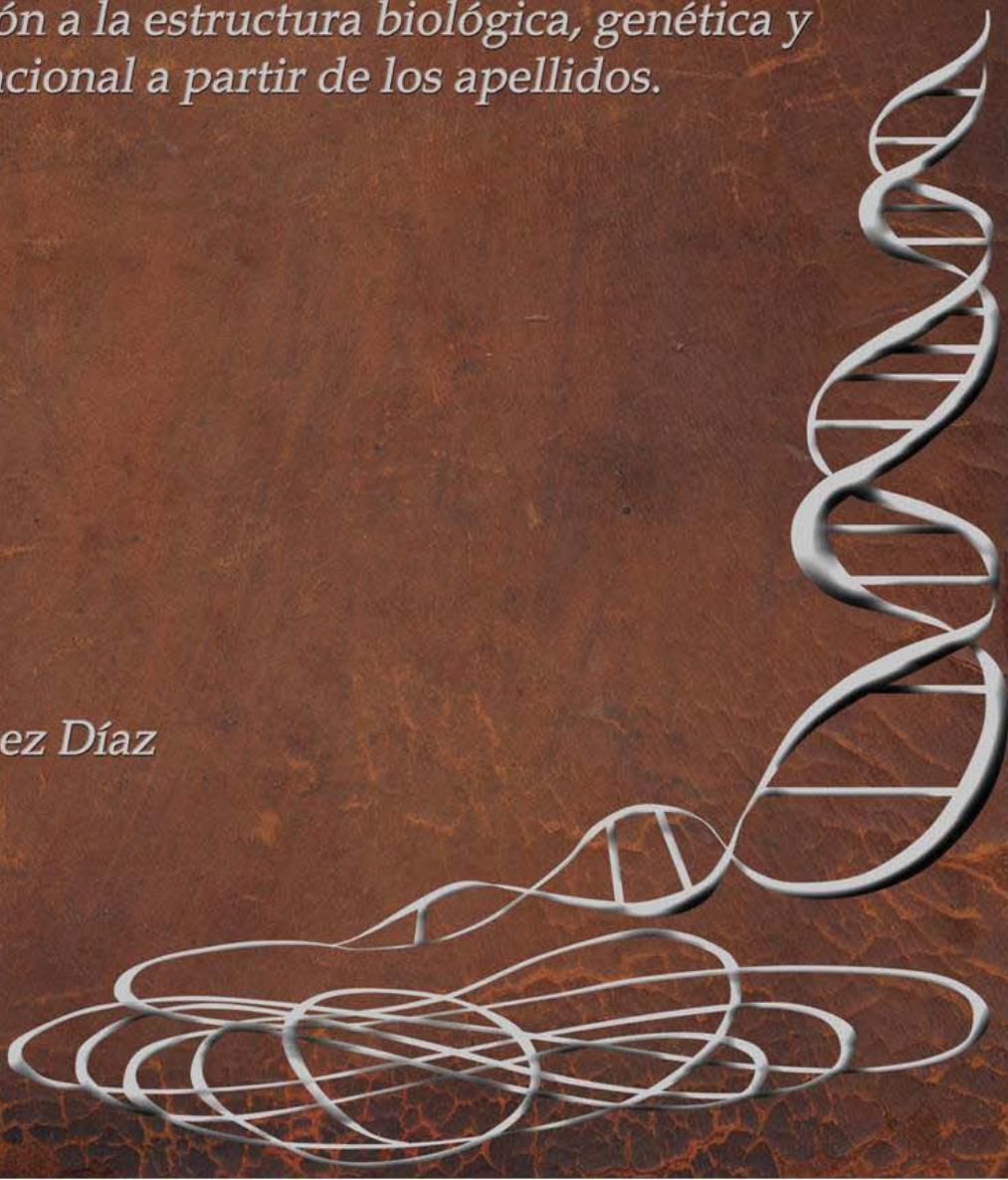
FACULTAD DE BIOLOGÍA

DEPARTAMENTO DE BIOLOGÍA ANIMAL, ECOLOGÍA,
PARASITOLOGÍA, EDAFOLOGÍA Y QUÍMICA AGRÍCOLA
ÁREA DE ANTOPOLOGÍA FÍSICA

VNiVERSiDAD
D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

LA POBLACIÓN ESPAÑOLA.

Aproximación a la estructura biológica, genética y poblacional a partir de los apellidos.



Roberto Rodríguez Díaz
Salamanca, 2015

FACULTAD DE BIOLOGÍA
DEPARTAMENTO DE BIOLOGÍA ANIMAL, ECOLOGÍA, PARASITOLOGÍA, EDAFOLOGÍA Y QUÍMICA AGRÍCOLA
ÁREA DE ANTOPOLOGÍA FÍSICA

LA POBLACIÓN ESPAÑOLA.

Aproximación a la estructura biológica, genética y poblacional a partir de los apellidos.

Memoria presentada por Roberto Rodríguez Díaz para optar al título de Doctor en Ciencias Biológicas, dirigida por los doctores
Doña María José Blanco Villegas y Don Franz Manni.

Salamanca 2015

Roberto Rodríguez Díaz

Los doctores Dña. María José Blanco Villegas del Departamento de Biología Animal, Parasitología, Ecología, Edafología y Química Agrícola de la Universidad de Salamanca y D. Franz Manni del Musée de l'Homme - National Museum of Natural History,

AUTORIZAN:

La presentación, para su lectura, de la Tesis Doctoral titulada "La población española. Aproximación a la estructura biológica, genética y poblacional a partir de los apellidos.", realizada por Don Roberto Rodríguez Díaz bajo su dirección en la Universidad de Salamanca.

Firmado:

María José Blanco Villegas

Franz Manni

Agradecimientos

No puedo empezar esta página de otra forma que expresar mi agradecimiento a Mache. Por muchas más cosas que su dirección académica.

Lo siguiente que quiero es pedir perdón, porque seguro que en estas palabras apresuradas me olvidaré de alguien de quien no debería olvidarme.

Gracias a mis directores de tesis, a Mache y a Franz, por lo que he tenido oportunidad de aprender estando cerca de ellos.

A Patri por su ayuda y paciencia y a mi madre por su apoyo. A toda mi familia especialmente a los que me gustaría que estuviesen aquí. A todos mis amigos por preocuparse por mi y apoyarme.

A los compañeros de departamento, a Víctor por toda la ayuda que siempre me ha prestado, a Valen y a Pablito, por compartir divagaciones y cafés, a Miguel, por las lecciones de historia.

"Si tu intención es describir la verdad,

hazlo con sencillez y la elegancia déjasela al sastre"

Albert Einstein

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

0. LA POBLACIÓN ESPAÑOLA Y SUS APELLIDOS

A fin de facilitar su lectura, en este trabajo, se ha optado por incluir la introducción, el material y los métodos en cada uno de los capítulos. Sin embargo parece conveniente añadir este capítulo previo con la explicación de algunas características de la base de datos utilizada y la población española.

LA POBLACIÓN ESPAÑOLA Y SUS APELLIDOS

La base de datos de los apellidos españoles está construida a partir de los datos contenidos en el padrón continuo de 2008. La biología de las islas responde a sus características especiales y deben ser tenidas en cuenta de manera independiente. Por ello, el trabajo se centra en la población peninsular.

La población española en 2008

En el año 2008 la población española es de 45.593.385 habitantes, 42.055.095 de ellos en el territorio peninsular. Desde 1787 la población española ha venido aumentando de manera constante (Figura 1). Sin embargo esta tendencia suave y constante no se observa igual a nivel provincial.

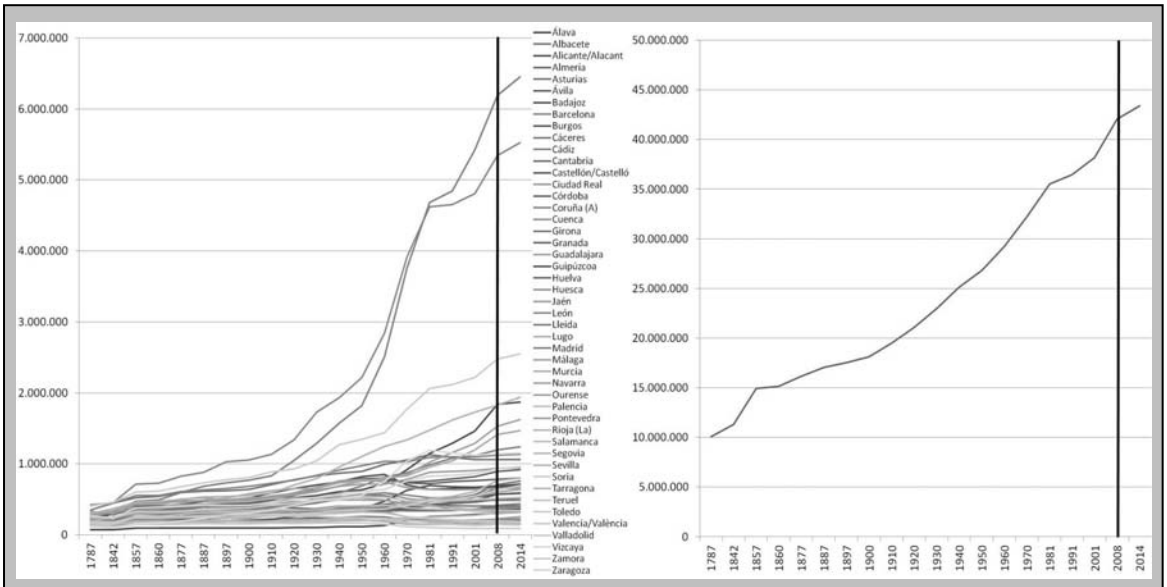


Figura 1. Evolución de la población española en cada provincia (Izquierda) y a nivel nacional (Derecha).

La población de cada provincia se mantiene en constante y suave aumento hasta los años 1950 y 1960 (Figura 1). A partir de esos años algunas provincias empiezan a perder población o a aumentarla más lentamente en favor de otras provincias que aumentan su población muy rápidamente. El cambio se produce bruscamente y obedece a un fenómeno migratorio que se produce hacia esos años hacia los centros de industrialización (Fuster y Colantonio, 2002). Este fenómeno cambia sustancialmente una población que hasta entonces era básicamente sedentaria y altera la distribución de la población.

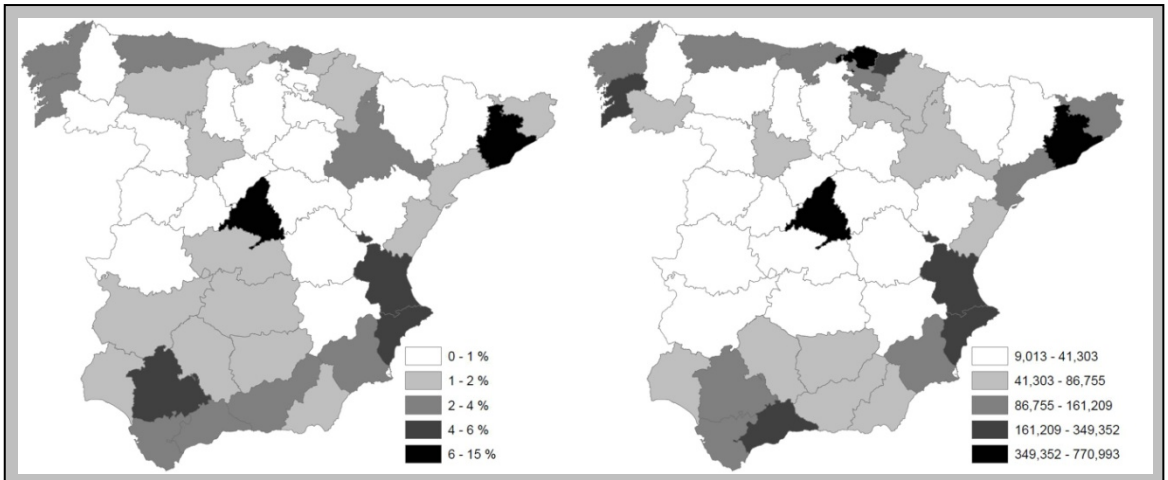


Figura 2. Porcentaje de la población total que representa cada provincia (Izquierda). Densidad poblacional de cada provincia (Derecha).

Probablemente como resultado de estos desiguales movimientos, la distribución geográfica de la población es muy desigual (Figura 2). El 50 % de la población española vive sólo en siete provincias Murcia, Málaga, Sevilla, Alicante, Valencia, Barcelona y Madrid, de hecho, estas dos últimas (Barcelona y Madrid) representan ellas solas el 27,42 % de la población peninsular española.

