SURNAME PATTERNS IN GALICIA

María J. Ginzo-Villamayor, Rosa M. Crujeiras and Xulio Sousa

Universidad de Santiago de Compostela

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
INSTITUTO DA LINGUA GALEGA (ILG)

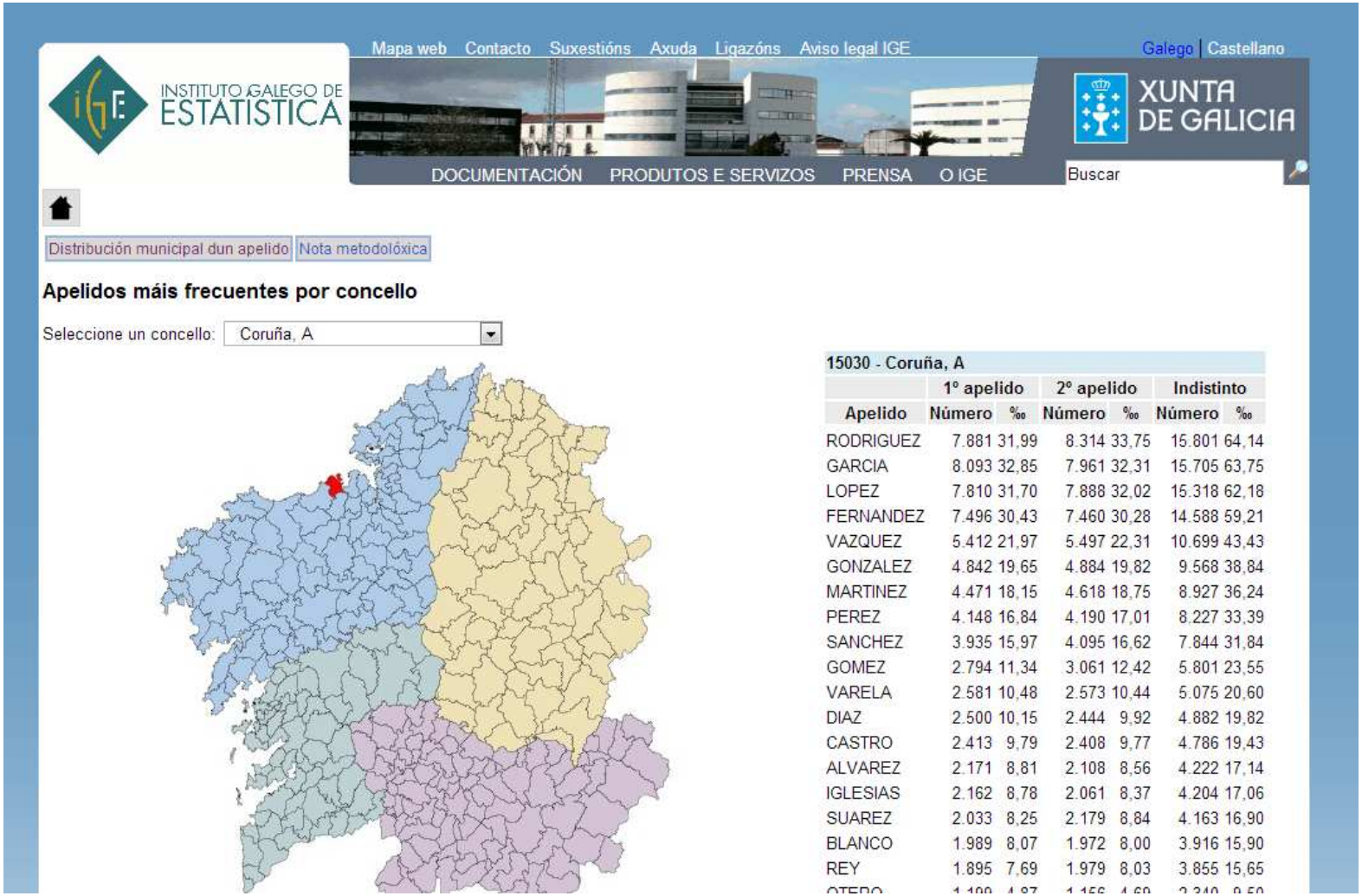
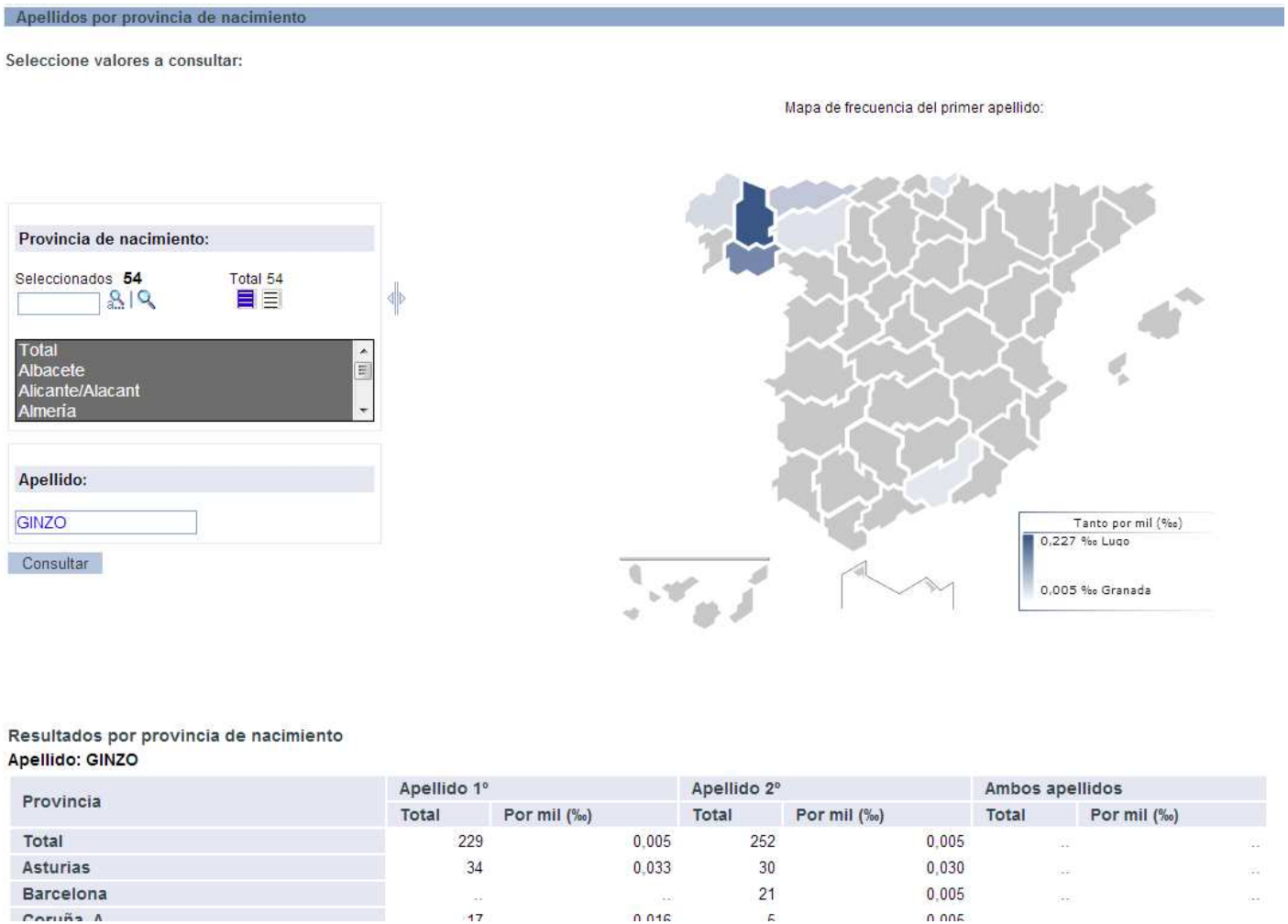


