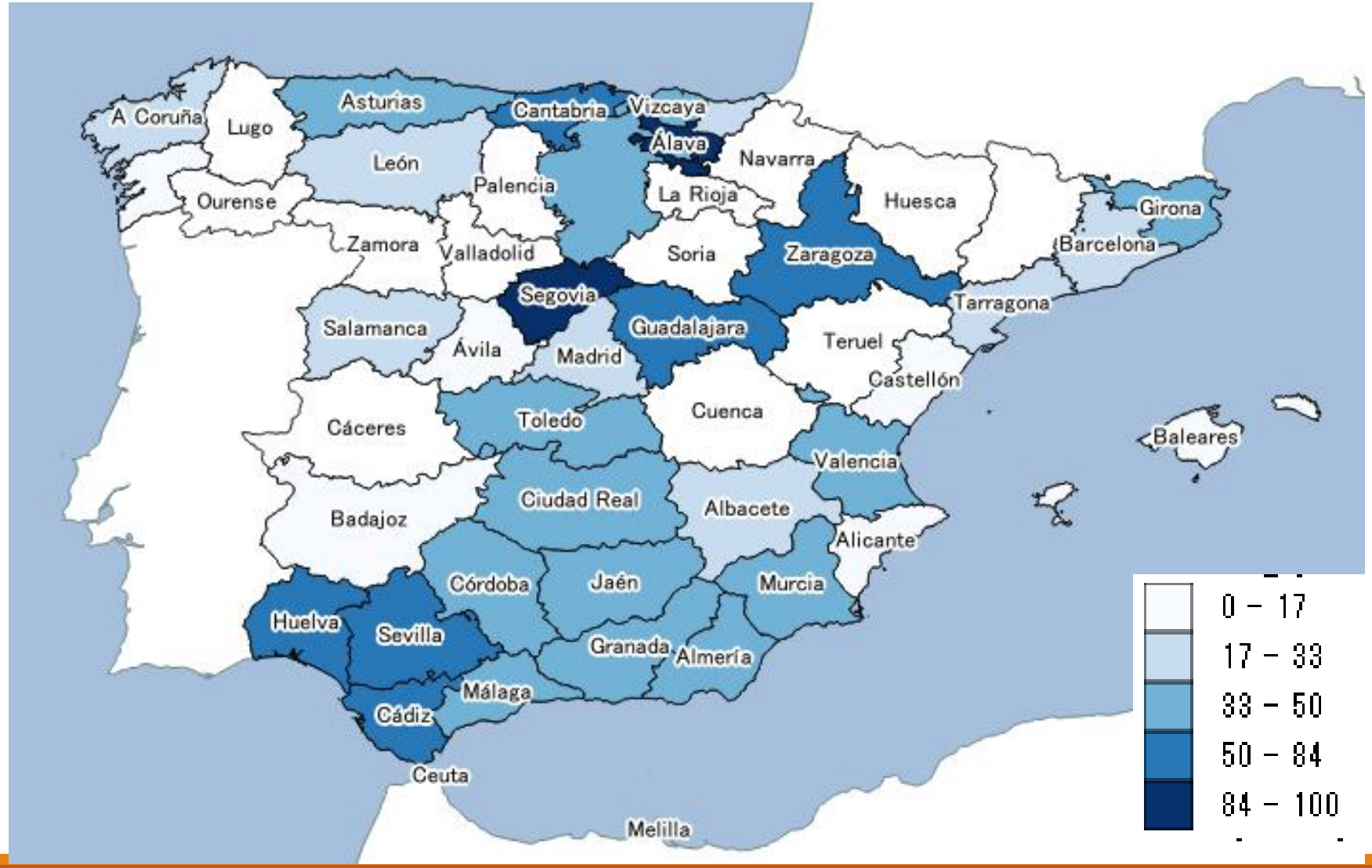


pescao (frecuencia absoluta vs. relativa)