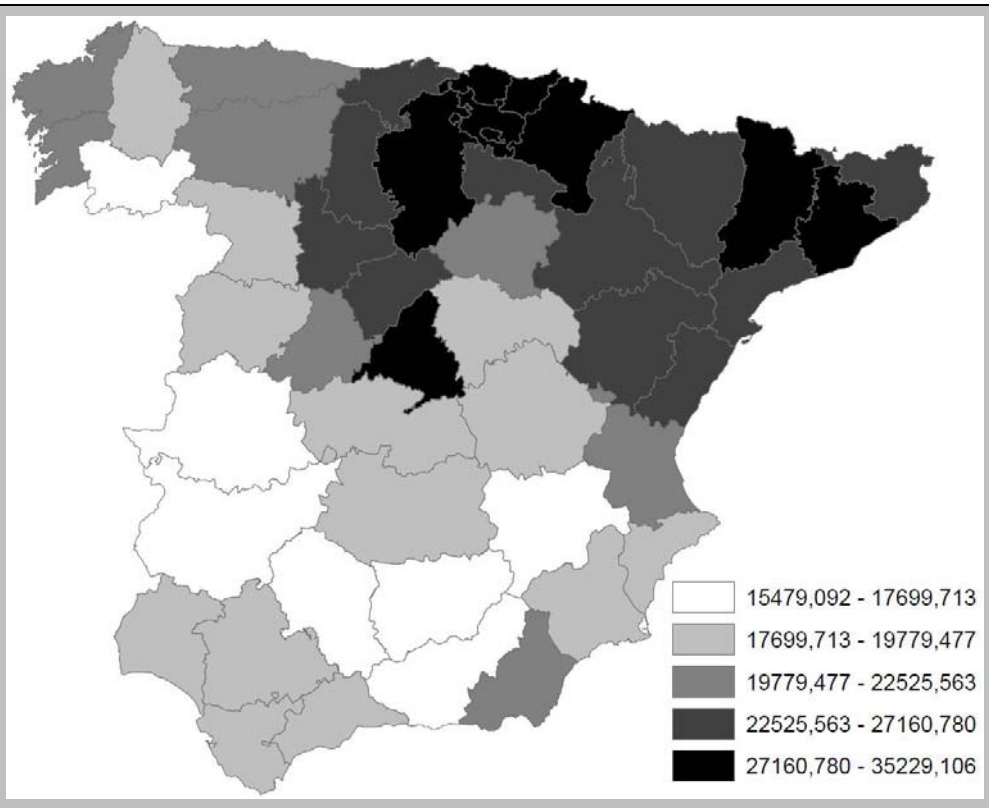


Figura 3. PIB en Euros por persona.

La población se concentra en zonas de alta densidad poblacional (Figura 2) distribuidas por las costas mientras que el centro peninsular (con la única excepción de Madrid) muestra una densidad poblacional mucho más baja.

Cabría esperar que esta distribución poblacional guardase alguna relación con la distribución de los recursos económicos. Sin embargo, esto no resulta exactamente así. Si observamos la distribución del PIB en euros por habitante (Figura 3).

Por lo general si parece que el nivel económico de las zonas costeras es más alto que el del interior. Pero ese no es el patrón de distribución de la riqueza dominante. Parece que se concentra principalmente en la mitad Norte y muy especialmente en el Norte-Oeste.

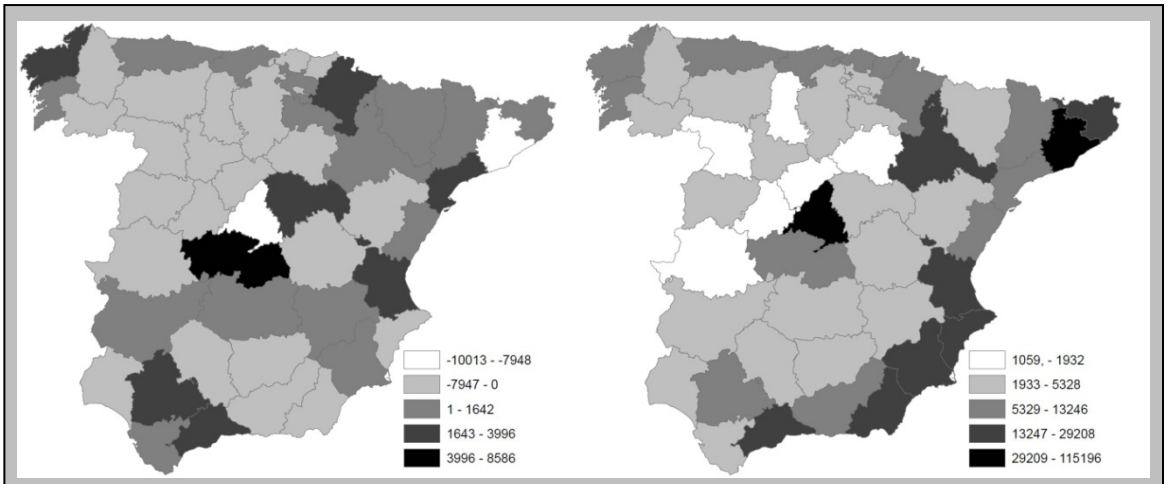


Figura 4. Saldos migratorios durante 2008. A la Izquierda el saldo migratorio interior. A la Derecha el saldo migratorio exterior.

Los movimientos migratorios exteriores a lo largo de 2008 (Figura 4) si guardan una similitud evidente con la densidad poblacional. Se concentran muy especialmente en la costa mediterránea y en menor medida en la cantábrica, las zonas más densamente pobladas de la península. En el centro peninsular, sin embargo, sólo Madrid (que por otro lado es el principal centro receptor) tiene alguna importancia en la zona centro.

En cuanto a las migraciones interiores, muestran un patrón más complejo (Figura 4). Prácticamente todo el noroeste peninsular y el sur pierden población en favor principalmente del noreste, la costa cantábrica, Sevilla y sus alrededores y los alrededores de Madrid. Es interesante que en el año 2008 Madrid y Barcelona hayan sido los principales centros de recepción de población extranjera y, sin embargo, presenten un saldo migratorio negativo en cuanto a movimientos internos.

Los apellidos

Más adelante, en cada capítulo se profundizará en el tratamiento de la base de datos que se ha hecho para cada caso particular. Por esa razón no se profundizará demasiado aquí pero hay algunos aspectos que conviene aclarar previamente.

La base de datos con la que se ha trabajado está compuesta por todos los apellidos que aparecen al menos 5 veces en al menos un municipio en el padrón continuo de 2008. En la base de datos en bruto sólo viene reflejada la frecuencia con la que un apellido aparece en un determinado municipio como primer apellido, como segundo apellido o en ambos lugares. Sabemos que en el caso del sistema de apellidos español en el que se usan dos apellidos, resulta recomendable utilizar ambos ya que multiplicamos la cantidad de información disponible y la robustez de los análisis (Pettener et al., 1998; Colantonio et al., 2003; Blanco-Villegas et al., 2004; Dipierri et al., 2011; Mateos y Tucker, 2008; Dipierri et al., 2011).

El resultado es una base de 56.976.706 datos, correspondientes a 87.148 formas de apellidos diferentes. El problema es que la base de datos en bruto contenía multitud de errores. Por ejemplo:

- Los apellidos con nombres se santo podían aparecer como una sola palabra o como dos palabras (Sanmartín o San Martín).
- El mismo apellido aparece con o sin tilde (Díaz o Diaz).
- El mismo apellido aparece con o sin un artículo delante (Torre o De la Torre).
- El mismo apellido aparece con diferentes grafías (Giménez o Jiménez).
- Algunos apellidos aparecen fusionados (García de la Torre).
- ...

Todos estos errores fueron depurados minuciosamente. Para ello se utilizó apoyo cartográfico y bibliográfico (Fauré et al., 2001; Solís, 2002) para establecer un doble control a la hora de la corrección. De esta manera sólo se corrigieron aquellas formas de las que podíamos estar seguros lingüísticamente que correspondían a un mismo apellido y aquellas formas que geográficamente son característicos de una misma población. En total se revisaron 24.252 formas y corrigieron 12.920, quedando 74.228 apellidos diferentes.

Una vez corregida la base de datos se han eliminado los apellidos menos frecuentes. Con ello se pretende evitar ruido que distorsione los resultados

pero perder la mínima información posible. Se ha elegido un umbral de 20 apariciones mínimo en toda la península.

Se ha elegido este umbral mínimo por dos razones. La primera es que la fecundidad media en 2008 según los datos del INE (www.ine.es) es de 4,12 hijos para una mujer de naturaleza española y 6,32 para una mujer inmigrante. Según estos números, una unidad familiar media española constará de 6-7 integrantes y una inmigrante de 8-9. Eligiendo 20 apariciones como umbral mínimo, nos garantizamos que el apellido representa al menos a dos familias.

Por otro lado, eliminar los apellidos que aparecen menos de 20 veces supone eliminar unos 41.000 apellidos de los 74.228 que tenemos, pero menos de 300.000 de los 56.976.706 datos que tenemos, por lo que apenas perdemos información, en torno al 0,6 % de la información que manejamos.

Por último se eliminó toda la información correspondiente a las poblaciones situadas fuera de la península. El resultado es una base de datos con 51.419.788 datos, correspondientes a 33.453 apellidos diferentes.

España es un país con una distribución de apellidos particular. Ya el propio sistema de doble apellido resulta peculiar, pero además cuenta con una diversidad relativamente baja. Hemos tomado como ejemplo los 20 apellidos más frecuentes de la población peninsular para realizar una primera aproximación a su distribución.

Los 20 apellidos más frecuentes de la España peninsular (Tabla 1), representan al 32,4 % de la población. Una cifra que puede tomarse como estimadora de la diversidad genética. Tomándolos como referencia, cabría esperar encontrar la mayor diversidad en Aragón, Cataluña y Navarra, donde los 20 apellidos más frecuentes representan entre el 22 y el 23 % de la población. La menor diversidad es esperable encontrarla en Asturias donde sólo 20 apellidos representan a más de la mitad de la población (53,5 %).

En cuanto a la composición de apellidos, la única comunidad que comparte exactamente los mismos 20 apellidos más frecuentes es Madrid (Tabla 1). Tras ella, las más similares al resto de la península parecerían Extremadura y Castilla la Mancha con sólo 2 apellidos diferentes y Castilla y León, Andalucía y el País Vasco que sólo difieren en 3 apellidos . Por el contrario, las comunidades autónomas más diferenciadas parecen Galicia, con hasta 8 apellidos diferentes y Asturias con 7.