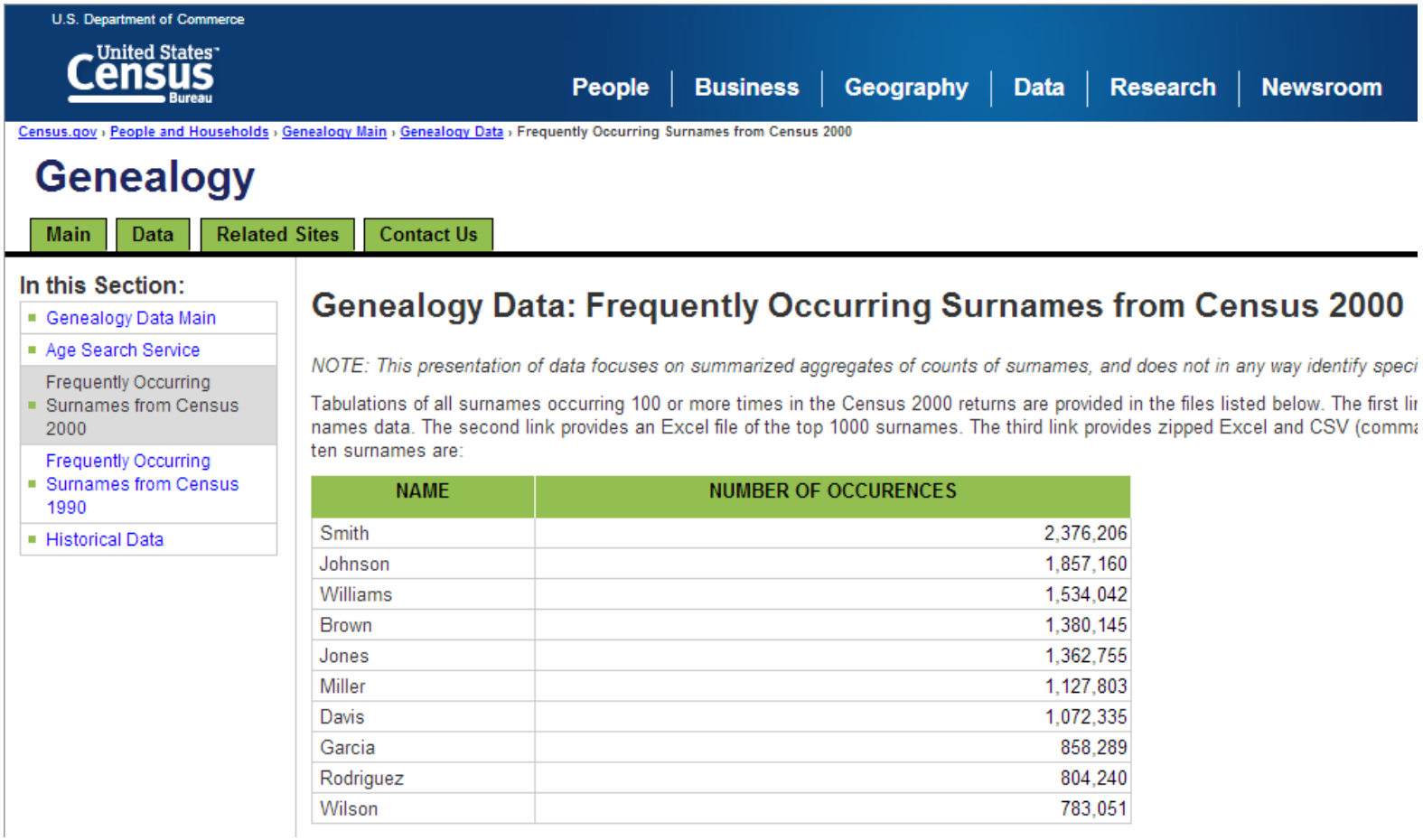
Figure: Surname frequency by councils (from http://www.ige.eu)



Resultados por provincia de nacimiento
Apellido: GINZO

Provincia	Apellido 1º		Apellido 2º		Ambos apellidos	
	Total	Por mil (‰)	Total	Por mil (‰)	Total	Por mil (‰)
Total	229	0,005	252	0,005	..	..
Asturias	34	0,033	30	0,030	..	..
Barcelona	..	..	21	0,005	..	..
Coruña, A	17	0,016	5	0,005		

Figure: Surname frequency map (from http://www.ine.es/)



U.S. Department of Commerce

United States Census Bureau

People | Business | Geography | Data | Research | Newsroom

Census.gov › People and Households › Genealogy Main › Genealogy Data › Frequently Occurring Surnames from Census 2000

Genealogy

Main | Data | Related Sites | Contact Us

In this Section:

- Genealogy Data Main
- Age Search Service
- Frequently Occurring Surnames from Census 2000
- Frequently Occurring Surnames from Census 1990
- Historical Data

Genealogy Data: Frequently Occurring Surnames from Census 2000

NOTE: This presentation of data focuses on summarized aggregates of counts of surnames, and does not in any way identify speci

Tabulations of all surnames occurring 100 or more times in the Census 2000 returns are provided in the files listed below. The first li names data. The second link provides an Excel file of the top 1000 surnames. The third link provides zipped Excel and CSV (comma ten surnames are:

NAME	NUMBER OF OCCURENCES
Smith	2,376,206
Johnson	1,857,160
Williams	1,534,042
Brown	1,380,145
Jones	1,362,755
Miller	1,127,803
Davis	1,072,335
Garcia	858,289
Rodriguez	804,240
Wilson	783,051

Figure: Frequently occurring surnames in USA 2000 (United States, Census Bureau)

Figure: Frequently occurring surnames in USA

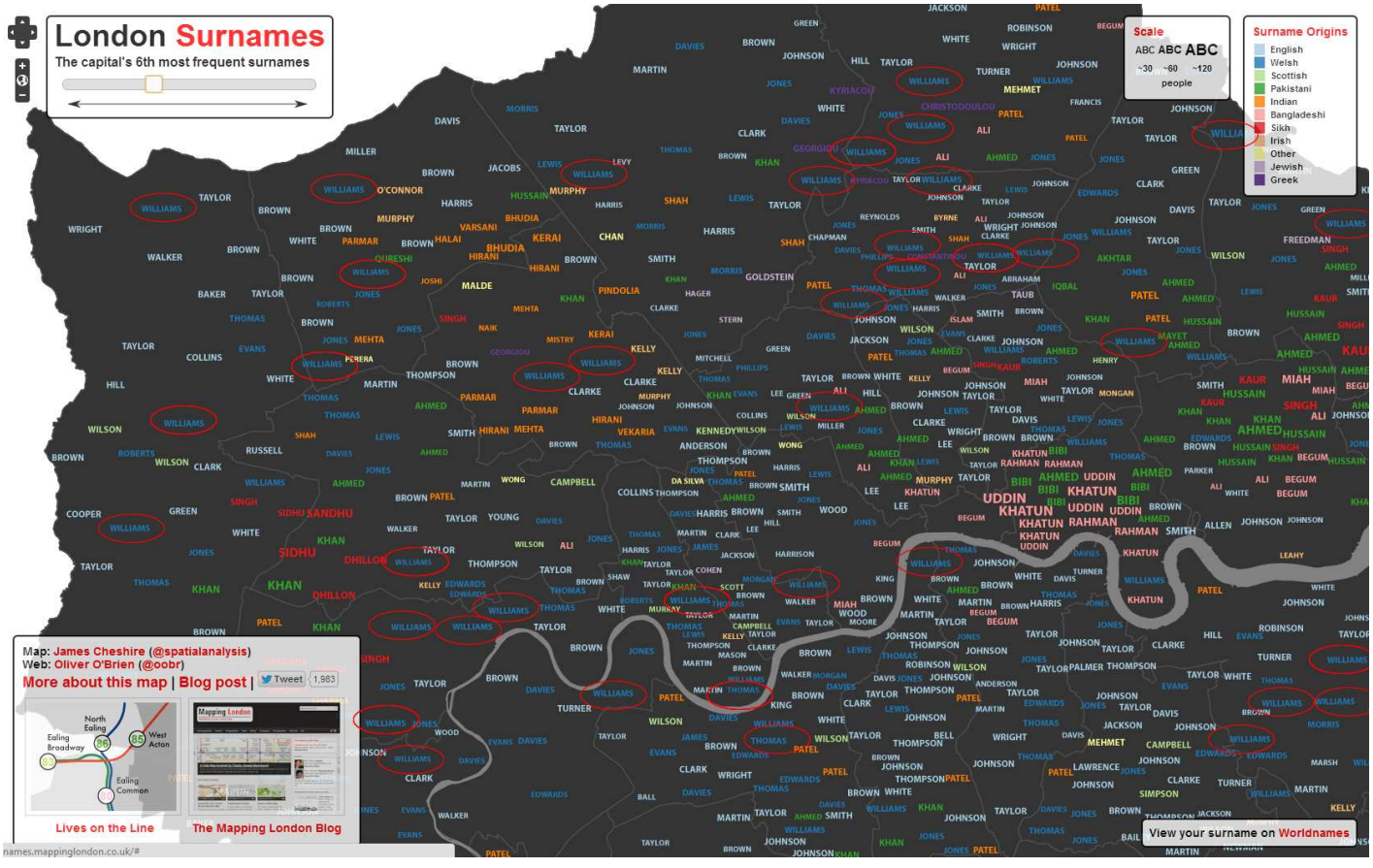


Figure: Frequently occurring surnames in London (from http://names.mappinglondon.co.uk/)

