XI Congreso Galego de Estatística e Investigación de Operacións
A Coruña, 24–25–26 de outubro de 2013

SURNAME PATTERNS IN GALICIA

María J. Ginzo-Villamayor[1], Rosa M. Crujeiras[1] and Xulio Sousa[2].

[1]Dpt. of Statistics and Operations Research. University of Santiago de Compostela (Spain)
[2]Dpt. of Galician Philology. University of Santiago de Compostela (Spain)

ABSTRACT

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. With the aim of identifying surname patterns, specially regionalized concentrations, different isonymy measures between regions can be computed and from these ones, clusters of surname zones can be constructed. In this work, an analysis of the surname patterns in Galicia is provided, considering a variety of isonymy indexes and different clustering methods.

Keywords: clusters, isonymy measures, lasker distance, surnames.

1. INTRODUCTION

The analysis of the geographical distribution of surnames allows to study the spatial and temporal human population structure. The spatial information obtainable from surnames, combined with their ubiquity, makes them a rich resource for regionalization studies. Despite showing distinctive geographical patterning, surnames have remained an underutilised source of information about population origins, migration and identity. Indeed, individuals who share location specific surnames are also likely to share a number of linguistic, genetic, historical and social characteristics as well as common ancestry.

The identification of surname patterns through isonymy (possession of the same surname) measures have been previously addressed by several authors. Cheshire *et al.* (2010) show a 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) have studied the surname space of the Czech Republic, finding clear parallelism between their network representation and ethno–cultura boundaries in this country. Recently, Mikerezi *et al.* (2013) describes the isonymic structure of Albania. From a different perspective, Cheshire and Longley (2012) consider a bidimensional kernel smoother to detect areas in Great Britain where certain surnames are more concentrated. Using this methodology, the authors identify some clustered surnames, such as Bamber or Macleod. However, given that the distribution of surnames can be viewed as a spatial point process, a kernel intensity smoother may be more accurate for analyzing this type of data.

In order to study the isonymy structure of Galicia, the surname distribution of 2430512 people from the 2011 census accross the 315 councils was analyzed. Different isonymy measures (specifically, Lasker and Nei's distances and Fisher's α coefficient) have been considered, and from Lasker distances, surname clusters have been identified. From the resulting regionalization, cultural, linguistic and genealogocial information about the population in Galicia can be inferred. It should be noted that the analysis is both data rich and computationally intensive.

This short paper is organized as follows. In Section 2, a brief description of the methods is provided. Section 3 includes a summary of the results and some final comments.

2. ISONYMY MEASURES

Surname (dis)similarity among regions can be quantified by different measures. Consider index $i = 1, \ldots, n$ for denoting 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 in them is denoted by S_{ij} ($S_{ij} = S_i \cup S_j$). The total number of surnames in a certain region i is denoted by n_i. Surnames will be denoted by indices k and l.

As mentioned in the Introduction, isonymy refers to the possession of the same surname, being a premise in genetics that individuals with the same surname are more likely to share the same family lineage, so isonymy indicates biological relation. With the notation introduced above, isonymy (as an internal measure, within a region i) is defined as

$$I_i = \sum_{k \in S_i} p_{ki}^2, \tag{1}$$

where p_{ki} denotes the relative frequency of surname k in region i. High values of isonymy are possible in a population where there are relatively few surnames, and low values of isonymy are obtained when the number of surnames is relatively large. From an analogy with genetics, as it happens for alleles, drift of surnames is proportional to time, and then small values of isonymy suggests recent immigration or settlement.

Another useful measure is Fisher's α (Barrai, 1996), which estimates the number of surnames having equal frequency. A small value of this coefficient indicates large inbreeding and drift, whereas a large value results for migration and low inbreeding. This measure is a diversity index and for large samples, as this case, it can be define as $\alpha_i = 1/I_i$, for an specific region i.