Nacional		Galicia		Asturias		Cantabria		Castilla y León		País Vasco		La Rioja		Navarra		Aragón		Cataluña		Extremadura		Castilla La Mancha		Madrid		Valencia		Murcia		Andalucía	
Ap.	%	Ap.	%	Ap.	%	Ap.	%	Ap.	%	Ap.	%	Ap.	%	Ap.	%	Ap.	%	Ap.	%	Ap.	%	Ap.	%	Ap.	%	Ap.	%	Ap.	%
García	4,2%	Rodríguez	4,7%	Fernández	9,1%	Fernández	5,2%	García	6,0%	García	3,2%	Martínez	5,2%	García	2,7%	García	2,7%	García	3,0%	Sánchez	3,8%	García	5,6%	García	4,9%	García	4,0%	Martínez	6,7%	García	4,0%
Fernández	2,8%	Fernández	4,5%	García	7,4%	García	4,3%	González	4,2%	Fernández	2,5%	García	3,9%	Martínez	2,7%	Pérez	2,0%	Martínez	2,1%	García	3,7%	Sánchez	3,8%	Sánchez	3,0%	Martínez	3,7%	García	5,4%	Rodríguez	2,9%
González	2,5%	González	3,6%	González	5,8%	González	4,2%	Martín	3,4%	González	2,4%	Fernández	2,9%	Pérez	2,0%	Martínez	1,9%	López	1,9%	González	3,2%	López	3,6%	Fernández	2,8%	Pérez	2,1%	Sánchez	4,3%	López	2,7%
Rodríguez	2,5%	López	3,3%	Álvarez	4,8%	Gutiérrez	3,1%	Fernández	3,3%	López	1,9%	Pérez	2,8%	Jiménez	1,9%	López	1,7%	Sánchez	1,8%	Rodríguez	2,8%	Martínez	3,1%	González	2,7%	López	2,1%	López	4,1%	Sánchez	2,7%
López	2,5%	García	3,2%	Rodríguez	4,4%	Ruiz	2,6%	Rodríguez	3,1%	Martínez	1,9%	Jiménez	2,4%	López	1,7%	Sánchez	1,5%	Fernández	1,7%	Fernández	2,2%	Fernández	2,8%	López	2,6%	Sánchez	2,0%	Pérez	2,5%	Fernández	2,6%
Martínez	2,4%	Pérez	2,6%	Suárez	2,6%	Pérez	2,5%	Sánchez	2,5%	Ruiz	1,8%	López	2,0%	Fernández	1,7%	Jiménez	1,3%	Rodríguez	1,7%	Martín	2,1%	González	2,3%	Rodríguez	2,5%	González	1,3%	Fernández	2,1%	González	2,3%
Sánchez	2,3%	Martínez	2,3%	Martínez	2,5%	Pérez	2,6%	Pérez	2,5%	Rodríguez	1,8%	González	2,0%	Sánchez	1,2%	Martín	0,9%	González	1,6%	Pérez	1,7%	Gómez	2,2%	Martín	2,4%	Fernández	1,3%	Hernández	1,8%	Pérez	2,2%
Pérez	2,1%	Vázquez	2,0%	López	2,3%	Martínez	2,3%	Martínez	2,1%	Sánchez	1,4%	González	2,0%	Sánchez	1,2%	Martín	0,9%	Pérez	1,5%	Gómez	1,6%	Rodríguez	2,1%	Martínez	2,0%	Rodríguez	1,3%	González	1,5%	Jiménez	1,9%
Gómez	1,4%	Álvarez	1,6%	Pérez	2,3%	López	2,2%	López	2,0%	Martín	1,3%	Sáenz	1,9%	Rodríguez	1,1%	Gómez	0,9%	Jiménez	1,0%	López	1,4%	Martín	1,7%	Pérez	1,9%	Gómez	1,2%	Ruiz	1,4%	Ruiz	1,7%
Martín	1,3%	Gómez	1,4%	Díaz	2,1%	Rodríguez	2,0%	Hernández	1,8%	Gómez	1,2%	Rodríguez	1,4%	Ruiz	1,0%	González	0,9%	Gómez	0,9%	Díaz	1,3%	Pérez	1,7%	Gómez	1,7%	Navarro	1,1%	Jiménez	1,3%	Martínez	1,7%
Jiménez	1,2%	Castro	1,2%	Menéndez	2,1%	Sánchez	1,5%	Alonso	1,6%	Ruiz	1,1%	Gómez	1,2%	Gabi	0,9%	Hernández	0,9%	Ruiz	0,9%	Moreno	1,3%	Jiménez	1,6%	Jiménez	1,4%	Navarro	0,9%	Gómez	1,3%	Martín	1,5%
Ruiz	1,1%	Iglesias	1,1%	Alonso	1,6%	Díaz	1,4%	Álvarez	1,5%	Alonso	0,8%	Moreno	0,9%	Echevarría	0,7%	Fernández	0,8%	Moreno	0,8%	Hernández	1,1%	Díaz	1,5%	Díaz	1,2%	Ruiz	0,9%	Gómez	1,3%	Moreno	1,4%
Díaz	0,9%	Sánchez	1,0%	Sánchez	1,6%	Alonso	1,3%	Gómez	1,5%	Hernández	0,7%	Gil	0,8%	Hernández	0,7%	Gil	0,8%	Martín	0,7%	Jiménez	1,1%	Moreno	1,4%	Moreno	1,1%	Hernández	0,8%	Moreno	1,1%	Romero	1,1%
Moreno	0,9%	Díaz	1,0%	Iglesias	1,0%	Ortiz	0,9%	Jiménez	1,0%	Echevarría	0,7%	Hernández	0,8%	Moreno	0,6%	Rodríguez	0,7%	Muñoz	0,7%	Muñoz	0,9%	Ruiz	1,4%	Hernández	1,1%	Hernández	0,8%	Moreno	1,1%	Romero	1,1%
Álvarez	0,8%	Blanco	0,9%	Gutiérrez	0,9%	Martín	0,8%	Díez	1,0%	Jiménez	0,7%	Alonso	0,8%	Sanz	0,6%	Sanz	0,7%	Hernández	0,6%	Domínguez	0,9%	Muñoz	1,3%	Muñoz	1,1%	Muñoz	0,7%	Muñoz	1,0%	Muñoz	1,1%
Muñoz	0,8%	Alonso	0,9%	Blanco	0,8%	Sáiz	0,8%	Gutiérrez	0,9%	Álvarez	0,7%	Sánchez	0,7%	Gómez	0,6%	Navarro	0,7%	Díaz	0,6%	Martínez	0,9%	Romero	0,8%	Ruiz	1,0%	Gil	0,5%	Marín	1,0%	Díaz	1,0%
Hernández	0,8%	Varela	0,9%	Gómez	0,7%	Díez	0,7%	Blanco	0,8%	Gutiérrez	0,6%	Pascual	0,7%	Díaz	0,6%	Ruiz	0,7%	Torres	0,5%	Romero	0,8%	Serrano	0,8%	Alvarez	0,7%	Torres	0,5%	Díaz	0,7%	Domínguez	0,7%
Romero	0,6%	Otero	0,9%	Méndez	0,6%	Cobo	0,7%	Sanz	0,8%	Díaz	0,6%	Díez	0,6%	Gil	0,5%	Serrano	0,6%	Vidal	0,5%	Alvarez	0,7%	Navarro	0,7%	Alonso	0,7%	Díaz	0,5%	Molina	0,7%	Ramírez	0,6%
Alonso	0,6%	Domínguez	0,6%	Vega	0,6%	Alvarez	0,7%	Muñoz	0,6%	Aguirre	0,5%	Marín	0,6%	Martín	0,5%	Moreno	0,6%	Marí	0,5%	Ramos	0,6%	Hernández	0,7%	Gutiérrez	0,6%	Romero	0,5%	Nicolás	0,6%	Gutiérrez	0,6%
Gutiérrez	0,5%	Rey	0,7%	Vázquez	0,6%	Sainz	0,7%	Ruiz	0,6%	Bilbao	0,5%	Alvarez	0,6%	Iriarte	0,4%	Muñoz	0,5%	Navarro	0,5%	Ruiz	0,6%	Gutiérrez	0,5%	Romero	0,6%	Pastor	0,5%	Cánovas	0,5%	Torres	0,6%
32,4%		38,6%		53,5%		40,7%		41,1%		26,2%		34,3%		23,4%		22,1%		23,3%		32,9%		39,5%		36,0%		26,7%		40,3%		34,9%	

Tabla 1. Frecuencias de los 20 apellidos más repetidos en cada Comunidad autónoma. En gris los apellidos que se encuentran entre los 20 más frecuentes en toda la población española.

Otro aspecto interesante de la distribución de los apellidos es su origen. Se han catalogado los apellidos en diferentes tipos para observar su distribución. Exactamente lo que se ha hecho es tomar los 1000 apellidos más frecuentes en cada provincia y catalogarlos dentro de cuatro categorías en función de su origen:

- Apellidos procedentes de nombres personales.
- Apellidos procedentes de nombres de profesiones.
- Apellidos procedentes de apodos.
- Apellidos procedentes de lugares geográficos.

Parece que la distribución de los diferentes tipos de apellidos podría mostrar patrones regionales y reflejar particularidades culturales de cada zona (Darlu et al., 2012).

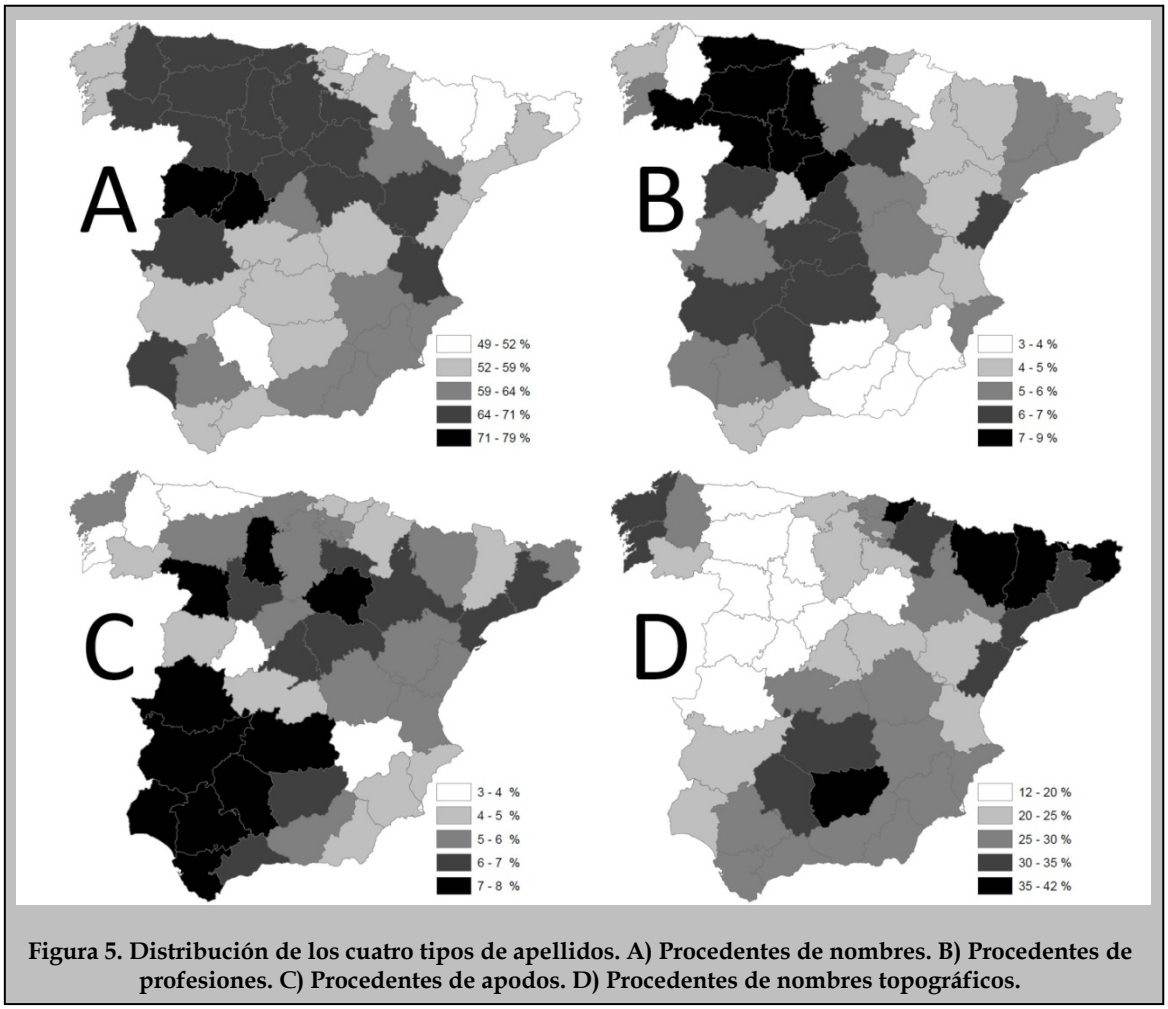


Figura 5. Distribución de los cuatro tipos de apellidos. A) Procedentes de nombres. B) Procedentes de profesiones. C) Procedentes de apodos. D) Procedentes de nombres topográficos.

El tipo de apellidos más numeroso son los que tienen su origen en nombres propios (representan en promedio el 61,98 % de los apellidos) a diferencia de lo que sucede en Alemania donde los más frecuentes son los apellidos procedentes de profesiones o de Holanda donde lo son los procedentes de topónimos (Bloothooft et al, 2014). En cualquier caso, parece que

lo más frecuente en Europa, con la excepción de Alemania es que los apellidos más frecuentes sean los procedentes de nombres o de topónimos. Este tipo de apellidos (Figura 4, A) es mucho más numeroso en el noroeste peninsular, con picos que llegan al 79 % en Ávila y el 77% en Salamanca.

El segundo tipo más numeroso son los procedentes de nombres de lugares (representan como promedio el 26,76 %). Son más importantes en toda la franja costera mediterránea y el noreste, con frecuencias del 42 % en Guipúzcoa y del 39 % en Huesca, Gerona y Lleida.

Notablemente menos importante son los otros dos tipos de apellidos. Los apellidos procedentes de nombres de profesiones representan en promedio el 5,68 % y son más frecuentes en la franja Oeste, especialmente en el noroeste (Figura 4, B) con frecuencias de hasta el 9 % en Zamora o en León. Por último, los menos frecuentes son los apellidos procedentes de apodos (son el 5,58 %). Son más frecuentes en el Suroeste peninsular (Figura 4, C).

Bibliografía

Blanco-Villegas MJ, Boattini A, Otero HR, Pettener D (2004) Inbreeding patterns in La Cabrera, Spain: Dispensations, multiple consanguinity analysis, and isonymy. Hum Biol 76: 191–210.

Bloothooft G, Brouwer L, Chareille P, Darlu P, Degioanni A, Dräger K, Germain J, Lisa A, Rodríguez R (2014) European Surname Typology Project. XXV International Congress of Onomastic Sciences.

Colantonio SE, Lasker GW, Kaplan BA, Fuster V (2003) Use of surname models in human population biology: a review of recent developments. Hum Biol 75:785-807.

Darlu P, Bloothooft G, Boattini A, Brouwer L, Brouwer M, Brunet G, Chareille P, Cheshire J, Coates R, Dräger K, Desjardins B, Hanks P, Longley P, Mandemakers K, Mateos P, Pettener D, Useli A, Manni F (2012) The family name as socio-cultural feature and genetic metaphor: from concepts to methods. Hum Biol. 84: 169-214.