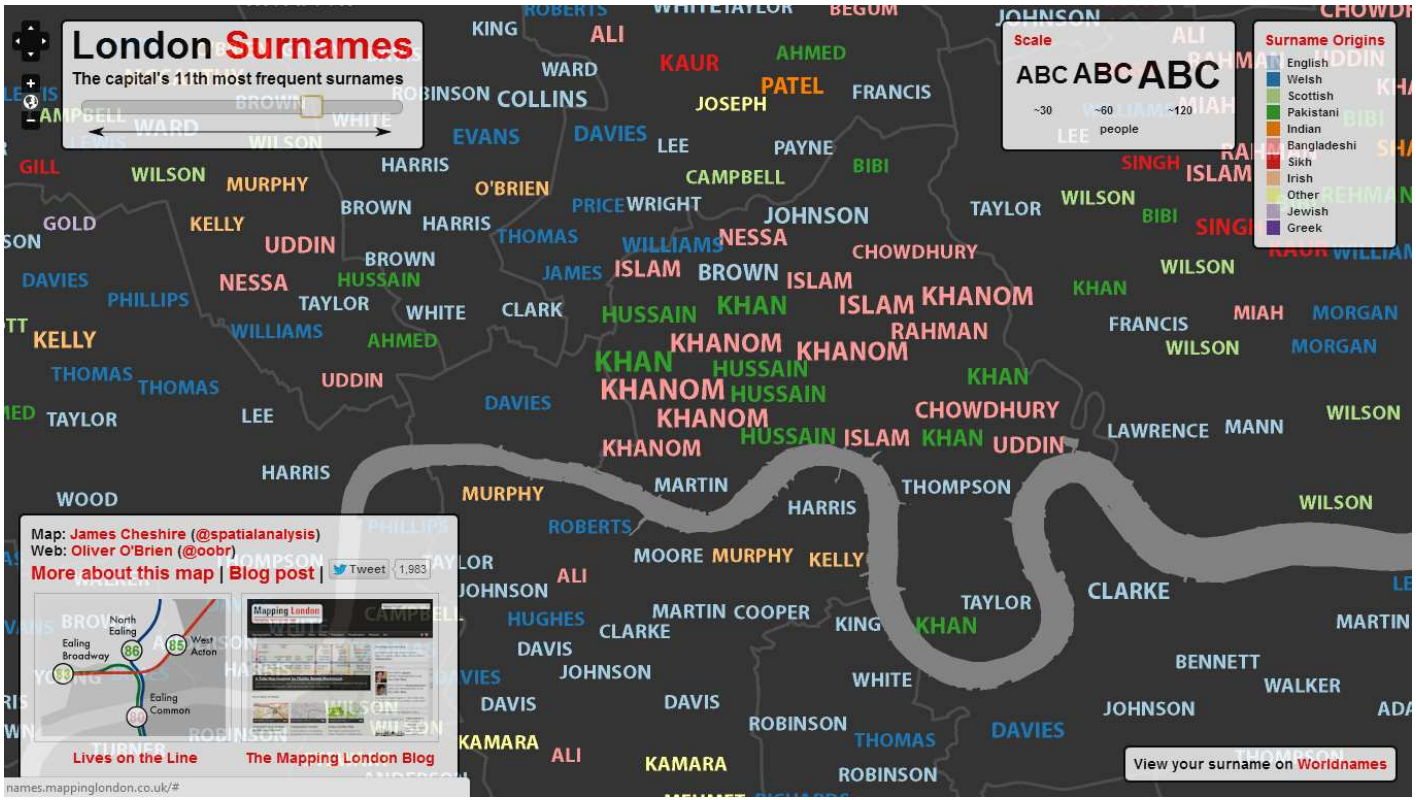


Figure: Frequently occurring surnames in London (from http://names.mappinglondon.co.uk/)

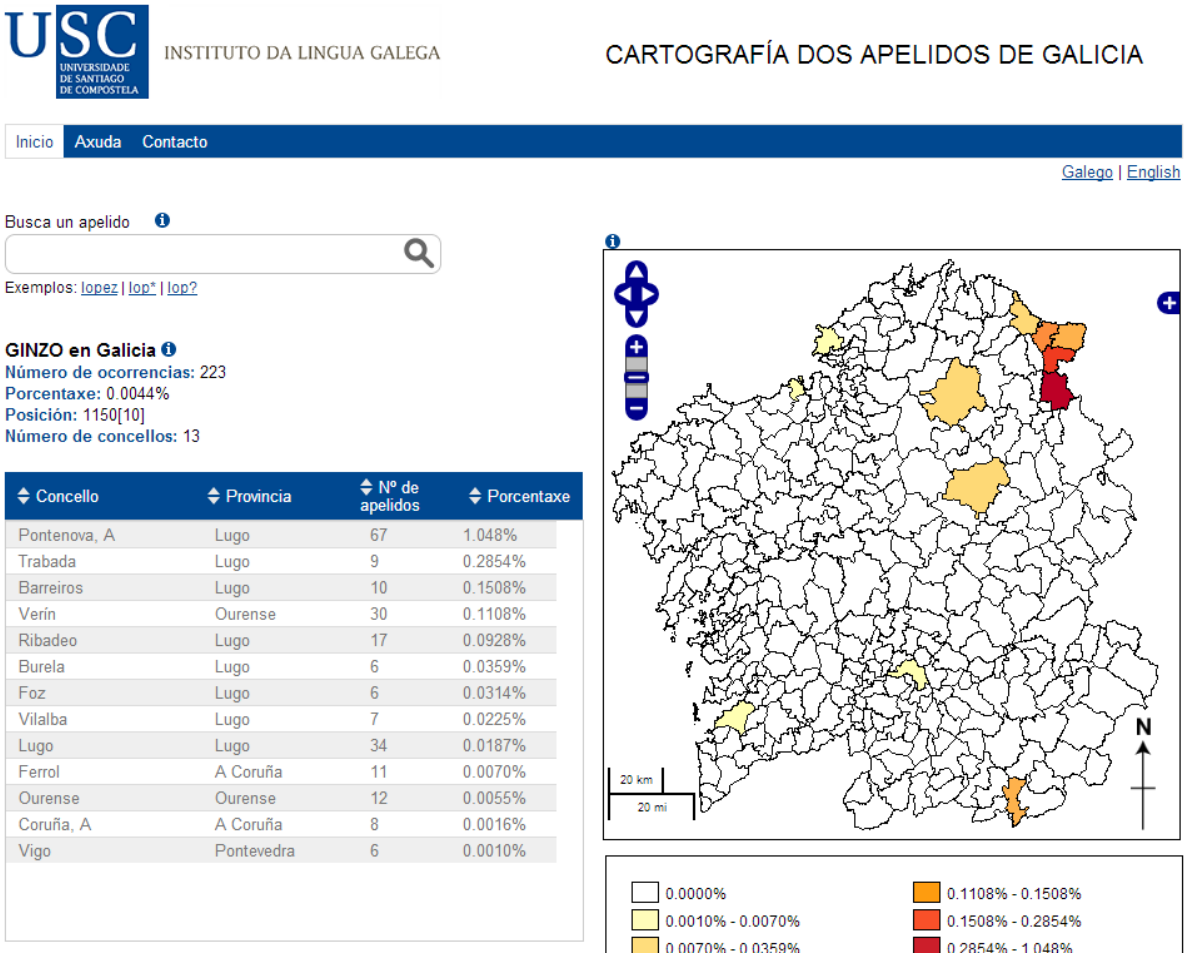


Figure: Surname Ginzo in Galicia (from http://ilg.usc.es/)

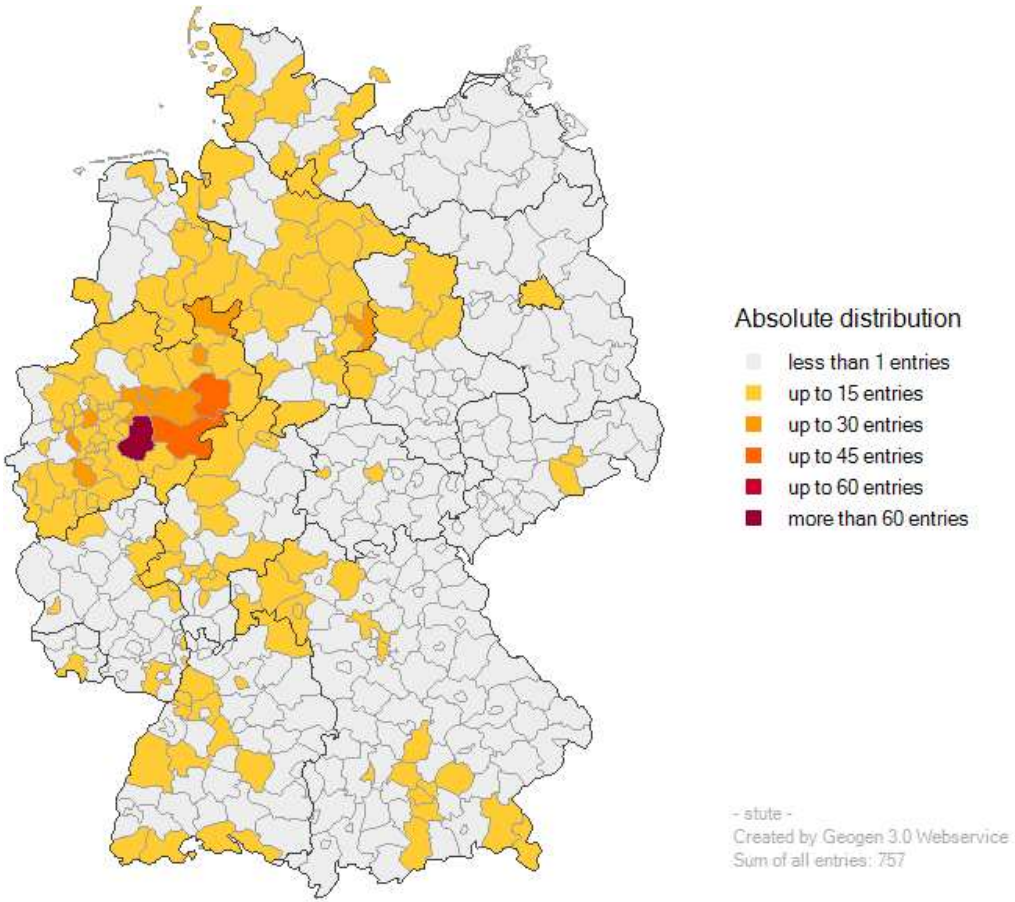
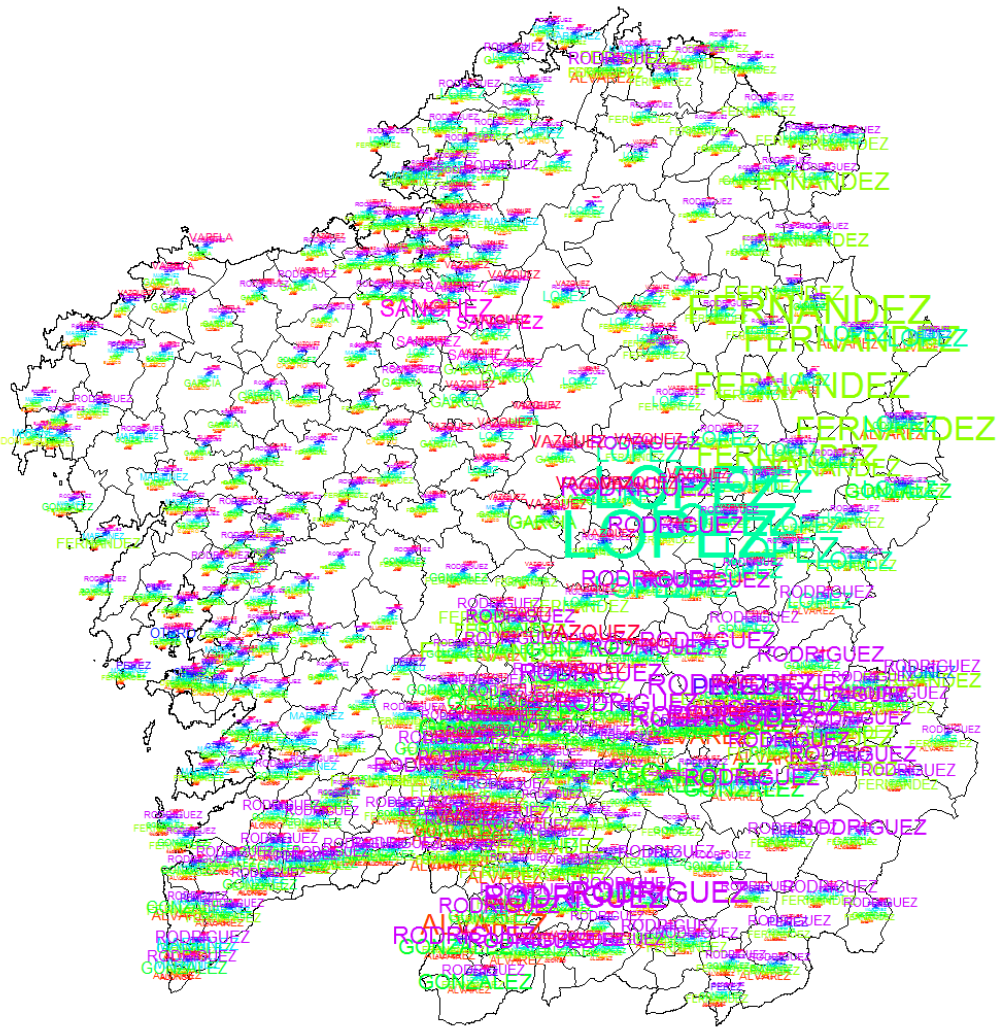