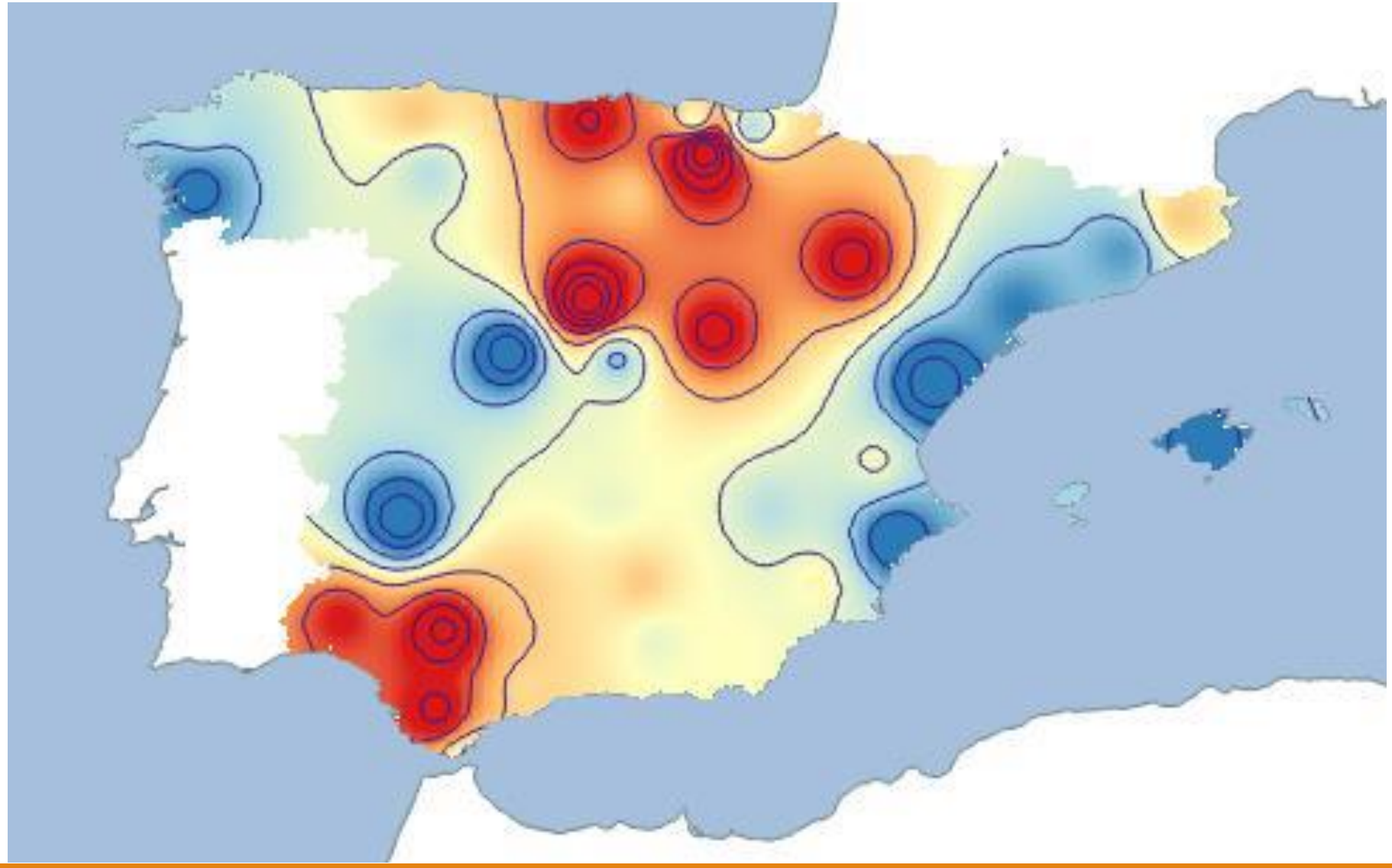




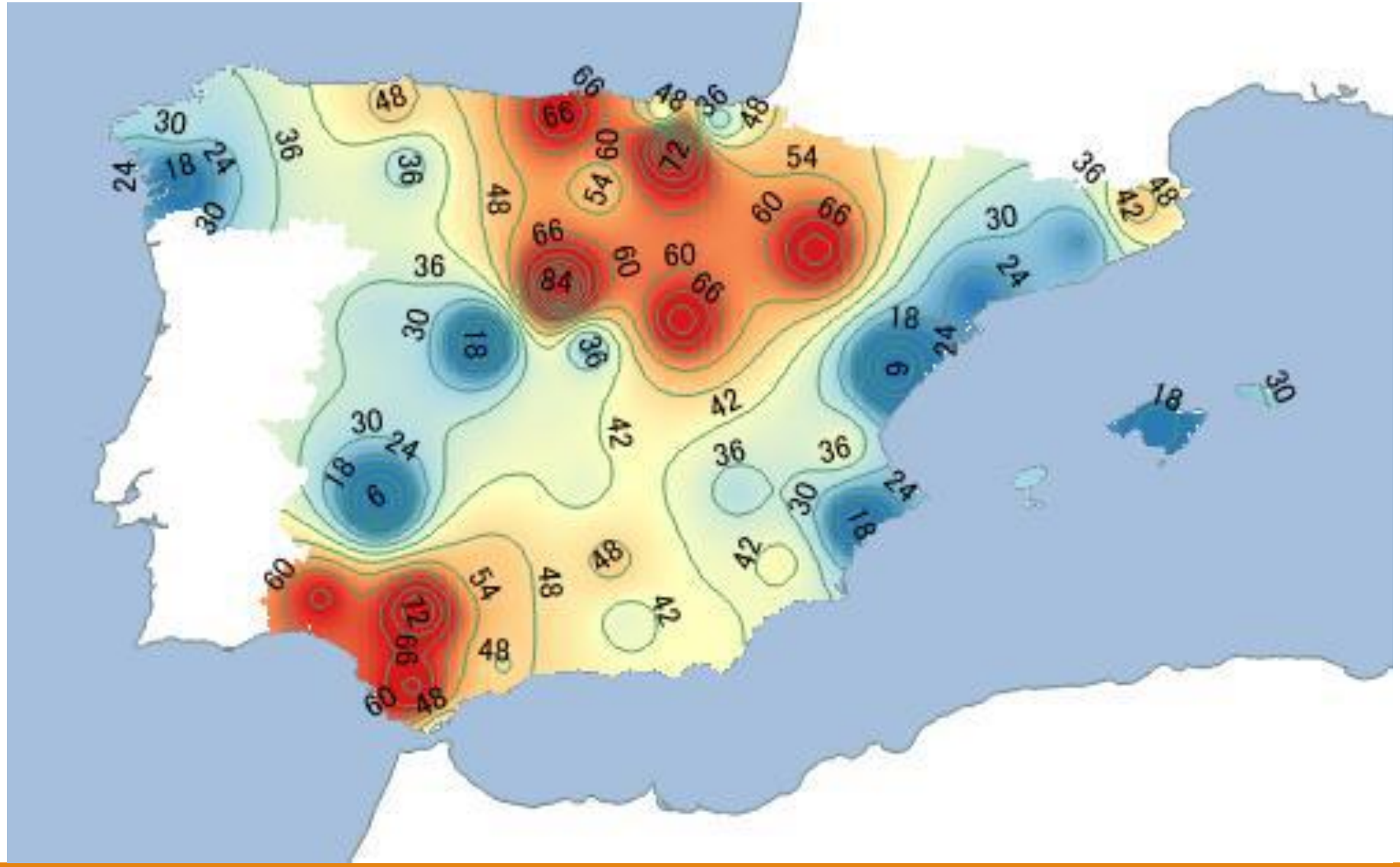
cuñao (frecuencia relativa por provincias)