Dipierri J, Rodríguez-Larralde A, Alfaro E, Scapoli C, Mamolini E, et al. (2011) A Study of the Population of Paraguay through Isonymy. Ann Hum Genet 75: 678-687.

Dipierri J, Rodríguez-Larralde A, Alfaro E, Scapoli C, Mamolini E, et al. (2011) A Study of the Population of Paraguay through Isonymy. Ann Hum Genet 75: 678-687.

Faure R, Ribes MA, García A (2001) Diccionario de apellidos españoles. Madrid: Espasa-Calpe.

Fuster V, Colantonio SE. 2002. Consanguinity in Spain: socioeconomic, demographic, and geographic influences. Hum.Biol 74: 301-315.

Mateos P, Tucker DK (2008) Forenames and Surnames in Spain in 2004. Names: A Journal of Onomastics, 56(3), 165–184.

Pettener D, Pastor S, Tarazona-Santos E (1998) Surnames and genetic structure of a high-altitude quechua community from the Ichu river valley peruvian central Andes 1825-1914. Hum Biol 70: 865-87.

Solís JA (2002) El gran libro de los apellidos. La Coruña: El arca de papel.

1. ESTRUCTURA DE LA POBLACIÓN ESPAÑOLA

ESTRUCTURA DE LA POBLACIÓN ESPAÑOLA

Resumen

Introducción: En la población española tenemos presentes una serie de características (como el alto grado de aislamiento, la diversidad lingüística, la complejidad orográfica,...) que unidas al hecho de que el sistema de apellidos español (tanto el sistema del doble apellido como los apellidos españoles en sí mismos) está ampliamente distribuido por el mundo, la convierten en un sujeto de especial interés para los estudios poblacionales.

El objetivo principal de este capítulo es determinar la estructura de la población española a través del uso de los apellidos. La presencia de aproximaciones anteriores y la disponibilidad de estudios genéticos y lingüísticos permiten también evaluar las técnicas empleadas.

Material y métodos: La base de datos de partida es el padrón continuo de 2008, que contiene casi 57 millones de datos (87.000 apellidos diferentes).

Los mapas autoorganizados de Kohonen (SOM) nos permiten identificar el origen geográfico de los apellidos y separar aquellos apellidos sin origen claro. Usando esta información para estimar el parentesco entre las provincias españolas de la manera más fiable posible. Este parentesco se analiza mediante la aplicación del algoritmo de Monmonier.

Resultados: Los resultados obtenidos muestran que la española es una población con una estructura muy conservada determinada, principalmente, por factores de tipo geográfico. Pese a la diversidad existente, los factores de tipo lingüístico y étnico, parecen haber jugado un papel secundario.

La influencia continuada de estos factores ha dado lugar a una población dividida en dos grupos: el arco Cántabro-atlántico y el arco Mediterráneo.

Introducción

El origen de los apellidos en Europa se remonta a la edad media. En el caso de España la aparición de los primeros apellidos es anterior al siglo IX y eran utilizados únicamente por las clases dominantes. La adopción del sistema hereditario de apellidos no se extendería entre todas las clases sociales hasta comienzos del siglo XIII. A partir del siglo XV el sistema de apellidos queda totalmente consolidado, en parte gracias a la obligación de registrar todos los nacimientos y defunciones (Faure et al., 2001; Mateos y Tucker, 2008). Además de su temprana aparición, el sistema español de apellidos tiene la particularidad de usar dos apellidos; el primero heredado de la madre, el segundo del padre (Pettener et al., 1998; Colantonio et al., 2003; Blanco-Villegas et al., 2004; Mateos y Tucker, 2008; Dipierri et al., 2011).

Recientemente, la disponibilidad de las grandes bases de datos y de potentes recursos informáticos se han unido para permitir el análisis grupos poblacionales más grandes, apareciendo análisis a nivel nacional (Rodríguez-Larralde et al., 1998; Barrai et al., 2000; Rodríguez-Larralde et al., 2000; Rodríguez-Larralde et al., 2003; Barrai et al., 2004; Dipierri et al., 2005; Manni et al., 2005; Dipierri et al., 2011; Rodríguez-Larralde et al., 2011) e incluso a nivel continental (Scapoli et al., 2007; Cheshire et al., 2011). En estos estudios, normalmente se espera que el uso de bases de datos tan grandes minimice las posibles desviaciones que pueden desviarse del uso de los apellidos como estimadores (Barrai et al., 2000; Rodríguez-Larralde et al., 2003). Pero muchas de las conclusiones obtenidas en este tipo de estudios parten de la asunción de que los apellidos son monofiléticos (Rodríguez-Larralde et al., 2003; Manni et al., 2005). Dicha asunción resulta uno de los puntos más problemáticos del uso de los apellidos desde que las primeras dudas sobre el carácter monofilético de los apellidos fueron expuestas (Rogers, 1991).

Actualmente este aspecto del estudio de los apellidos sigue siendo un tema de controversia. De cualquier manera estudios recientes demuestran que mas allá

de la controversia y del carácter monofilético o polifilético de los apellidos, la diversidad de linajes Y dentro de un apellido resulta mucho menor que la existente dentro de un muestreo al azar (Sykes e Irven, 2000) y que apellidos con una distribución geográfica marcada corresponden a uno o unos pocos linajes paternos y dichos linajes comparten origen con el apellido en cuestión. Por lo tanto, si somos capaces de analizar la distribución geográfica de un apellido hasta el punto de identificar su origen geográfico, podemos tener una certeza razonable de que sus portadores comparten un origen geográfico y un mismo pool genético (Manni et al., 2005; Balanovskaia et al., 2011). Esta característica de los apellidos hace de ellos, más que unos estimadores genéticos, unos marcadores geográficos extremadamente útiles para el estudio y comparación de las poblaciones humanas.

Convencionalmente, las bases de datos que se han utilizado para este tipo de trabajos han sido listines telefónicos, como algunos autores han puesto de manifiesto (Rodríguez-Larralde et al., 2003; Cheshire et al., 2011), estas bases de datos suponen un recurso magnífico para el estudio de poblaciones, pero soportan algunos sesgos.

Actualmente, existen posibilidades que nos permiten dar un paso más en el análisis en el análisis de los apellidos y utilizarlos como marcadores geográficos seguros. Por un lado recientes trabajos han analizado la distribución geográfica de los apellidos (Manni et al., 2005), aplicando las redes neuronales (Kohonen, 1982 y 1984) a las bases de datos nacionales se consigue identificar y separar aquellos apellidos que muestran una clara distribución geográfica con un origen distinguible, de los que muestran una distribución geográfica demasiado amplia y carente de un patrón que nos permita identificar su origen geográfico (Manni et al., 2005; Boattini et al., 2011).

Por otro lado, en la actualidad, se puede acceder a bases de datos censales. Estas bases de datos no han sido explotadas hasta la fecha y carecen de los sesgos

presentes en los registros telefónicos o de votantes (Rodríguez-Larralde et al., 2003; Cheshire et al., 2011).

El presente capítulo toma estos dos puntos (el uso de las redes neuronales y de las bases de datos censales) para abordar el estudio de la población española. Una población que resulta de especial interés por diferentes razones. Desde un punto de vista estrictamente técnico la disponibilidad de los datos censales ha sido crucial. Por otro lado, para la población española, existen en la literatura científica algunos estudios de apellidos tanto a nivel nacional (Rodríguez-Larralde et al., 2003) como a nivel europeo (Scapoli et al., 2007; Cheshire et al., 2011) que nos pueden servir como comparativa. A estos, más recientemente se les ha sumado un estudio genético de la población española (Adams et al., 2008). Todos estos estudios previos suponen un magnifico punto de partida para analizar la utilidad de las metodologías a aplicar sobre una base de datos inédita y conseguir una aproximación lo más profunda y fiable posible a la estructura de la población española.

Más allá del interés metodológico, los procesos microevolutivos y los factores que los han condicionado resultan uno de los aspectos más interesantes de este tipo de estudios. Un amplio abanico de condicionantes se han identificado en diferentes estudios, pero podemos considerarlos principalmente de tipo lingüístico (Barrai et al., 2002, 2004; Manni et al., 2008), étnico (Fiorini et al., 2007), socio-económico (Caravello y Tasso 1999, 2002) y geográficos (Esparza et al., 2006; Boattini et al., 2006, 2007). Todos estos factores se encuentran muy presentes en la población española. Lingüísticamente cuatro idiomas oficiales coexisten en el territorio español, incluso solapándose en algunas regiones (Garcia, 2007). En el caso particular de la población española existe además un estudio dialectométrico (Goebl, 2010) que nos permitirá analizar con detalle la influencia de los factores lingüísticos. En términos étnicos, el País Vasco es uno de las más antiguas y distintivas poblaciones en toda Europa (Bertranpetit et al., 1995; Belle et al., 2006). En lo que a los factores geográficos respecta, la península ibérica es una de las zonas más complejas del continente, con una

gran altitud media y un alto grado de aislamiento. Por último, a lo largo de la historia, el territorio español ha sido sujeto de sucesivas ocupaciones, disputas internas y fuertes cambios en la organización administrativa interna (Martínez, 2003). Por todo ello, la población española parece una magnífica población para un análisis de estas características.

Material y métodos

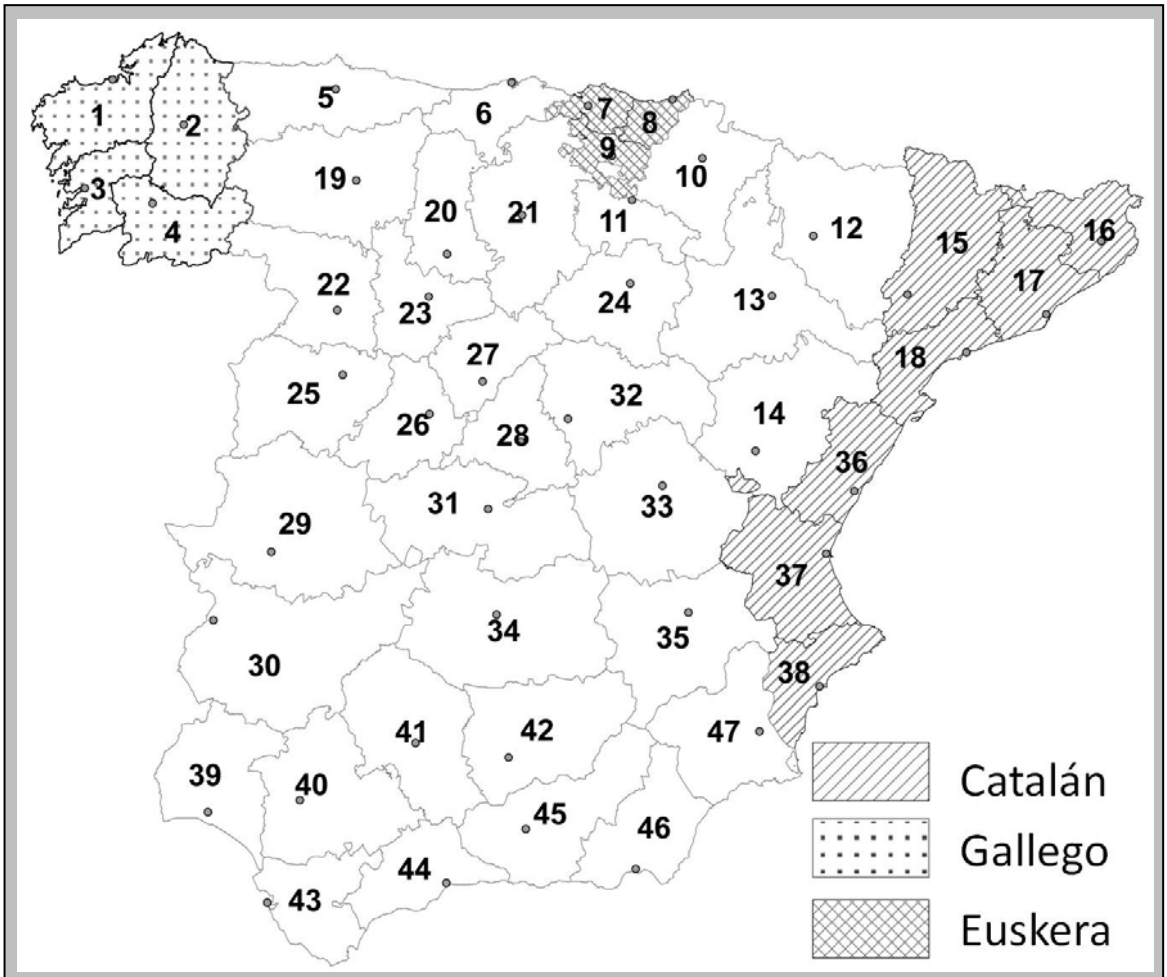


Figura 1. Mapa provincial y extensión geográfica de las lenguas cooficiales españolas.

Zona de estudio

España es un país situado en la península ibérica en el extremo sur-oeste del continente europeo, tiene una superficie de 504.645 Km2 y está rodeado por el mar por Norte, Sur, Noroeste y Este. Por tierra tiene fronteras con Portugal por el Oeste y Francia por el Noreste.