Figure: Surname Stute in Germany (from http://christoph.stoepel.net/geogen/en/Default.aspx)

- **Albania:** http://www.jeffdonofrio.net/Donofrio%20Albanese/.
- **The Netherlands:** http://www.meertens.knaw.nl/nfb/index.php?taal=eng.
- **Belgium:** http://www.familienaam.be/.
- **Italian:** http://www.cognomix.it/mappe\discretionary{-}{}{}dei\discretionary
- **Poland:** http://polishgeno.com/?p=60.
- **Chile:** http://apellidos.dechile.net/.
- ...

	Surname	Frequencies
1	RODRIGUEZ	105704
2	FERNANDEZ	99981
3	GONZALEZ	77929
4	LOPEZ	74866
5	GARCIA	69074
6	PEREZ	56082
7	MARTINEZ	51036
8	VAZQUEZ	45802
9	ALVAREZ	35512
10	GOMEZ	30712
11	CASTRO	27313
12	IGLESIAS	25899
13	DIAZ	22297
14	SANCHEZ	21650
15	BLANCO	20791
16	OTERO	19943
17	ALONSO	19918
18	VARELA	19851
19	DOMINGUEZ	18888
20	REY	16403
21	SUAREZ	14728
22	LORENZO	13365
23	PIÑEIRO	11934
24	PEREIRA	10957
25	VIDAL	10663

Table: Frequently occurring surnames from Census 2011. Top 25.



Surname
LOPEZ
FERNANDEZ
RODRIGUEZ
VAZQUEZ
GONZALEZ
GARCIA
DIAZ

Table: Frequently occurring surnames from Census 2011 in Lugo

Figure: Frequently occurring surnames from Census 2011 in Lugo.

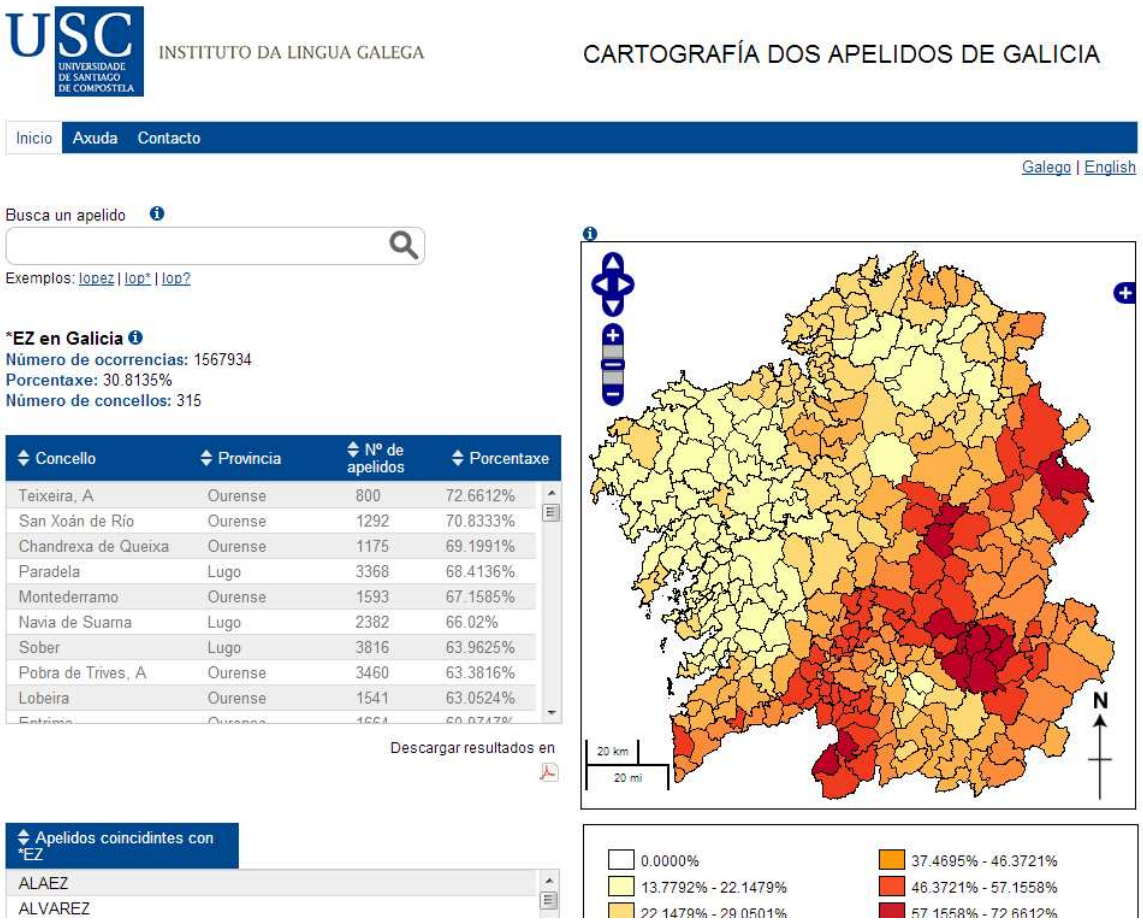


Figure: Search from a specific surname (from http://ilg.usc.es/cag/)

Motivation

- Surnames (family names) can be used as a source of information for population characteristics, given that the analysis of surname patterns provides information about long–term and short–term dynamics of population movements.

Objective

- By constructing clusters of surname zones, from different isonymy measures between regions, we aim to identify surname patterns, specially regionalized concentrations.

Some previous works

Isonymy measures in Galicia

References

- Cheshire *et al.* (2010): showed strong relationship between district surname and geographic locations in Great Britain, constructing clusters from surrounding districts based on Lasker distances.
- Boattini *et al.* (2010, 2012): analyzed the geographic location of different Italian surnames using neural networks, which allow for distinguishing monophyletic and polyphyletic surnames.
- Novotnỳ *et al.* (2012): studied the surname space of the Czech Republic, finding clear parallelism between their network representation and ethno–cultura boundaries in this country.
- Mikerezi *et al.* (2013): described the isonymic structure of Albania.