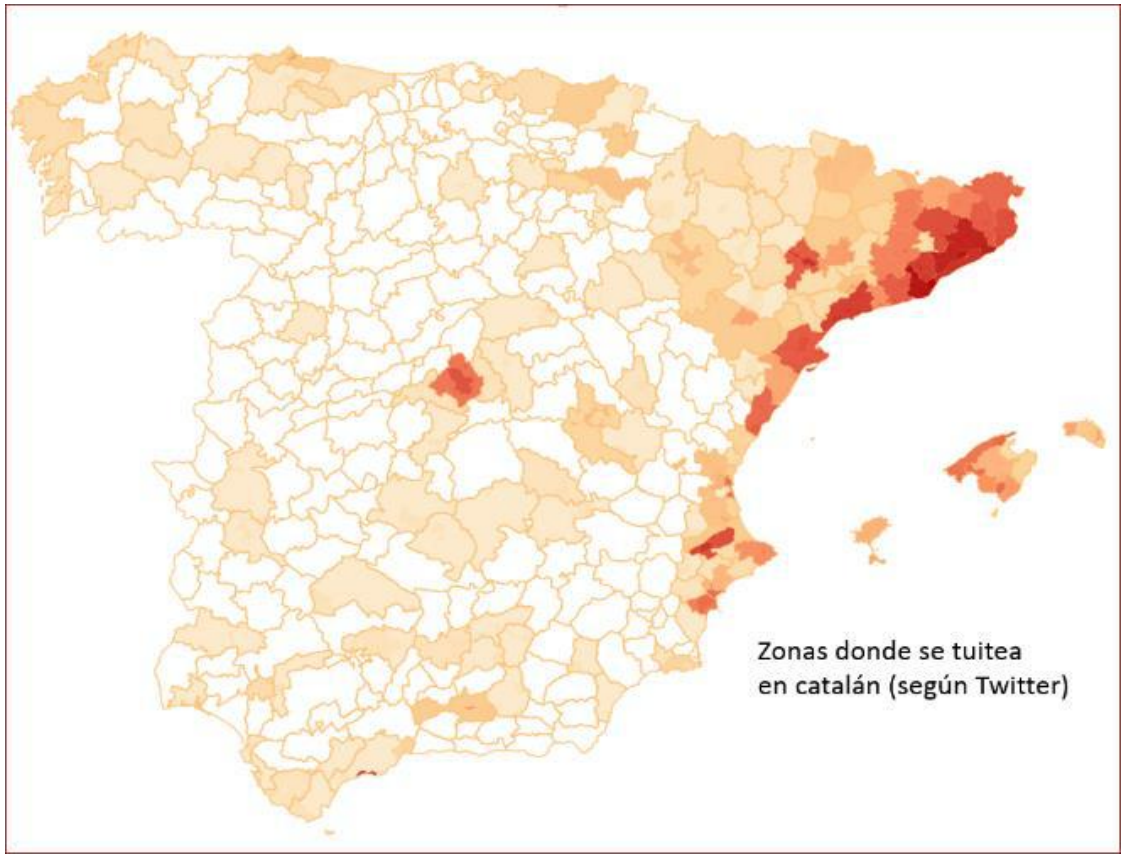
cuñao (valores interpolados ratio/provincias)