Geográficamente, hablamos de un territorio situado a bastante altitud sobre el nivel del mar, con un promedio de 660 m y de marcado carácter montañoso en comparación con el promedio del resto de países europeos.

Tiene una población de 47 millones de habitantes desigualmente repartidos, que se concentran principalmente en los entornos costeros, dejando el interior

del país con una baja densidad poblacional (excepción hecha de Madrid, capital administrativa).

Administrativamente, la España peninsular está organizada en 15 comunidades autónomas y 47 Provincias. El idioma oficial de España es el español o castellano, pero comparte espacio con otras lenguas cooficiales en algunos territorios (Figura 1).

Isonimia

Desde el trabajo pionero de Crow y Mange (Crow y Mange, 1965) el análisis de los apellidos (isonimia) ha tenido una importante aportación al estudio de procesos microevolutivos en las poblaciones humanas. En concreto, en este capítulo, nos centraremos en el uso de estas técnicas para el estudio de las relaciones entre poblaciones.

Genéticamente hablando, se puede considerar que dos poblaciones diferentes serán tanto más próximas cuanto mayor sea la probabilidad de que dos alelos para un mismo locus en dichas poblaciones sean idénticos por descendencia (Morton et al., 1971). El proceso de herencia del apellido emula al proceso de herencia genética del cromosoma Y (Chakraborty y Schwartz, 1990). Pero más allá de si podemos o no considerar que un apellido está ligado sólo a un linaje Y, podemos estar seguros de que un apellido con un origen geográfico delimitado, estará caracterizando a un acervo genético propio de esa área geográfica y que lo distingue de otros grupos (Sykes e Irven, 2000), lo que nos permite utilizar a los apellidos como marcadores geográficos o poblacionales (Balanovskaia et al., 2011). Estos son los principios básicos del método de la isonimia. En otras palabras, podemos considerar que dos poblaciones estarán tanto más emparentadas cuanto mayor es el número de apellidos en común y cuanto más parecida es su distribución (Lasker y Kaplan, 1985).

Base de datos

La base de datos de apellidos, se ha obtenido del instituto nacional de estadística, e incluye los datos correspondientes al padrón continuo de población del año 2008 en las 47 provincias peninsulares. Los datos son públicos, sin embargo no están accesibles y la información debe solicitarse al INE mediante el procedimiento establecido. Un resumen de esta información está disponible en la página del INE (www.ine.es). La base de datos incluye los apellidos de toda la población censada en cada municipio en 2008, siempre y cuando aparezcan más de 5 veces en un mismo municipio. Esto último excluye a los apellidos menos frecuentes entre los que está más representada por ejemplo la migración reciente.

	Nº
Población española en el año 2008	*45.593.385*
Datos totales en la base de datos	*56.976.706*
Datos totales en la base de datos corregida	*51.419.788*
Datos totales con origen identificado	*16.687.001*
Formas diferentes de apellidos en la base de datos	*87.148*
Formas diferentes de apellidos en la base de datos corregida	*33.753*
Formas de apellidos con origen identificado	*31.433*

Tabla 1. Explicación de la base de datos y su procesado.

La población registrada en España (Tabla 1) en el padrón 2008 fue de 45.593.385 habitantes. El sistema español de apellidos utiliza dos apellidos, el primero heredado del padre y el segundo de la madre. Dada la disponibilidad de los dos apellidos de cada individuo, la base de datos se construyó utilizando tanto el primero como el segundo apellido algo que según diversos autores, multiplica la cantidad de información y contribuye a la robustez del análisis (Pettener et al., 1998; Colantonio et al., 2003; Blanco-Villegas et al., 2004; Dipierri et al., 2011; Mateos y Tucker, 2008; Dipierri et al., 2011).

Sin embargo los registros que contiene la base de datos no son exactamente el doble del número de habitantes. La base de datos del INE sólo registra aquellos apellidos que aparecen más de cinco veces en un mismo municipio. Por otro

lado existe una parte de la población que no porta los dos apellidos (por ejemplo la inmigración reciente de países con otro sistema de apellidos, o las adopciones o los hijos ilegítimos que en ocasiones tampoco llevan los dos apellidos...). En definitiva hemos trabajado con una base de datos de 56.976.706 registros que representan a 87.148 apellidos diferentes (Tabla 1).

Corrección de la base de datos

La base de datos inicial fue minuciosamente revisada. Se encontraron apellidos repetidos, diferentes grafías del mismo apellido, espacios entre las palabras, errores ortográficos y formas compuestas. Todas estas incidencias representan un problema a la hora del procesado estadístico. Para evitar estos problemas, todos los apellidos fueron revisados manualmente y con apoyo bibliográfico (Faure et al., 2001; Solís, 2002) y cartográfico; y corregidos cuando fue necesario.

En el siguiente paso del depurado se seleccionaron sólo los datos referidos a la población peninsular. Las islas y los territorios extra-peninsulares tienen una relación particular con el resto del territorio español. Para evitar los efectos estadísticos de estas particularidades se ha trabajado sólo con la población peninsular.

Para evitar el exceso de ruido en los procedimientos estadísticos al identificar el origen geográfico de cada grupo de apellidos, se han eliminado los datos correspondientes a los apellidos que aparecen menos de 20 veces en la base de datos.

La base de datos final, después del depurado, contiene 51.419.788 datos, correspondientes a 33.753 apellidos diferentes que se distribuyen por las 47 provincias peninsulares del territorio español (Tabla 1).

Procesado de los datos

Terminado el depurado se ha construido una matriz de doble entrada, en la que las filas (i) corresponden a cada apellido y las columnas (j) a cada provincia. De

esta forma cada casilla (ij) corresponde a la frecuencia que representa cada apellido en la población total de cada provincia.

Posteriormente se ha realizado una transformación de las frecuencias en dos pasos (Boattini et al., 2011):

- En un primer paso, se pretende evitar que las poblaciones más pequeñas tengan un peso excesivo, aplicando la siguiente expresión:

$$f_i = \frac{fabs_{ij}}{\log(pop_j)}$$

  Siendo $fabs_{ij}$ la frecuencia absoluta del apellido "i" en la provincia "j" y pop_j es la población total de la provincia "j"

- En un segundo paso, se pretende evitar que los apellidos se agrupen en función de lo numerosos que sean, empleando a la expresión:

$$wf_i = \frac{f_i}{\Sigma f_i}$$

  Donde f_i es el resultado de la expresión anterior.

Agrupamiento de los apellidos

Los apellidos se han agrupado en función de su distribución geográfica utilizando un procedimiento estadístico tipo Cluster de minería de datos, los mapas autoorganizados de Kohonen o SOM (Kohonen, 1982; 1984).

Los SOM son redes neuronales de aprendizaje no supervisado que permiten obtener un reconocimiento estadístico de patrones. En este caso se han utilizado para el reconocimiento de patrones en la distribución geográfica de los apellidos. Este es un procedimiento de reciente aplicación en el campo de la Biodemografía y permite agrupar los apellidos en función de su distribución y analizar esta para identificar sus orígenes (Manni et al., 2005).

En nuestro caso el software utilizado ha sido el paquete de R-project "Kohonen" (Wehrens y Buydens, 2007). Utilizando este software hemos clasificado los apellidos en una matriz rectangular. El tamaño de dicha matriz debe establecerse. El criterio a seguir es elegir un tamaño que no sea tan grande que imposibilite su interpretación ni tan pequeño que los resultados no resulten representativos. Para conseguir esto, el criterio adoptado ha sido utilizar el tamaño de matriz más pequeño en el aparezca una celda vacía (Boattini et al., 2011) que en nuestro caso es de 12 celdas de ancho por 12 de alto, con 1.000 repeticiones, de esta forma utilizamos el tamaño más pequeño para el que todos los grupos ya son representativos (empiezan a aparecer celdas vacías).

En definitiva, el SOM consta de una capa de entrada de 33.753 vectores (un vector por cada apellido) y una capa de salida de 144 celdas (grupos de apellidos con similar distribución geográfica).

Finalmente, la distribución geográfica de cada grupo de apellidos se ha representado gráficamente mediante mapas de gradientes utilizando el software Arcgis 10.0. La distribución de cada apellido muestra un lugar en el que la concentración de sujetos que lo portan es mayor, ese es el lugar de origen del apellido y por lo tanto del grupo poblacional (Manni et al., 2005; Boattini et al., 2011) al que caracteriza, y que compartirá por un acervo genético común (Sykes e Irven, 2000; Balanovskaia et al., 2011). Analizando esta distribución geográfica, podemos establecer tres grupos de apellidos en base a su origen:

- Apellidos que muestran un claro pico de mayor concentración claro. Estos apellidos pueden considerarse monofiléticos en el sentido de que caracterizan a un grupo poblacional concreto con un mismo pool genético.
- Apellidos que muestran varios picos de mayor concentración en ubicaciones diferente. Estos apellidos no pueden asociarse a una única población y por lo tanto no se pueden considerar marcadores geográficos seguros.

- Por último, apellidos que no muestran una distribución clara y que no puede distinguirse un origen. Tampoco pueden asociarse a una población y tampoco pueden considerarse marcadores seguros.

Sólo el primero de los grupos se considera que puede caracterizar a la población a la que va asociado y que por lo tanto contiene información fiable (Manni et al., 2005).

Distancias genéticas

Completada la fase anterior, se han empleado solo los apellidos cuyos orígenes se han podido identificar para calcular el parentesco entre poblaciones. Para ello se han calculado las relaciones genéticas entre las poblaciones con el paquete de R-project "Biodem" (Boattini et al., 2006). Concretamente, se han calculado:

- Índice de Lasker (Lasker, 1977), dado por la expresión:

$$R = \frac{(S_{i1}\, S_{i2})}{(\,2\, n_{\,1}\, n_2)}$$

  Donde S_{i1} y S_{i2} sin las frecuencias absolutas del apellido "i" en las poblaciones 1 y 2 respectivamente y $n_{\,1}$ y n_2 los tamaños de las poblaciones 1 y 2.

- Índice de Relethford (Morton, 1973; Relethford, 1988), dado por la expresión:

$$d^2 = I_{ii} + I_{jj} - 2I_{ij}$$

  Donde I_{ii} es el grado de isonimia al azar dentro de la población "i", I_{jj} es el grado de isonimia al azar dentro de la población "j", e I_{ij} es el grado de isonimia al azar entre las poblaciones "i" y "j".

- Índice de Hedrick (Hedrick, 1971; Weiss, 1980) que viene dado por la expresión:

$$H_{ij} = \frac{\Sigma_k(p_{ki}p_{kj})}{1/2\Sigma_k(p_{ki}^2 + p_{kj}^2)}$$

Donde p_{ki} y p_{kj} son las frecuencias del apellido o la procedencia "k" en las poblaciones "i" y "j" respectivamente.

El resultado de este procedimiento es una matriz cuadrada por cada coeficiente, con un número de filas y columnas igual al número de provincias incluidas en el estudio (47x47). En esta matriz, cada casilla indica la similitud o la distancia genética entre las poblaciones de la fila y columna correspondientes.

Con el objeto de conocer el índice más adecuado para el caso particular que nos ocupa, se han estudiado las diferencias entre los coeficientes mediante un test de Mantel (Mantel, 1967) a tres realizado con el software con el paquete de R-Project "ade4" (Thioulouse et al., 1997). El resultado ha sido estadísticamente significativo (p-value= 0,001), las tres matrices de relaciones genéticas ofrecen resultados equiparables.

Se han realizado análisis exploratorios con los tres coeficientes de parentesco y, los resultados son similares. Por todo esto, se ha optado por emplear el coeficiente cuyos resultados eran más claros, que en el caso que nos ocupa es el coeficiente de Hedrick.

Representación de la estructura poblacional

La matriz de parentesco de Hedrick se ha estudiado en busca de discontinuidades y agrupaciones que muestren la estructura de la población española. Para empezar, el parentesco se ha representado mediante un MDS y, para un análisis más claro de las posibles aglomeraciones se ha realizado un

gráfico de densidades de Kernel con el paquete "MASS" (Venables y Ripley, 2002) de R-Project.

Para el cálculo de las barreras genéticas se necesita una matriz de distancias, pero el índice de Hedrick es un coeficiente de parentesco. Por esta razón la matriz de Hedrick se ha transformado en una matriz de distancias euclídeas mediante una transformación estándar (Mardia, 1979), según la expresión:

$$d_{ij} = \sqrt{(c_{ii} - 2c_{ij} + c_{jj})}$$

Una vez obtenida esta matriz de distancias se han identificado las barreras genéticas empleando el algoritmo de Monmonier (Monmonier, 1973). Esta técnica permite visualizar los datos contenidos en una matriz de distancias sobre un mapa geográfico identificando barreras genéticas (Barbujani et al., 1996; Manni et al., 2001; Palmé et al., 2003; Manni et al., 2004). Para su aplicación se ha utilizado el paquete de R-Project "adegenet" (Jombart, 2008). Este procedimiento parte de las coordenadas geográficas y compara las distancias geográficas con las distancias genéticas, identifica los pares en los que el desfase entre ambas distancias es mayor y ahí, sitúa las barreras genéticas (Manni et al., 2004).