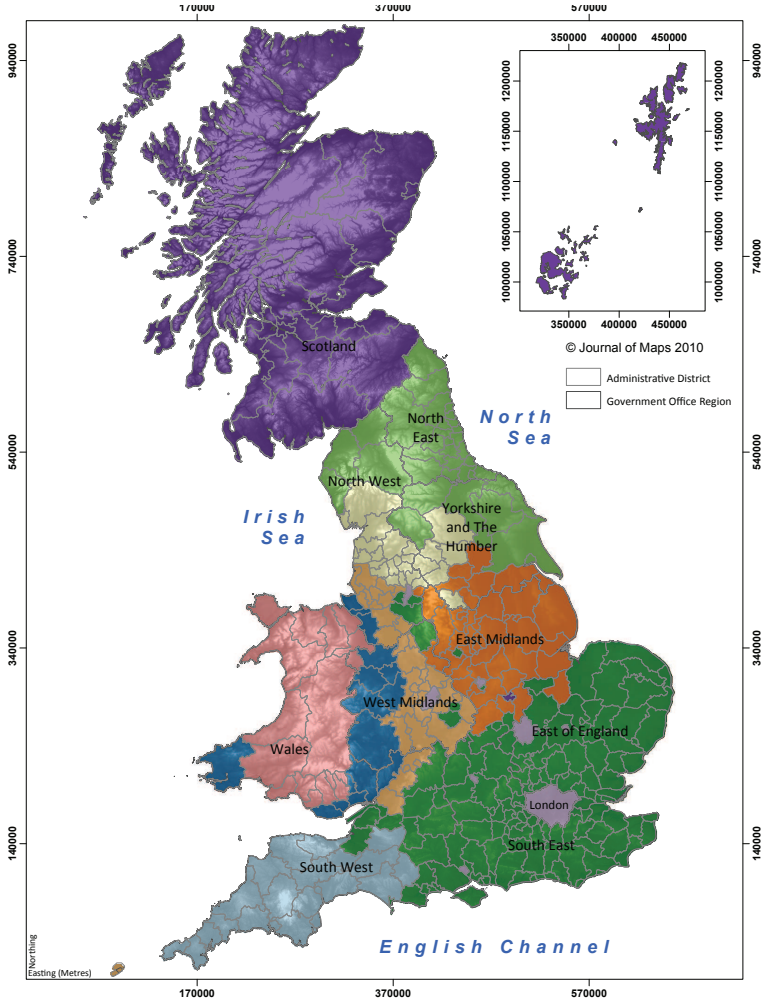


Figure: Cheshire e Longley (2011). Regions of surnames in GB.

Some notation

- Surname (dis)similarity among regions can be quantified by different measures.
- Index $i = 1, \ldots, n$ denotes a certain geographical region (for two regions, (i, j)).
- Each region has an associated collection S_i of surnames, and for a pair of regions, the collection of all the surnames is denoted by S_{ij}.
- The total number of surnames in a certain region i is denoted by n_i. Surnames will be denoted by index k.

Our data

- For the analysis of the Galician data, the regions considered were the 315 councils in Galicia.
- Continuous Municipal Census (January 1, 2011): 2,795.422 people.
- Number of different surnames: 20.754, corresponding to 2,430.512 people in 315 councils.
- **Warning:** Surnames that appear only in a council were removed, as well as those ones below and above the 5% and 95% quantiles of the distribution of number of councils.
- Data have been provided by Instituto Galego de Estatística (IGE).

Isonymy

Identification of surname patterns is usually made by isonymy (possession of the same surname). For a region i, the **isonymy** is defined:

$$I_i = \sum_{k \in S_i} p_{ki}^2,$$

where p_{ki} is the relative frequency of surname k in region i.

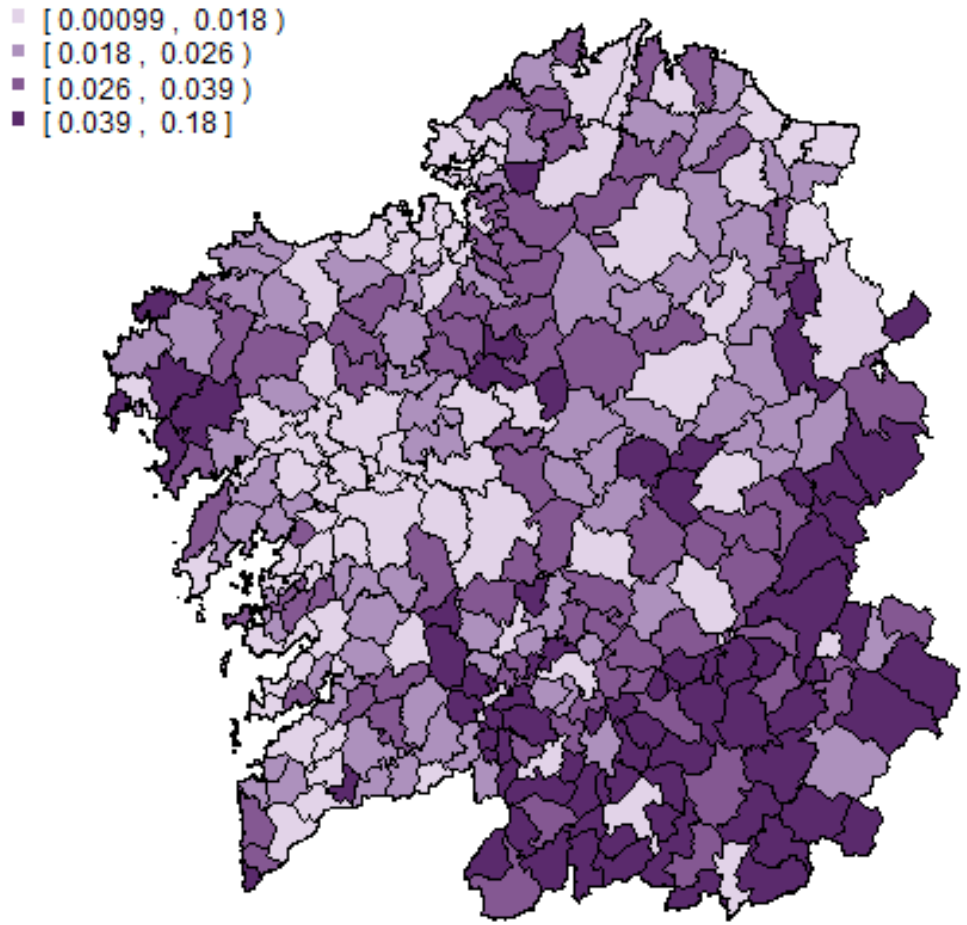


Figure: Isonymy for first surname in Galicia.

Isonymy for similarity

- Isonymy between two regions i and j: $I_{ij} = \sum_{k \in S_{ij}} p_{k_i} p_{k_j}$
- Euclidean distance: $E = \sqrt{1 - \sum_{k \in S_{ij}} \sqrt{p_{ki} p_{kj}}}$
- Lasker's distances: $L = -\log(I_{ij})$
- Nei's distance: $N = -\log\left(\frac{I_{ij}}{\sqrt{I_i I_j}}\right)$

- Nei's distance is highly correlated with google distance, computed from the councils centroids.
- Lasker and Euclidean distances do not present a strong correlation.
- Similar conclusions in other works, (for instance Mikerezi *et al.* (2013) with the surnames from Albania).

		UTM Distance	Google Distance	Isonymy Between	Lasker Distance	Nei Distance	Euclidean Distance
UTM Distance	Correlation	1	0.96800	-0.32953	0.51747	0.47856	0.51747
	Std. Error	0	0.00113	0.00425	0.00385	0.00395	0.00385
Google Distance	Correlation	0.96800	1	-0.32852	0.50847	0.48306	0.50847
	Std. Error	0.00113	0.00000	0.00425	0.00387	0.00394	0.00387
Isonymy Between	Correlation	-0.32953	-0.32852	1	-0.53697	-0.47007	-0.53697
	Std. Error	0.00425	0.00425	0.00000	0.00379	0.00397	0.00379
Lasker Distance	Correlation	0.51747	0.50847	-0.53697	1	0.96007	1
	Std. Error	0.00385	0.00387	0.00379	0.00000	0.00126	0.00000
Nei Distance	Correlation	0.47856	0.48306	-0.47007	0.96007	1	0.96007
	Std. Error	0.00395	0.00394	0.00397	0.00126	0.00000	0.00126
Euclidean Distance	Correlation	0.51747	0.50847	-0.53697	1	0.96007	1
	Std. Error	0.00385	0.00387	0.00379	0	0.00126	0

Table: Correlation matrix between distances.

- Once the aforementioned measures are obtained, the final output is a graphical representation of the different surname regions obtained by $\cdots$

Multivariate Analysis

- Representing the clusters given by dendrograms constructed from the matrices of Lasker's distances (Cheshire *et al.*, 2010), so the basic information of splitting or merging clusters is the similarity or isonymic distance between areas.
- The basic information for splitting or merging clusters is the similarity or distance between the clusters, and this distance can be obtained by different methods, such as complete linkage or Ward's procedure.