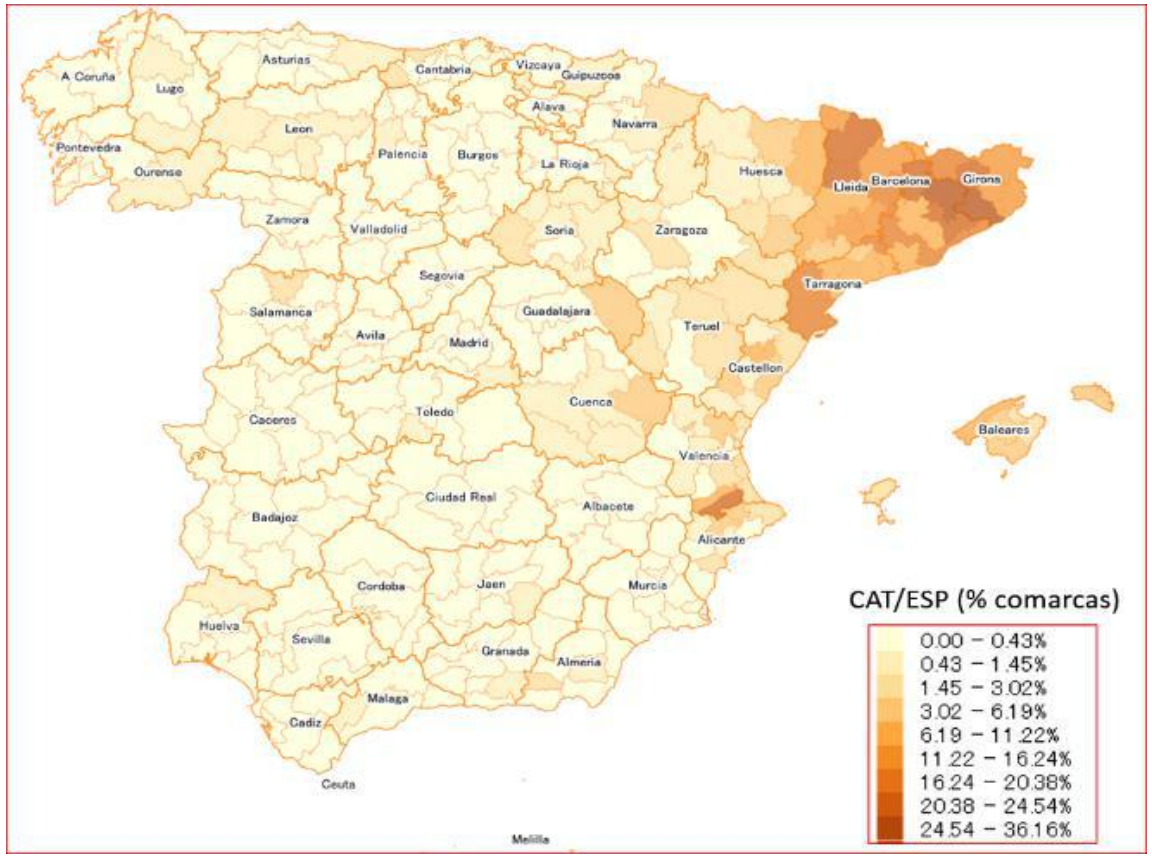
cuñao (valores interpolados ratio/provincias)



Mapas temáticos (2)