En nuestro caso, los datos de partida son las coordenadas de las 47 capitales de provincia españolas. El hecho de seleccionar las capitales de provincia como puntos de referencia para calcular las distancias responde a un doble criterio. Por un lado, desde un punto de vista histórico demográfico, las capitales de provincia son los centros administrativos históricos de cada provincia. Por otro lado, geográficamente, el cálculo exploratorio de las áreas de influencia de las capitales de provincia, ha mostrado un mapa extremadamente similar al mapa de provincias.

Partiendo de estos datos, el programa detecta el mayor desfase entre distancias genéticas y distancias geográficas y por ahí se empieza a dibujar la barrera. El

trazado de las barreras se ha iniciado dos veces en dos puntos diferentes para obtener dos barreras (Manni et al., 2004). Además de este procedimiento ordinario, el software "adegenet", permite realizar un proceso optimizado. De esta forma, se puede identificar la barrera que tiene mayor relevancia en toda su extensión y no sólo en su punto inicial. En este proceso el software realiza un número dado de repeticiones (en nuestro caso han sido 100). En cada una de ellas inicia el trazado por un punto diferente y de todos los trazados selecciona el más significativo. De esta forma se identifica la barrera más importante en todo su trazado con independencia del punto por el que empieza.

Para estudiar las diferencias que puedan existir en la composición de apellidos entre los grupos resultantes, se ha realizado un análisis de muestras apareadas. Dado que los apellidos no se ajustan a los parámetros de normalidad se ha optado por el contraste no paramétrico de Wilcoxon (Wilcoxon, 1945) y se ha realizado utilizando el paquete "MASS" de R-Project (Venables y Ripley, 2002).

Comparación entre distancias

Para calcular las distancias entre las poblaciones se han utilizado las coordenadas de las capitales de provincia y se ha estimado la distancia en línea recta entre ellas usando el paquete "maptools" de R-Project (Nicholas et al., 2012).

Una vez calculadas estas distancias, se ha estudiado la correlación entre ambas distancias mediante un test de Mantel usando el paquete "ade4" de R-Project (Thioulouse et al., 1997). Además del test de Mantel, hemos realizado una regresión lineal usando el paquete "stats" de R-Project (R Development Core Team, 2011) para obtener unos resultados comparables a los disponibles previamente en la literatura (Rodríguez-Larralde et al., 2003).

Por último se ha realizado también un test de Mantel para comparar las distancias obtenidas mediante la isonimia con las disponibles en la literatura calculadas usando haplotipos Y-STR (Adams et al., 2008).

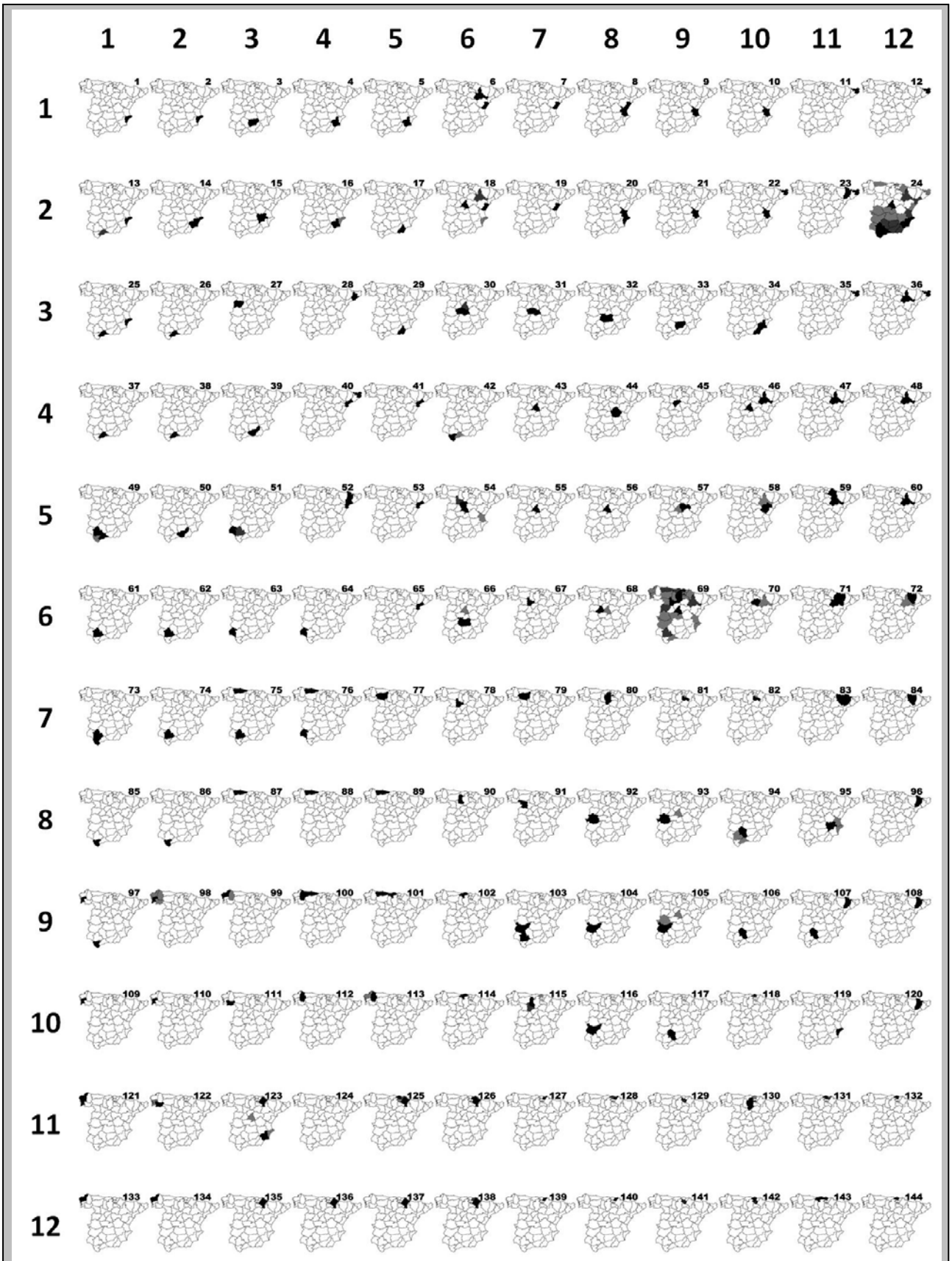


Figura 2. Matriz resultante de realizar los SOMs. Cada celda es un mapa de gradiente del territorio peninsular español que representa la distribución geográfica del correspondiente grupo de apellidos.

Resultados

SOM

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	
	379	347	249	328	340	128	370	96	561	268	366	792	Nº de apellidos
1	109.913	328.049	226.727	225.552	93.148	47.302	96.345	70.696	129.934	557.946	81.971	87.851	Nº de datos
	38	38	42	47	47	36	36	36	37	37	16	16	Origen
	330	130	111	403	250	367	267	144	333	44	103	903	Nº de apellidos
2	51.482	118.756	26.249	604.140	152.672	102.035	128.261	313.877	242.288	73.291	28.804	29.331.181	Nº de datos
	38	38	35	47	46	??	36	37	37	??	15	??	Origen
	188	251	182	129	146	227	228	164	186	123	204	58	Nº de apellidos
3	21.933	455.480	33.015	10.691	22.587	92.956	38.158	27.253	26.493	67.768	133.137	7.433	Nº de datos
	??	44	25	17	46	31	31	34	42	47	16	16	Origen
	273	246	179	124	326	236	283	158	83	167	349	802	Nº de apellidos
4	29.890	84.016	29.022	64.576	87.355	381.156	23.088	32.552	9.750	20.590	408.258	111.898	Nº de datos
	44	44	45	18	18	43	28	33	27	??	13	13	Origen
	80	199	101	109	422	178	197	338	160	167	84	566	Nº de apellidos
5	69.273	142.155	36.528	48.741	60.216	102.726	12.266	100.614	21.775	31.273	27.163	148.503	Nº de datos
	44	45	39	15	18	27	28	28	32	14	13	13	Origen
	478	320	163	116	230	218	203	92	525	76	241	290	Nº de apellidos
6	56.573	68.054	22.763	60.473	144.560	142.997	94.418	19.478	8.632.184	13.787	83.500	79.596	Nº de datos
	40	40	39	39	18	34	23	26	??	24	13	12	Origen
	110	274	25	15	206	147	164	110	223	162	92	189	Nº de apellidos
7	49.251	556.362	14.520	14.190	131.876	19.111	35.566	12.916	102.909	31.141	20.294	17.238	Nº de datos
	40	40	??	??	19	23	19	21	11	11	12	12	Origen
	447	259	237	423	150	63	196	164	223	194	139	254	Nº de apellidos
8	59.822	67.081	162.893	118.263	375.211	6.980	46.725	18.160	100.187	563.756	197.834	46.221	Nº de datos
	43	43	5	5	5	20	22	29	29	41	35	15	Origen
	17	230	232	40	48	265	106	233	213	174	8	474	Nº de apellidos
9	2.858	1.235.934	562.171	11.582	37.836	45.963	140.030	22.481	225.085	16.408	567	42.414	Nº de datos
	??	3	1	2	5	6	30	30	30	41	41	15	Origen
	460	343	145	226	214	191	151	267	220	195	328	199	Nº de apellidos
10	150.192	214.038	17.041	29.594	102.994	124.801	55.807	72.341	71.307	116.422	523.706	67.488	Nº de datos
	3	3	4	2	2	6	20	30	41	7	38	15	Origen
	184	158	69	0	88	119	344	170	150	210	135	401	Nº de apellidos
11	371.366	101.690	26.806	0	142.773	45.253	257.614	147.591	33.320	90.025	81.247	137.154	Nº de datos
	1	4	??	??	10	8	8	8	9	21	7	7	Origen
	596	366	214	485	446	162	408	138	129	135	83	644	Nº de apellidos
12	186.413	296.365	78.682	53.841	154.402	73.258	79.776	52.789	10.962	43.097	43.915	93.542	Nº de datos
	1	1	10	10	10	10	8	8	9	7	6	7	Origen

Tabla 2. Tabla resumen del SOM. Muestra el número de apellidos y el número de datos agrupados en cada celda; y el origen de cada grupo de apellidos.

Como paso previo indispensable al estudio de la estructura genética de la población española, se han estudiado los apellidos para optimizar su utilidad como estimadores de parentesco.

Los mapas autoorganizados de Kohonen (SOM) han permitido agrupar los apellidos en función de su distribución geográfica (Tabla 2, Figura 2). Una vez agrupados, en la Figura 2 se muestra la distribución de cada grupo. El resultado de este proceso, ha permitido identificar el origen geográfico de cada grupo de apellidos (Tabla 2, Figura 2). En total se ha identificado el origen de 31.433 de los 33.753 apellidos con los que se ha trabajado (16.687.001 datos). Se ha continuado el resto del trabajo con estos apellidos con el fin de evitar sesgos.

Aislamiento por distancia

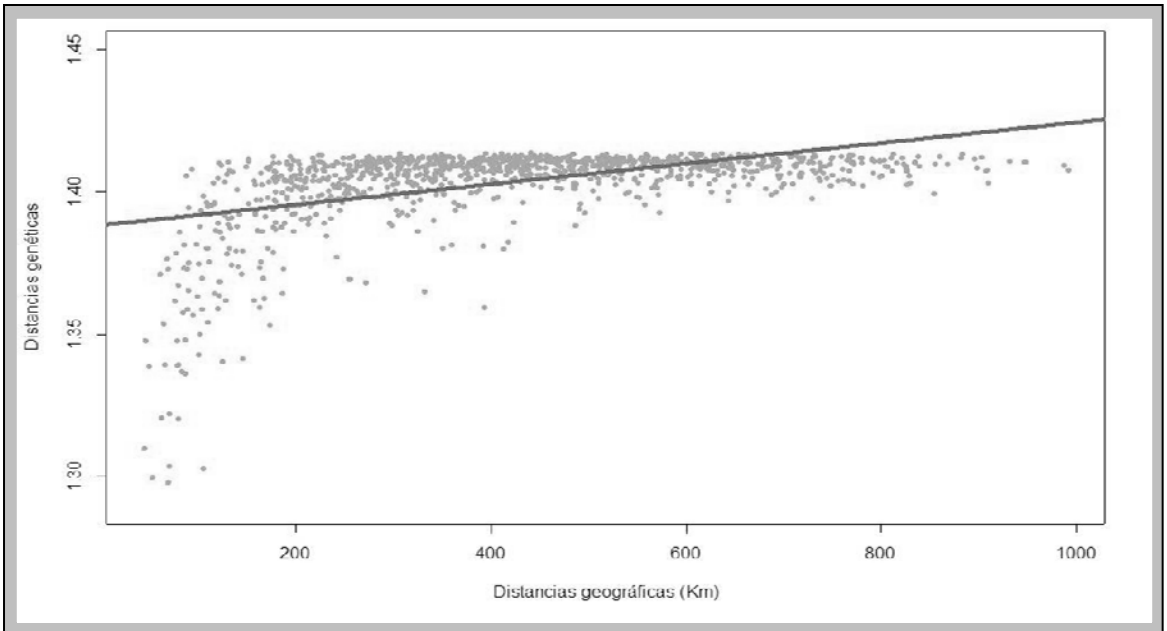


Figura 3. Relación entre distancias de Hedrick y distancias geográficas.