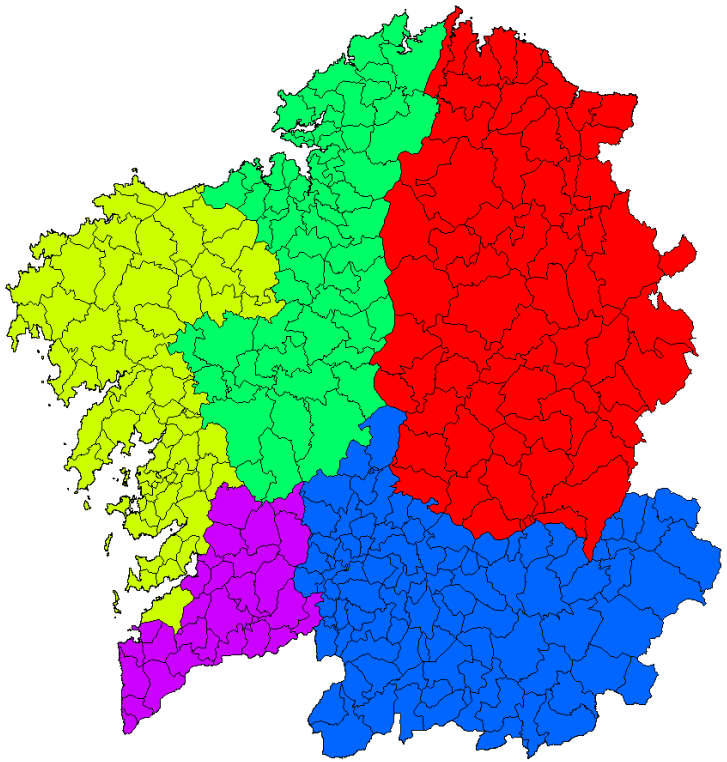


Figure: Surname clusters for Lasker's distances (5 groups)

Population

In Galicia, population movements towards urban areas began to be more important from the 70s of the last century. Consider the surname of people born in 1965 or earlier to become the new onomastic regionalization.

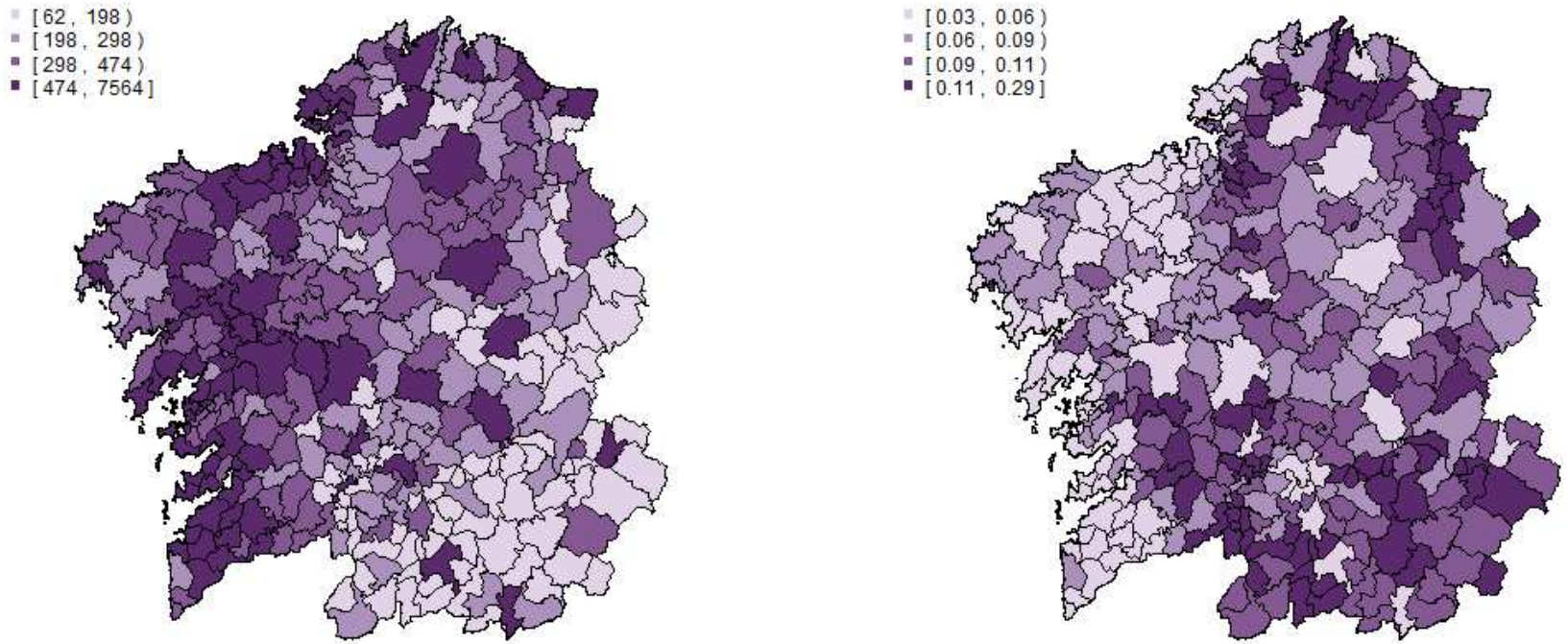


Figure: Number of different surnames by council. On the right part, the frequency is weighted by population. (People born in the year 1965 or earlier).

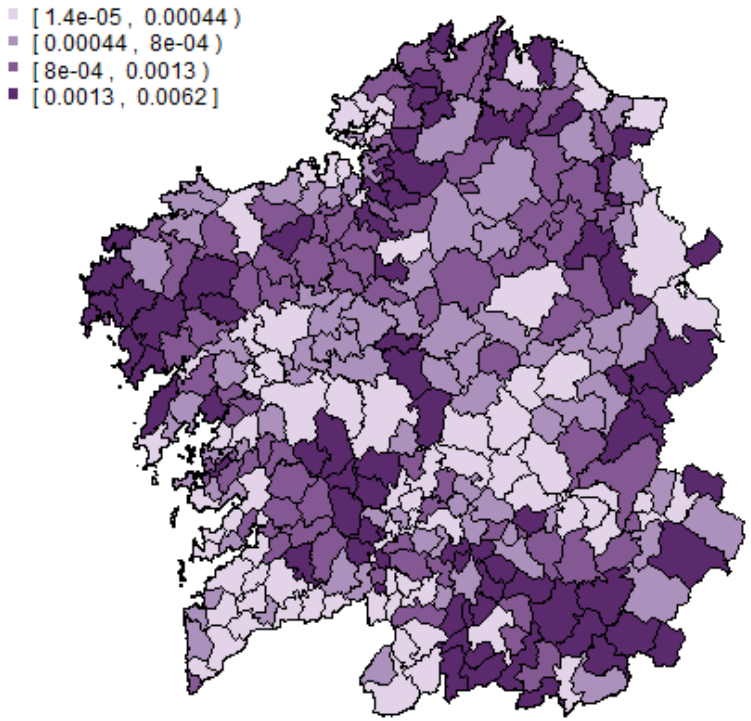


Figure: Isonymy for first surname of people born in the year 1965 or earlier.