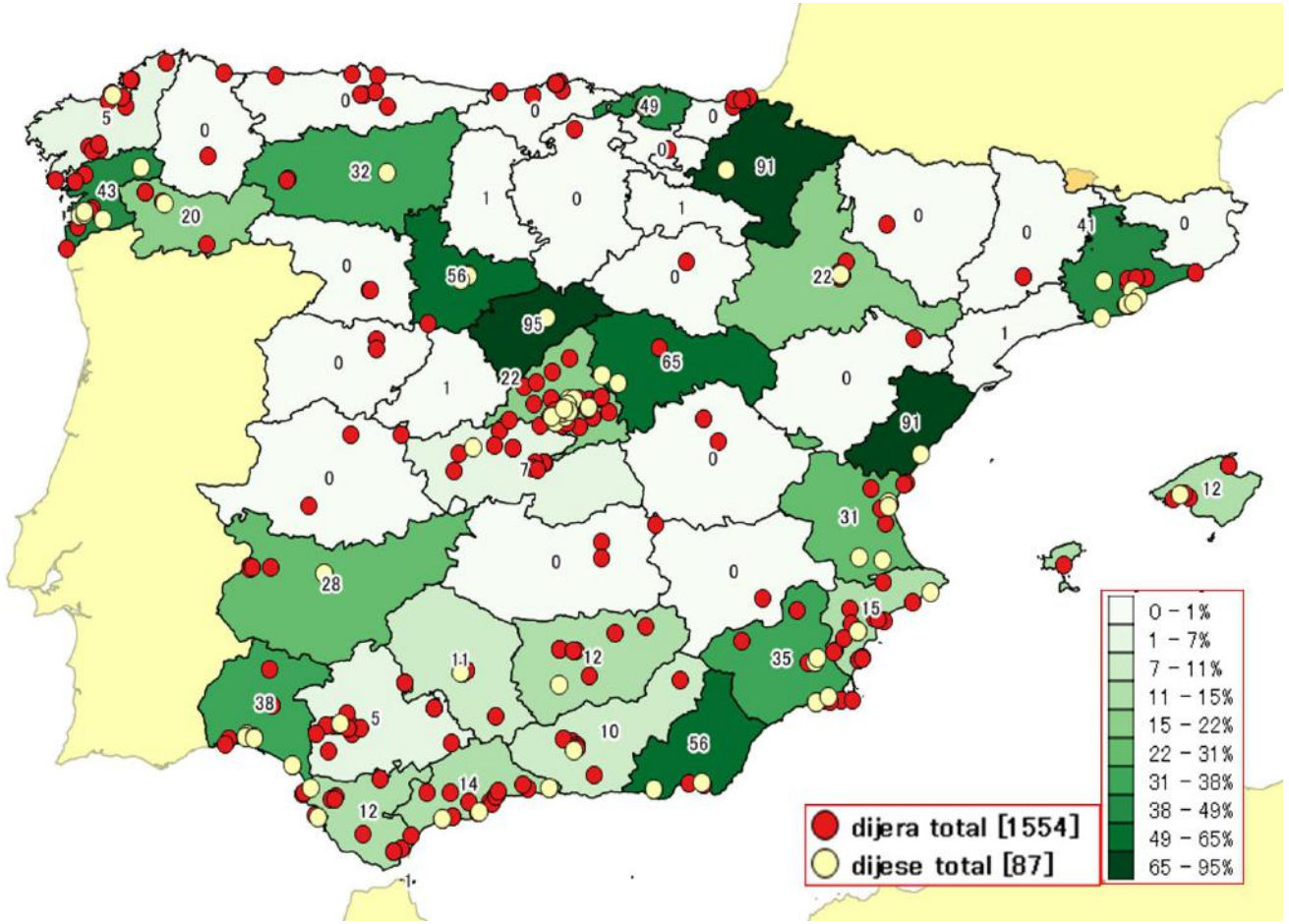


Tuits en catalán



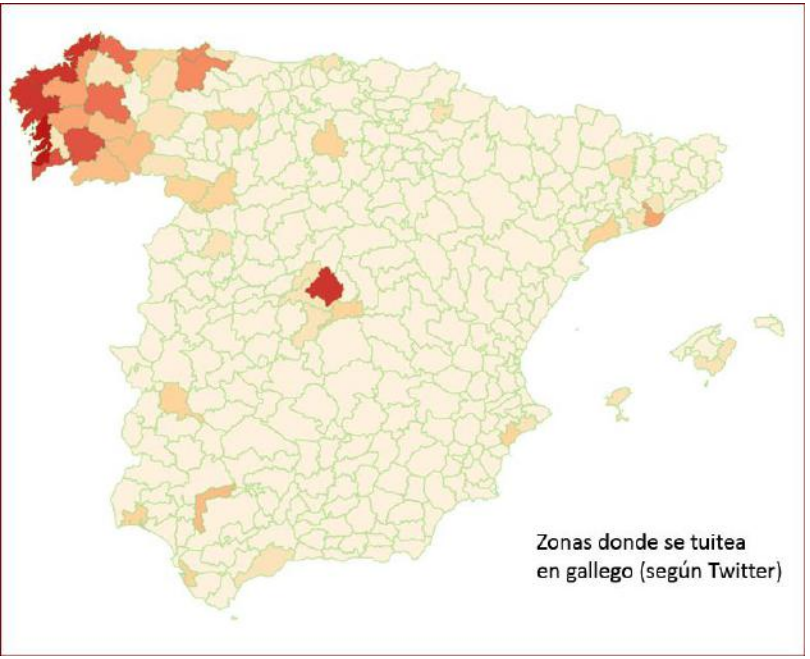
Proporción de tuits CAT/ESP