Isonymy can be also extended as a measure of population similarities between groups. Under the assumption of a common origin, isonymy between two regions i and j is defined as:

$$I_{ij} = \sum_{k \in S_{ij}} p_{k_i} p_{k_j}. \tag{2}$$

Other different measures of the isonymic distance between a pair of locations can be derived from (2). For instance, the Lasker distance is given by:

$$L = -\log(I_{ij}) \tag{3}$$

Lasker distances can be interpreted as a measure of similarity between to areas, where large distance indicate less similarity in surname composition. Nevertheless, Lasker distance is not the only option to quantify surname similarity. Other common coefficients are the Euclidean distance, introduced by Cavalli–Sforza and Edwards (1997) and Nei's distance (Nei, 1973), both of them given by:

$$E = \sqrt{1 - \sum_{k \in S_{ij}} \sqrt{p_{ki} p_{kj}}} \quad \text{and} \quad N = -\log\left(\frac{I_{ij}}{\sqrt{I_i I_j}}\right), \quad \text{respectively.} \tag{4}$$

Euclidean and Nei's distances have been developed for purely genetic data, but they can be applied to the frequencies of surnames, such as done by Mikerezi *et al.* (2013). In addition, in order to detect isolation by distance between locations i and j, the linear correlation of surname distances (Lasker's, Euclidean and Nei's) with their geographic distances can be computed.

Once the aforementioned measures are obtained, the final output is a graphical representation of the different surname regions. This is usually done by representing the clusters given by dendrograms constructed from the matrices of Lasker's distances (see 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.

For the analysis of the galician data, the locations considered were the 315 councils in Galicia, although different aggregations are also possible (by provinces and for the whole region). 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. In Figure 1, the distribution of surname frequencies (number of councils where the surname appears) is shown. It can be seen that this distribution is highly asymmetric, similar to other studies. In the horizontal axis, the different surnames should be located (labels are omitted, for the sake of clarity). Just as a summary, in the left part, surnames with highest appearance frequencies accross councils should be placed, corresponding to Alén and Chaves. On the contrary, in the right part of the axis, Aatioum and Abagashould appear (foreign surnames). Among the removed surnames are those ones with at least one representative in each council (Fernández, González, López, Pérez, Rodríguez). As explained above, also those surnames present in a single council are also ignored. Some examples are Cotofre, Larrán and Roales.

Figure 1: Distribution of surname frequencies, removing surnames that appear just in a council and those ones below and above the 5% and 95% quantiles of the appearance frequencies.

3. SOME RESULTS AND DISCUSSION

The analyzed data corresponds to 20754 different surnames in Galicia. Nei's distance (4) is highly correlated with geographic distance, computed from the councils centrodis. Lasker and Euclidean distances (from (3) and (4), respectively) do not present a strong correlation. The correlation matrix between distances is presented in Table 1.

		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
Geographical 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 1: Matrix of correlations between distances.

The effective surname number, Fisher's α (inverse of isonymy in (1)), in Galicia is 1857 for the whole country considered as a unit. The average over the 315 councils is 30. When compare all

Galicia with councils, these values indicate that the estimates of α from this source varies with the size of the area and of the population studied. This is in part explained by the difference in frequencies of common surnames in each subdivision. In Table 2, it can be observed that α is higher in large areas and populations as compared to smaller subdivisions.

	Council	Province	Galicia
I_i	0.03305	0.00140	0.00054
α	30	714	1857

Table 2: Isonymy (1) and α for Galicia. Average isonymy and α for councils and provinces.

Following the methodology presented in Section 2, clusters of councils based on Lasker distances between them can be obtained. Some preliminary results reveal a unique and evidence–based regional geography that is of use in improving our understanding of cultural and social history. This study found clear regionalisation patterns in surname frequency distributions, closely matching the historical borders for five diocesan boundaries. The resulting regionalisation demonstrates the utility of an innovative inductive approach to summarizing and analyzing large population datasets across cultural and geographic space, the outcomes of which can provide the basis to hypothesis generation about social, cultural and historic patterning. The research also contributes a range of methodological insights for future studies concerning spatial clustering of surnames.

REFERENCES

Barrai, I., Scapoli, Beretta, M.N., Mamolini, E. and Rodríguez–Larralde, A. (1996) Isonymy and the genetic structure of Switzerland I. The distributions of surnames. *Annals of Human Biology*, **23**, 431–455.

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.

Boattini, A., Pedrosi, M.E., Luiselli, D. and Pettener, D. (2010) Dissecting a human isolate: Novel sampling criteria for analysis of the genetic structure of the Val di Scalve (Italian Pre-Alps). *Annals of Human Biology*, **37**, 604–609.

Cavalli–Sforza, L.L. and Edwards, A.W.F. (1967) Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics*, **19**, 233–257.

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.

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.

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.

Nei, M. (1973) The theory and estimation of genetic distance. *Genetic structure of populations (ed. N. E. Morton)*. Hawaii: Hawaii University Press.