Utilizando los apellidos cuyo origen se ha conseguido identificar en la fase anterior, se ha realizado una primera aproximación a la estructura de la población española. Para ello se han calculado las relaciones de parentesco que hay entre las provincias españolas (Coeficiente de Hedrick). Una vez conocidas, se ha estudiado su relación con las distancias geográficas (Figura 3).

La matriz de distancias de Hedrick y la matriz correspondiente de distancias geográficas muestran una correlación significativa (test de Mantel, p-valor<0,01; regresión lineal, p-value <0,01; r = 0,491). La distancia de Hedrick entre las poblaciones aumenta muy rápido hasta un radio de 200 kilómetros, a partir de ahí aumenta más despacio para el resto de distancias (Figura 3).

Estructura de la población

Las distancias de Hedrick entre las 47 provincias peninsulares españolas, se han representado (Figura 4) en un MultiDimensional Scaling (MDS). En esta representación se observa cierta ordenación. Las provincias del oeste se han situado en la derecha del MDS (Figura 4, A), las provincias del norte español en

la zona superior (Figura 4, B), las del sur abajo y las del este en la parte infrior derecha (Figura 4, C) .

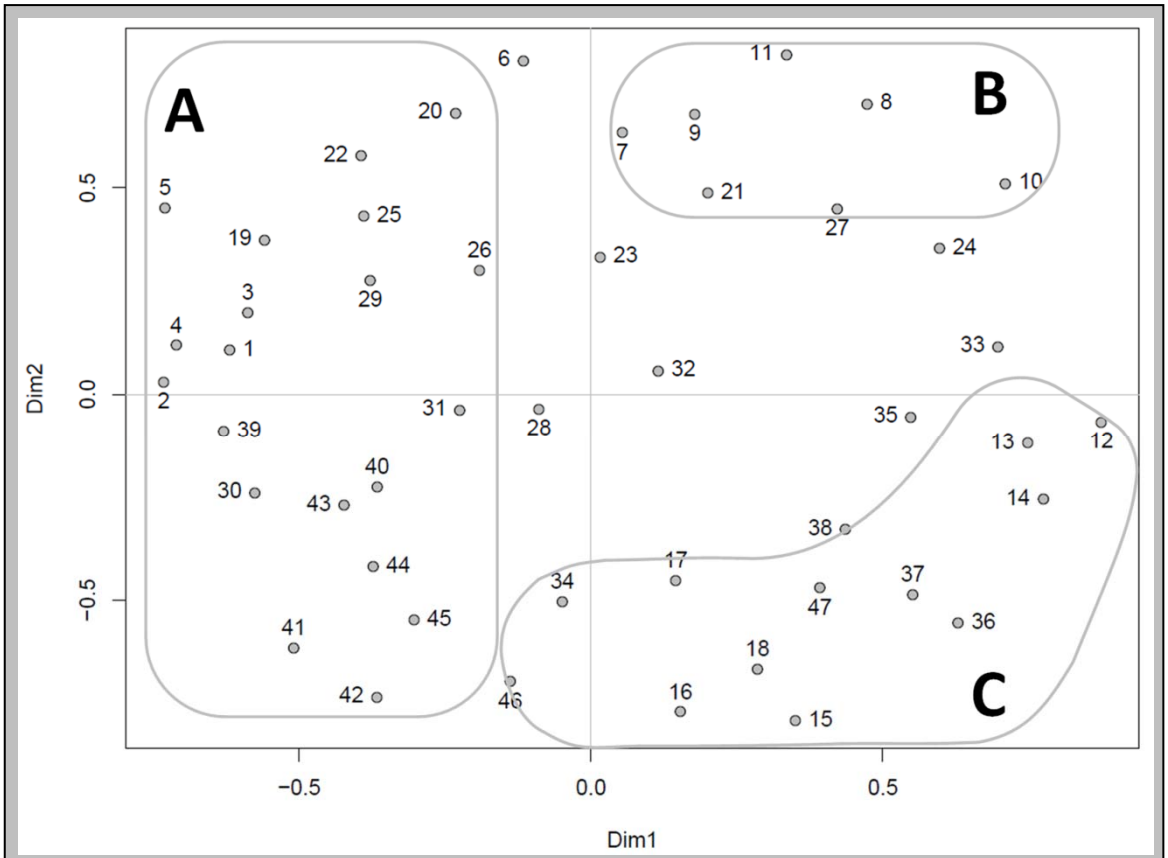


Figura 4. Estructura de la población española calculada mediante coeficiente de Hedrick y representada en un MDS (Multidimensional Scaling).

Con el objetivo de conocer si la estructura de la población española obedece homogéneamente a un modelo de aislamiento por distancia o existe rastro de influencia de factores que lo hayan alterado, se ha analizado el MDS en busca de posibles agrupaciones usando un gráfico de densidades de Kernel (Figura 5).

El gráfico de densidades incorpora a la representación del MDS la densidad de puntos. De esta manera las zonas más altas de corresponden a grupos de provincias más próximas, que tendrán una composición de apellidos más parecida; y las más bajas a zonas de transición. En el caso de la población española, podemos observar tres agrupamientos: el primero en la parte inferior izquierda (corresponde a las poblaciones del oeste - Figuras 4 y 5, A), el segundo en la zona superior (la zona en la que estaban ubicadas las provincias

del norte en el MDS - Figuras 5 y 6B) y el tercero en la zona inferior y derecha (las poblaciones del sureste - Figuras 5 y 6C).

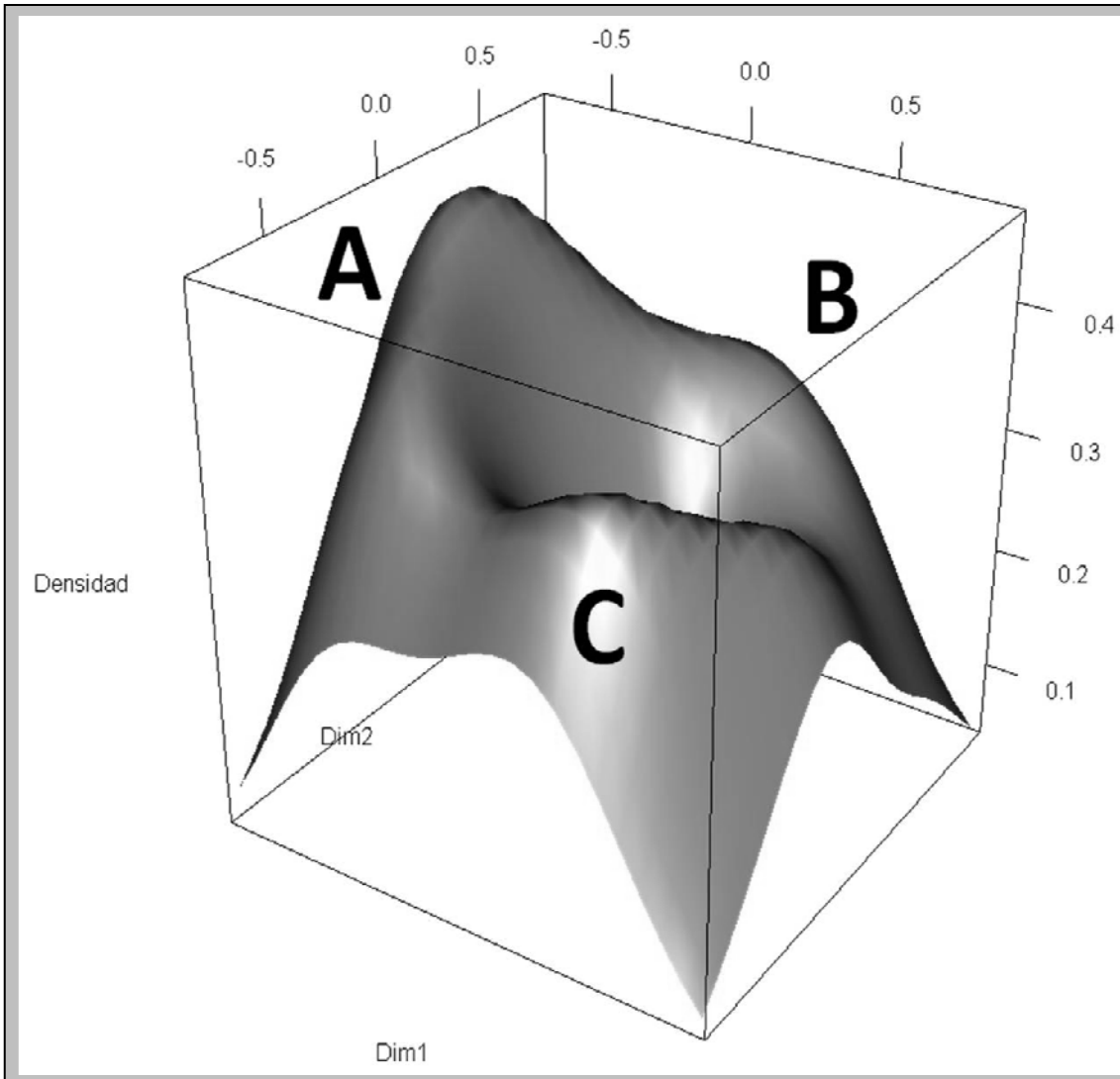


Figura 5. Gráfico de densidades de Kernel. El gráfico representa el MDS, pero añade la densidad de puntos en una tercera dimensión. Donde la densidad es más alta las poblaciones están agrupadas.

Con el propósito de analizar mejor las discontinuidades en la estructura de la población que sugieren las zonas con menor densidad del gráfico de Kernel (Figura 5) y localizarlas geográficamente, se han computado las dos primeras barreras empleando el algoritmo de Monmonier y el algoritmo "optimizado" de Monmonier (Figura 6).

Las dos primeras (Figura 6; Barrera 1, Barrera 2) se han obtenido en ese orden utilizando el procedimiento tradicional del algoritmo de Monmonier. Es decir, a

partir de los dos primeros puntos donde la diferencia entre distancia genética y distancia geográfica es mayor. Para la tercera barrera (Figura 6; Barrera Optimizada), sin embargo, se ha utilizado el procedimiento optimizado. Por este procedimiento se obtiene la barrera de mayor entidad.

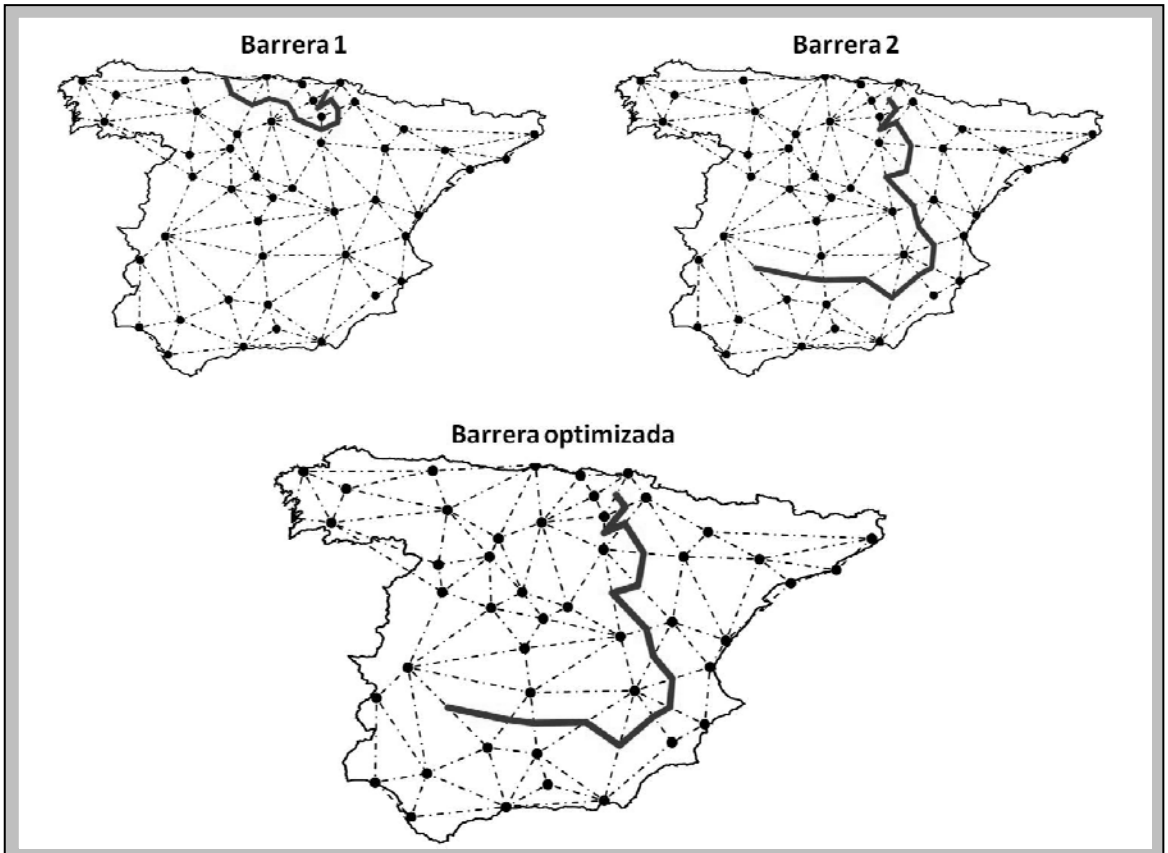


Figura 6. Barreras genéticas identificadas usando el algoritmo de Monmonier. Las barreras 1 y 2 se han calculado utilizando el algoritmo de Monmonier clásico. La barrera optimizada se ha calculado mediante el algoritmo optimizado de Monmonier.