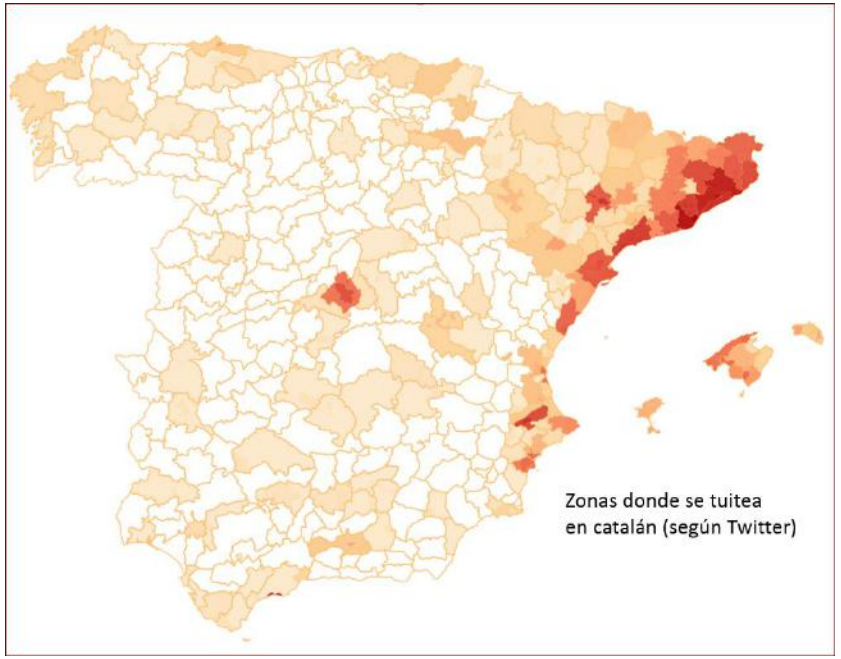
Mapa temático combinado con puntos

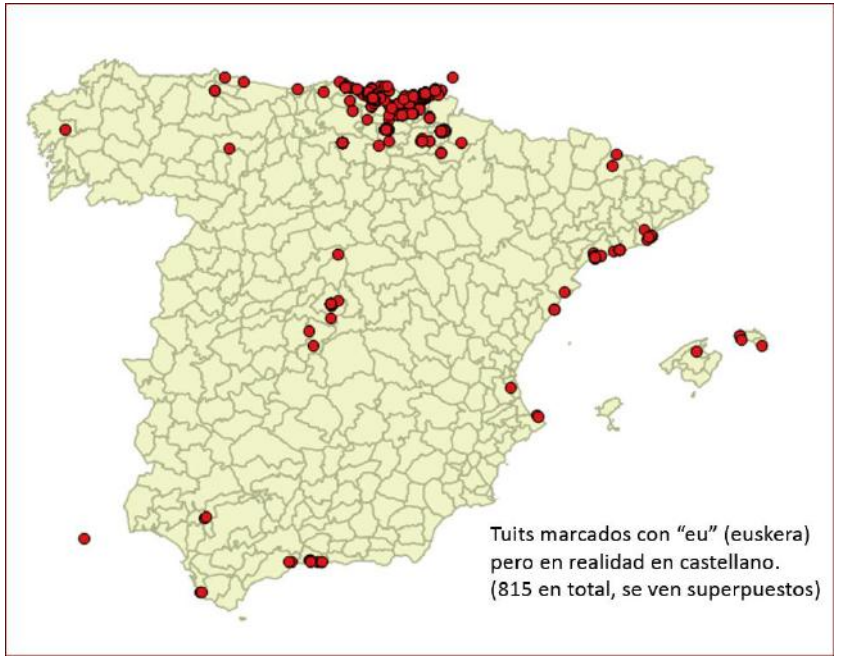


Dijera vs dijese

Contacto de lenguas