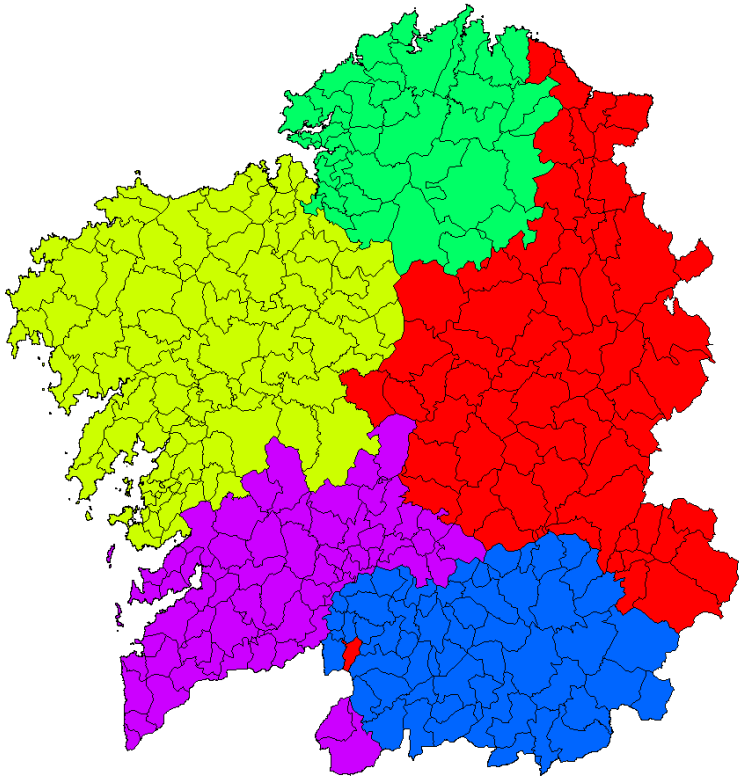
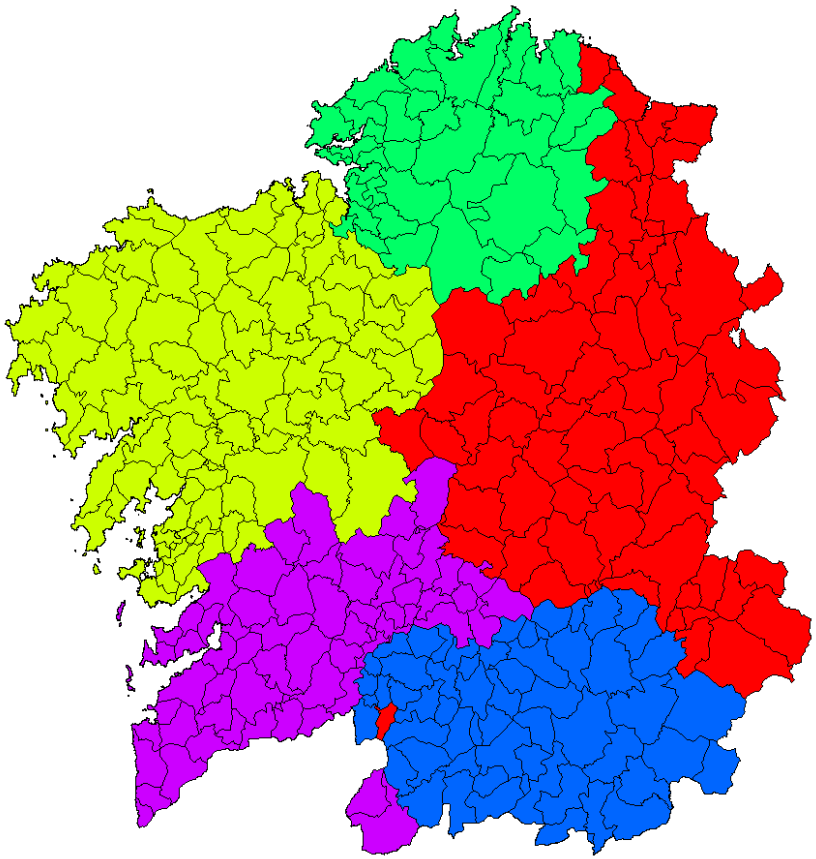
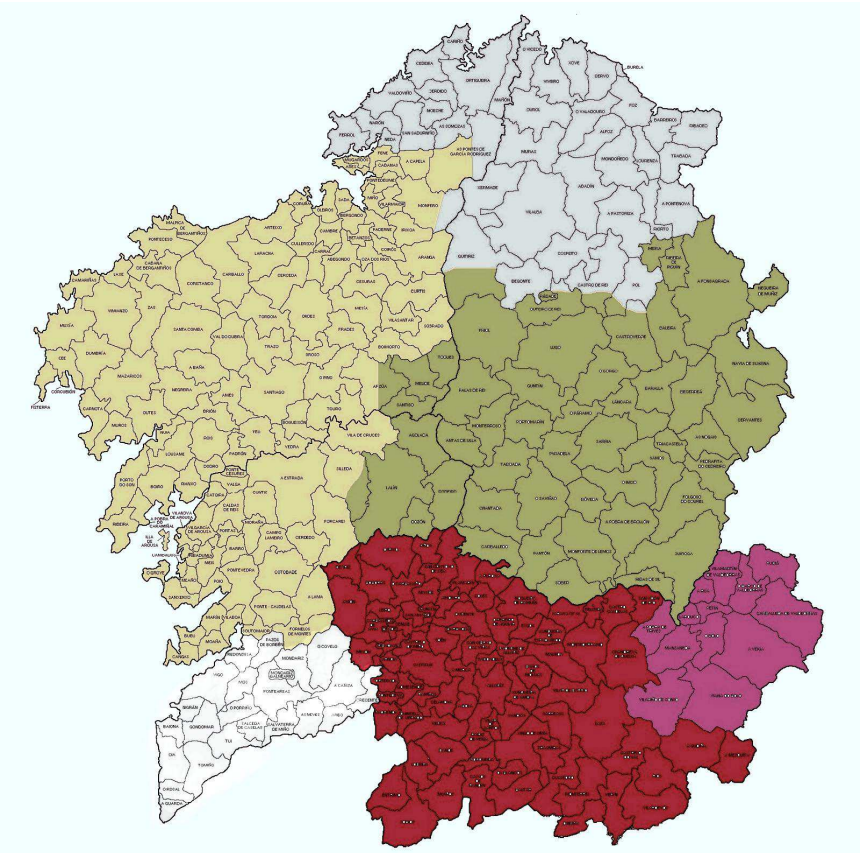
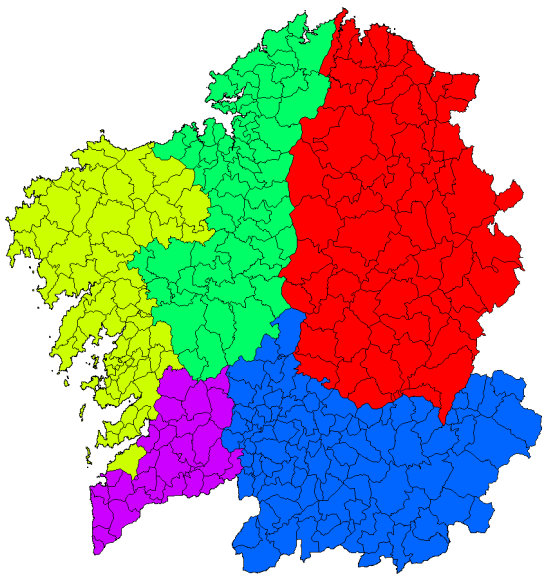
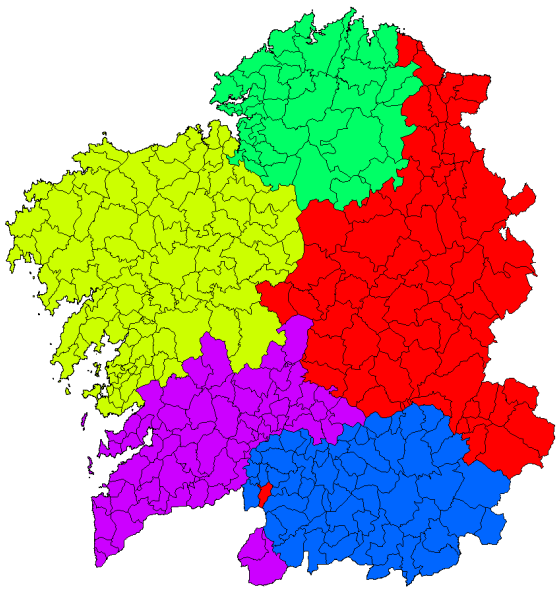


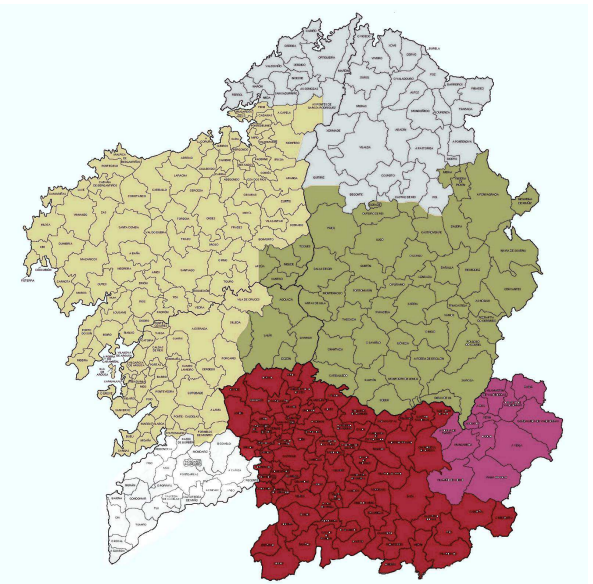
Figure: Surname clusters for Lasker's distance (5 groups). (People born in 1965 or earlier).











Some diversity indexes

- **Shannon index:** a higher index value indicates a greater biodiversity (maximum 5).

$$H_i = -\sum_{k \in Si} p_k \log p_k.$$

- **Simpson index:** it is often used to quantify the biodiversity of a habitat. It takes into account the number of species present, as well as the abundance of each species.

$$\text{Div}_i = \frac{n_i(n_i - 1)}{\sum_{k \in Si} n_k(n_k - 1)}$$

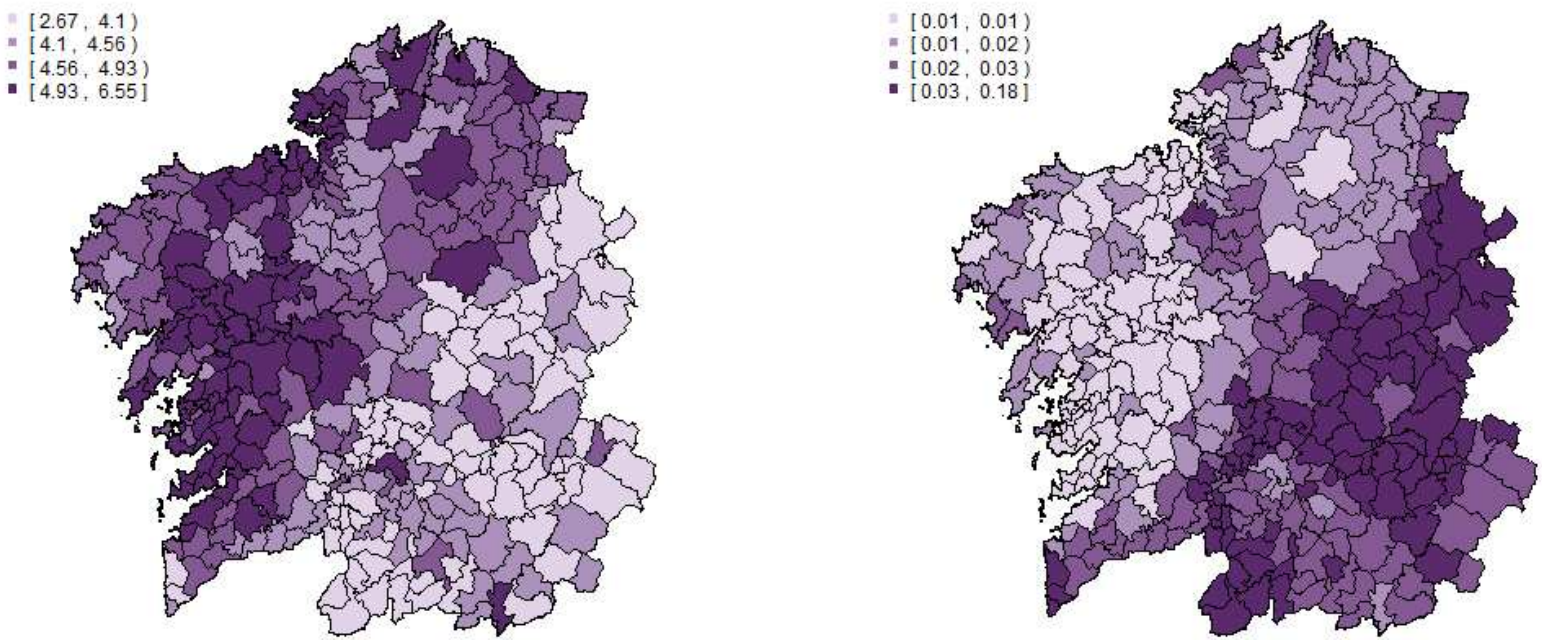


Figure: Shannon index (left) and Simpson index (right).