gallego catalán euskera

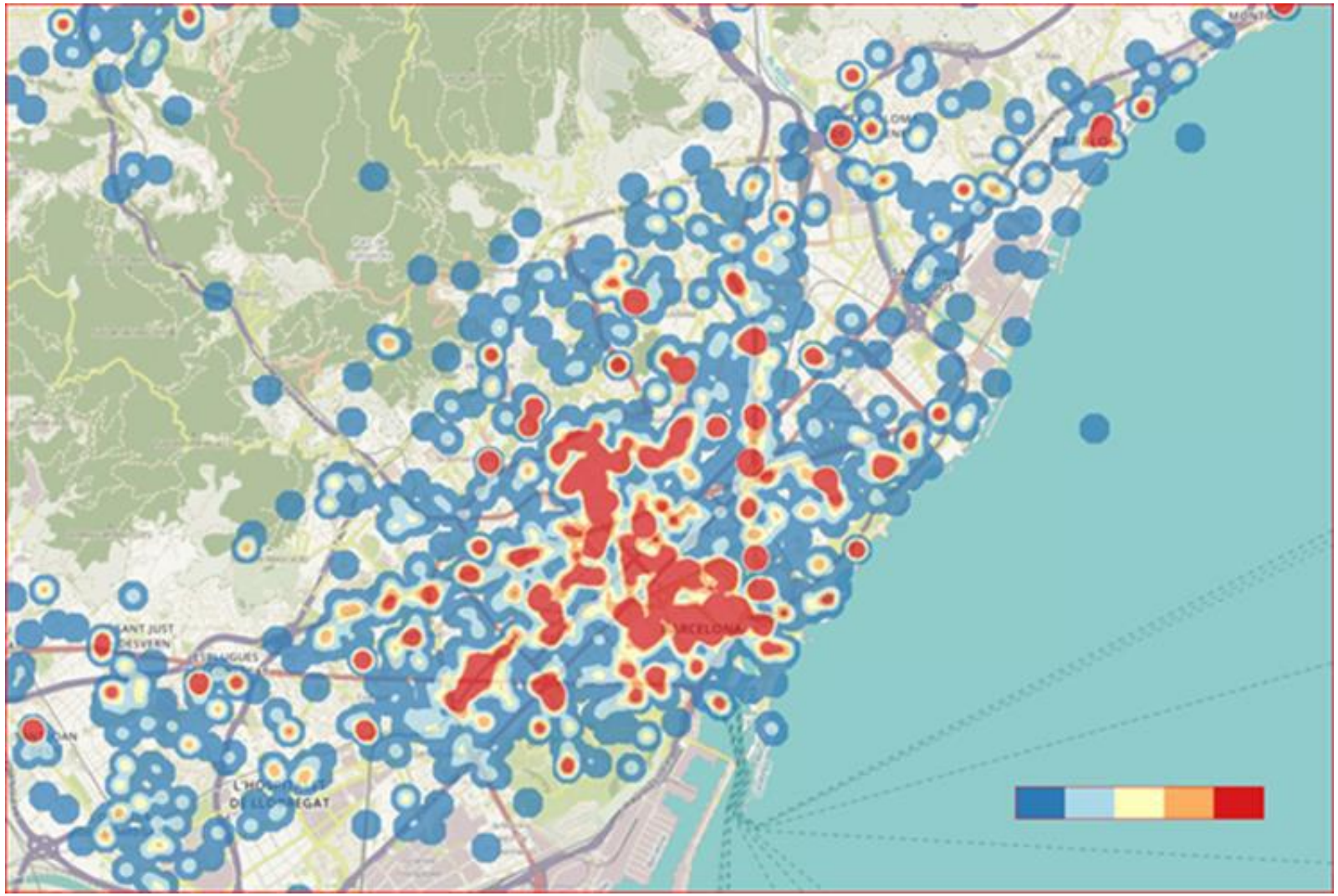






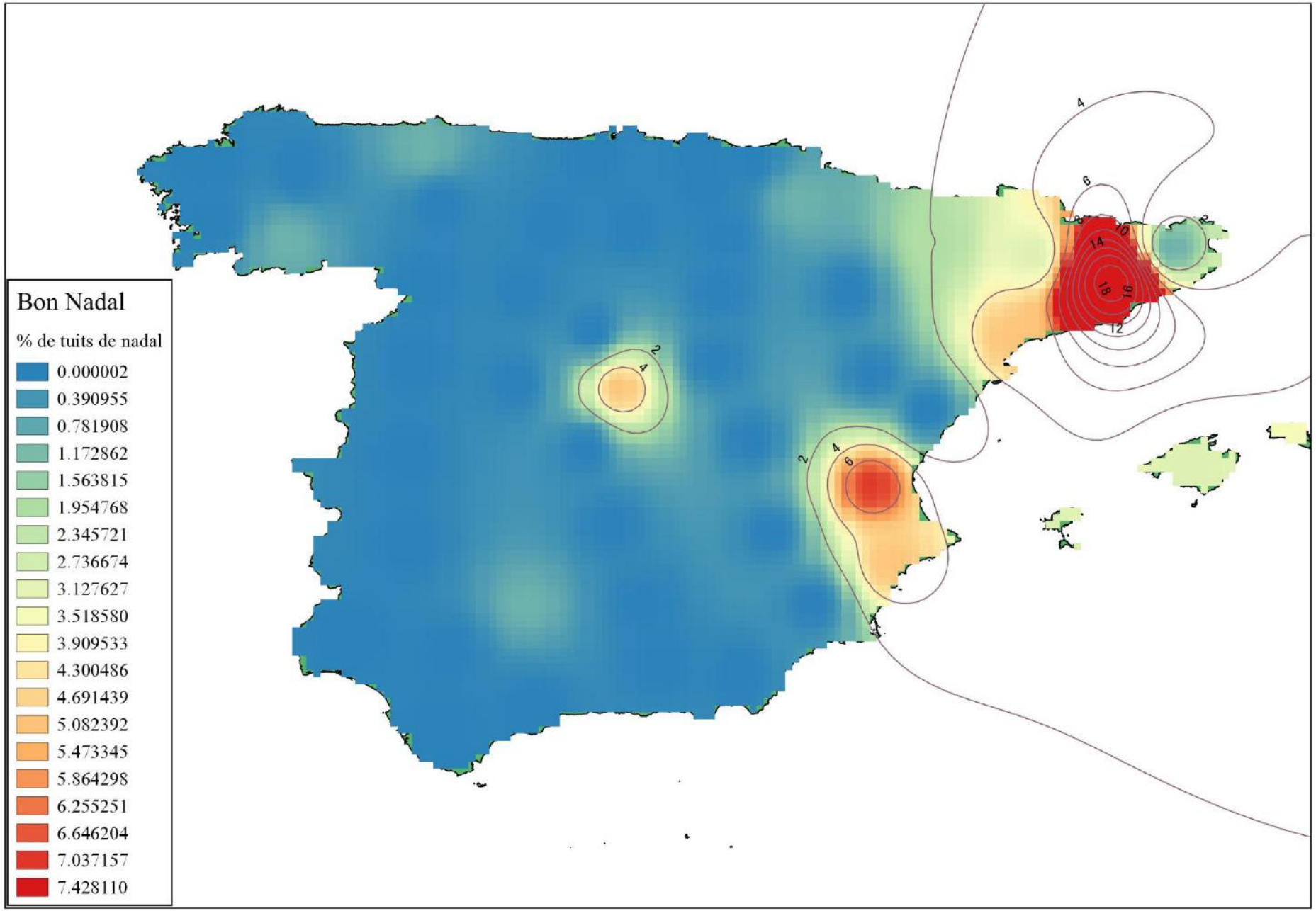
Mapa de calor (2)

Densidad de tuits
en catalán
en Barcelona



Interpolación de datos

Bon Nadal

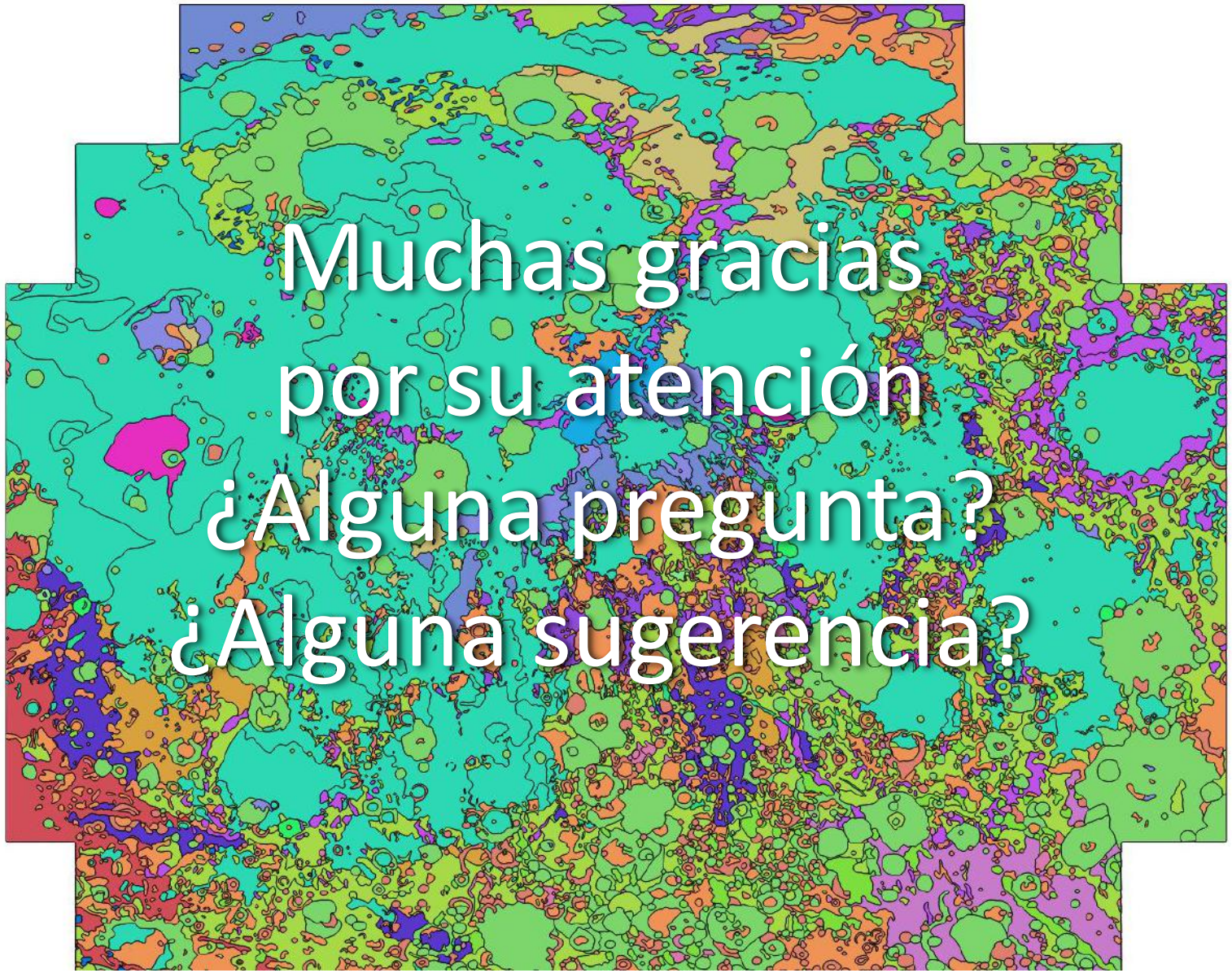


Algunas conclusiones

- Las redes sociales ofrecen una fuente de datos geocodificados y sincrónicos, de gran utilidad para la geolingüística. El lingüista variacionista puede (y debe) beneficiarse de las tecnologías de procesamiento de lenguajes naturales (NLP) y de los modernos sistemas de información geográfica (SIG).

- Se ha observado frecuentemente la coexistencia de todo tipo de variantes.

- La variación lingüística de algunos fenómenos, requiere una enorme cantidad de datos, posiblemente del orden de los miles de millones de palabras. Actualmente es posible.



Muchas gracias
por su atención
¿Alguna pregunta?
¿Alguna sugerencia?