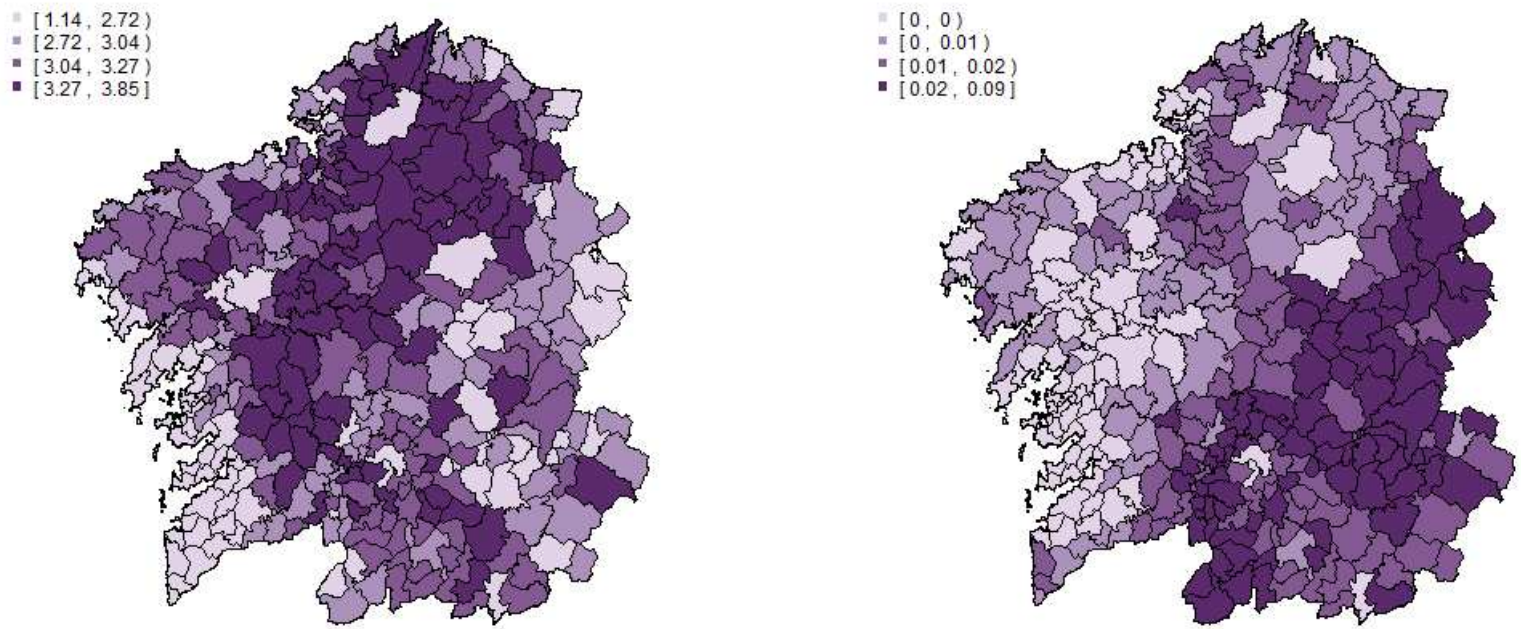


Figure: Shannon index (left) and Simpson index (right), people born in 1965 or earlier.

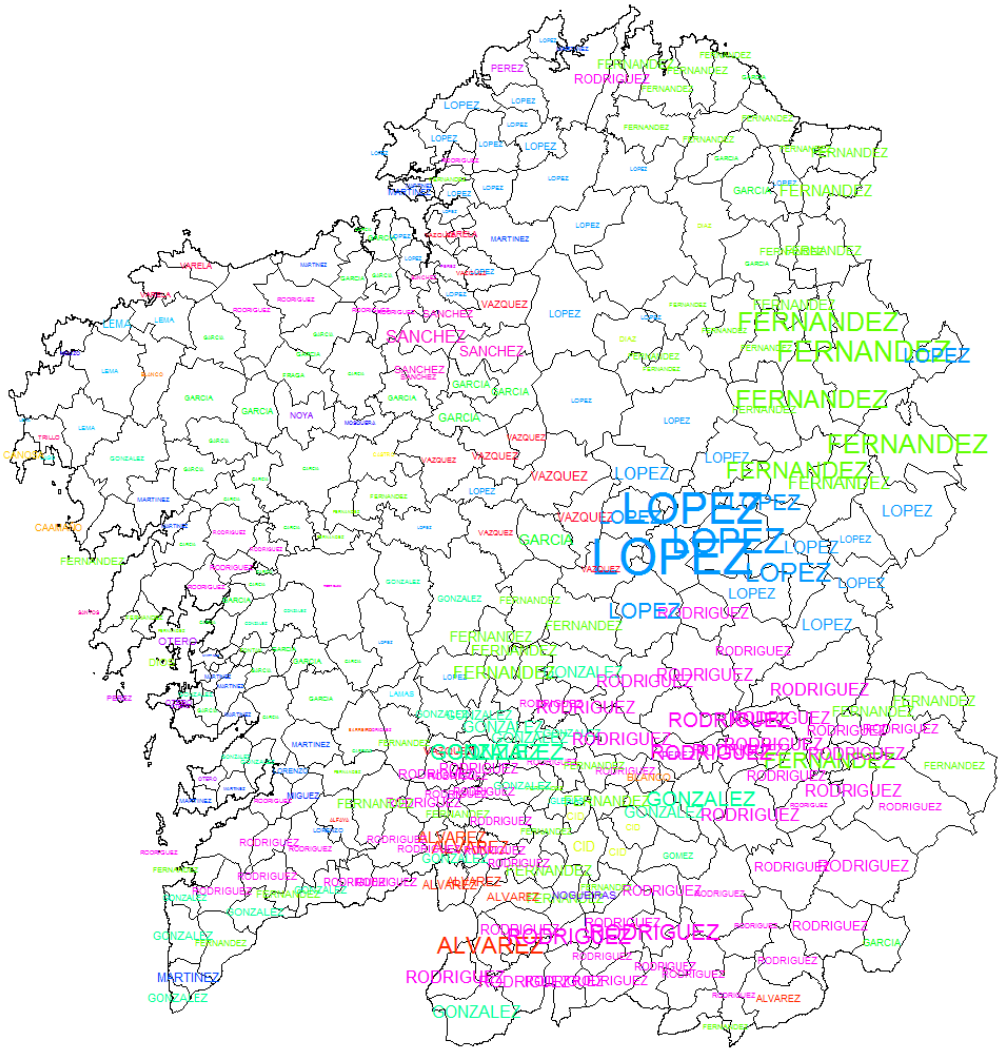


Figure: Frequently occurring surnames from Census 2011 by council.


Boattini, A., Lisa, A., Fiorani, O., Zei, G., Pettener, D. and Manni, F. (2012) General method to unravel ancient population structures through surnames, final validation on Italian data. *Human Biology*, **84**, 235–270.


Cheshire, J.A. and Longley, P.A. (2012) Identifying spatial concentrations of surnames. *International Journal of Geographical Information Science*, **26**, 309–325.


Cheshire, J.A., Longley, P.A. and Singleton, A.D. (2010) The surname regions of Great Britain. *Journal of Maps*, **6**, 401–409.


Mikerezi, I., Shina, E. Scapoli, C., Barbujani, G. Mamolini, E., Sandri, M., Carrieri, A., Rodríguez–Larralde, A. and Barrai, I. (2013) Surnames in Albania: a study of the population of Albania through isonymy. *Annals of Human Genetics*, **77**, 232–243.


Novotnỳ, J. and Cheshire, J. A. (2012) The Surname Space of the Czech Republic: Examining Population Structure by Network Analysis of Spatial Co-Occurrence of Surnames. *PloS one*, **7**, doi:10.1371/journal.pone.0048568.


Studeny, A. C. (2012) Quantifying biodiversity trends in time and space. *University of St Andrews*.

SURNAME PATTERNS IN GALICIA

María J. Ginzo-Villamayor, Rosa M. Crujeiras and Xulio Sousa

Universidad de Santiago de Compostela

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
INSTITUTO DA LINGUA GALEGA (ILG)