La primera barrera se sitúa en el norte de España y separa la población del País Vasco y Cantabria del resto de España. La segunda barrera, la de mayor entidad, separa la población española en dos grandes bloques: noroeste y sureste (Figura 6). Al calcular la barrera optimizada observamos que coincide exactamente con la segunda barrera (Figura 6), lo que implica que la mayor discontinuidad se encuentra entre las poblaciones del noroeste y el sureste. El resultado del test de Wilcoxon comparando la composición de apellidos a ambos lados de la barrera optimizada muestra un resultado significativo, es decir que las diferencias entre las poblaciones de uno y otro lado de la barrera son significativas.

Por otro lado, es importante señalar también que los grupos observados en el gráfico de Kernel (Figura 5) coinciden con los grupos que separan las barreras identificadas (Figura 6). La barrera 1 separa el grupo del norte del resto y la barrera optimizada (Figura 6) separa el noroeste y el sureste. Además las zonas de transición, coinciden con las poblaciones que están en los extremos de las barreras.

Discusión

En nuestro trabajo hemos analizado los apellidos españoles de acuerdo al padrón municipal de 2008. Esta es, probablemente, la primera vez que este tipo de datos se utilizan en un estudio a nivel nacional. Por eso merece la pena detenerse un momento en la discusión de su uso.

Hasta ahora las bases de datos que han venido utilizándose en este tipo de estudios son, principalmente, de dos tipos:

a) los listines telefónicos (Rodríguez-Larralde et al., 1998; Barrai et al., 2000; Rodríguez-Larralde et al., 2003; Barrai et al., 2004; Manni et al., 2005; Scapoli et al., 2007; Cheshire et al., 2011).
b) los listados de votantes (Rodríguez-Larralde et al., 2000; Dipierri et al., 2005; Cheshire et al., 2011; Dipierri et al., 2011; Rodríguez-Larralde et al., 2011).

Aunque el tamaño de las bases de datos minimiza sus efectos, ambos tipos de fuentes de información de datos presentan sesgos. En las bases de datos procedentes de los listines telefónicos sólo aparece el titular de la línea, lo que provoca que ciertos apellidos (ciertos sustratos sociales) tiendan a aparecer menos representados porque más individuos comparten una misma línea telefónica, de la misma forma, parecen estar más representados en estos listados los varones que las mujeres (Rodríguez-Larralde et al., 2003; Cheshire et al., 2011). En los listados de votantes se deja fuera a colectivos enteros sin derecho al voto, como los inmigrantes más recientes, menores o presos por ejemplo, con lo que se subestima la presencia de grupos poblacionales enteros (Cheshire et al., 2011).

En la base de datos procedente del padrón están presentes todos los individuos de la población, por esta razón se evitan los sesgos mencionados anteriormente. Sin embargo, pueden aparecer otros. Por ejemplo, están incluidos todos los rangos de edad, algo que no nos permite estimar si el lugar en el que están los

sujetos es la población reproductora en la que se van a integrar o el lugar de inserción definitivo de sus genes puede cambiar. De cualquier manera, parece que el uso de esta base de datos presenta la posibilidad de estudiar grandes entornos geográficos con unas ventajas y una fiabilidad que no ofrecen las otras fuentes de información tradicionales.

Identificación del origen geográfico de los apellidos

El uso de los apellidos como marcadores genéticos cuenta con un amplio respaldo bibliográfico en la literatura científica. Los estudios exploratorios que han tratado de testar el método comparando sus resultados con los obtenidos mediante otras aproximaciones bien sean genéticas (Sykes e Irven, 2000; King et al., 2006; Lisa et al., 2007; King y Jobling, 2009A, 2009B; Álvarez et al., 2010) o por reconstrucciones familiares (Gagnon y Heyer, 2001; Esparza et al., 2006; Boattini et al., 2007; Rodríguez-Díaz y Blanco-Villegas, 2010) respaldan su fiabilidad. Sin embargo no se deben obviar las limitaciones que se desprenden de las características del propio método. La mayoría centradas en la estabilidad de la asociación entre apellido y linaje del cromosoma Y. Dicha asociación no siempre se ha demostrado fiable (Graf et al., 2010) y no todos los apellidos parecen corresponderse con un sólo linaje Y fundador. Si bien es cierto que no siempre se puede establecer la asociación entre apellidos y linaje Y (Graf et al., 2010), sí parece cierto que la mayoría de los apellidos pueden corresponder a un único o a muy pocos linajes Y. Por lo tanto preferimos asimilar los apellidos a marcadores geográficos. Cada población está caracterizada por una composición de apellidos que va asociada al acervo genético común de esa población (Sykes e Irven, 2000; Balanovskaia et al., 2011).

En este contexto resulta de especial utilidad la aplicación de los SOM que agrupan los apellidos en función de su distribución geográfica (Figura 2). Esto nos permite identificar su origen geográfico y considerar que todos los apellidos pertenecientes a una misma celda (Figura 2) corresponden a un mismo grupo poblacional, tienen el mismo origen y han estado sujetos a los mismos

condicionantes históricos (Manni et al., 2005). Por otro lado, la fiabilidad de los apellidos para realizar estudios poblacionales está contrastada (King y Jobling, 2009A; Balanovskaia et al., 2011). En este caso, al identificar el origen de los apellidos y eliminar aquellos sin origen claro nos aseguramos de que los apellidos con los que trabajamos caracterizan a las poblaciones que se estudian.

En este caso se ha conseguido identificar el origen del 93,13 % de los apellidos (31.433 de 33.753) lo que representa el 32,45 % de los datos (16.687.001 de 51.419.788). Una vez asegurado que estos apellidos son marcadores adecuados para la población que queremos estudiar, son los que utilizamos para el resto de los cálculos.

Aislamiento por distancia

La comparación entre las distancias de Hedrick y geográficas (Figura 3) mediante un test de Mantel muestra una correlación entre distancias genéticas y geográfica (p-value < 0,01) lo que implica que la estructura de la población obedece a un modelo de aislamiento por distancia.

Sin embargo se deben analizar con mayor detenimiento estos resultados. Normalmente, se toma la relación entre distancias genéticas y distancias geográficas como un estimador del ajuste de la metodología empleada y como un indicador del sedentarismo de la población (Gagnon y Heyer, 2001; Manni, 2005 et al.; Esparza et al., 2006; Rodríguez-Díaz y Blanco-Villegas, 2010). Tomando en cuenta esta reflexión, merece la pena merece la pena comparar la relación entre las distancias genéticas observadas en este estudio y las geográficas, con lo disponible en la bibliografía (Rodríguez-Larralde et al., 2003; Scapoli et al., 2007). En estos últimos se ha utilizado una regresión lineal, de manera que la única forma de hacer nuestros resultados comparables ha sido utilizar también una regresión lineal. El resultado es que nuestra correlación (p-value < 0,01, r = 0,491) parece más estrecha que las observadas anteriormente (Rodríguez-Larralde et al., 2003; Scapoli et al., 2007). La explicación de este mejor ajuste, reside en el uso de la información de apellidos procedente del

padrón, que nos ha permitido evitar los sesgos de la base de datos (registro telefónico) usada en los anteriores estudios (Rodríguez-Larralde et al., 2003; Scapoli et al., 2007) y la eliminación del análisis de los apellidos que no hemos podido asociar a una población prevé la subestima de las distancias genéticas entre las provincias.

Estructura genética de la población española

El coeficiente de parentesco de Hedrick entre todas las provincias españolas se ha representado en un MDS (Figura 4). En el MDS ya se aprecia una organización. Las provincias del norte y oeste están situadas en la parte superior izquierda (Figura 4, A), las del norte en la parte superior derecha (Figura 4, B) y las provincias del sur y el este en la parte inferior derecha (Figura 4, C). Es interesante señalar también que estas zonas de mayor densidad no están separadas abruptamente, sino por zonas con menor densidad que podrían reflejar zonas de transición entre los grupos y que incluyen unas pocas provincias del Oeste y el Norte. Teniendo en cuenta el stress del MDS (Stress = 0,330 < 0,363), se puede considerar que esa organización no es debida al azar sino que es consecuencia de la influencia de factores externos (Sturroch y Rocha, 2000).

Se han redistribuido las provincias para coincidir con la organización geográfica planteada por Adams (Adams et al., 2008) para comparar la estructura poblacional que observamos con la revelada por ese estudio. Se han comparado ambas matrices de distancias mediante un test de Mantel y el resultado ha sido significativo ($r = 0,1648$, $p = 0,08$). Este resultado sugiere que tanto los apellidos como los marcadores Y-STR, ambos heredados por línea paterna, muestran un mismo dibujo de la estructura de la población española. El estudio de Adams (Adams et al., 2008), maneja unas divisiones geográficas muy diferentes a las que manejamos nosotros, basadas principalmente en la disponibilidad de muestras, Por lo tanto resulta imposible estimar el alcance real de la correlación entre la variabilidad del cromosoma Y y la variabilidad de los apellidos en

España. Pero esta relación sugiere que la estructura de la población española ha mantenido un alto grado de estabilidad con el paso de los siglos, algo ya apuntado en el trabajo de Adams (Adams et al., 2008). Lo que nos hablaría de que es una población especialmente aislada y estable, unas características difíciles de encontrar y que hacen de la población española un sujeto de estudio de especial interés.

Barreras genéticas

El análisis de Kernel (Figura 5) demuestra la existencia de tres grupos dentro de la población española, pero para definirlos claramente y estudiar las causas de esta estructura hemos utilizado el algoritmo de Monmonier (Figura 6), que nos permite ver si realmente existen discontinuidades entre estos grupos y, de ser así, ubicar dónde están estas discontinuidades; lo que nos permitirá identificarlas y analizar sus orígenes (Manni et al., 2004).

Efectivamente, el análisis de las barreras (Figura 6) muestra los mismos grupos que el gráfico de Kernel, pero clarifica su organización. La primera barrera (Figura 6, Barrera 1) separa al País Vasco, principalmente (Figuras 4 y 5, B), del resto de España un hecho conocido y documentado (Bertranpetit et al., 1995; Belle et al., 2006). Esta barrera parece ser el resultado de la acción conjunta de factores tanto lingüísticos como históricos y socioculturales, pero también de factores orográficos.

La segunda barrera (Figura 6, Barrera 2 y Barrera Optimizada) resulta ser la de mayor importancia y divide España en dos grupos, Centro-Noroeste y Sureste (Figuras 4 y 5, A y C). Por otro lado, los bordes de las barreras en los que no se ha apreciado discontinuidad, coinciden con las poblaciones situadas en las zonas de transición (menor densidad) en el gráfico de Kernel (Figure 5). Las diferencias entre la composición de apellidos que hay a cada lado de la barrera (test de Wilcoxon), resulta significativa (p-value < 0,05). De hecho los apellidos con origen probable en la mitad Norte Oeste solo son portados por el 11,99 % de

la población en la mitad Sur Este. Simétricamente sólo el 19,47 % de la población del Norte Oeste lleva apellidos originarios de Sur Este.

Esta organización, aunque guarda ciertos puntos en común, difiere en lo sustancial de lo encontrado por Rodríguez-Larralde (Rodríguez-Larralde et al., 2003) y se asemeja más a lo descrito posteriormente por Scapoli (Scapoli et al., 2007). Estas divergencias pueden ser consecuencia directa de la diferente metodología empleada. En los casos anteriores, como ya se ha comentado, por un lado los autores utilizaron la base de datos derivada del listín telefónico y por el otro no se hizo un tratamiento previo de los apellidos que permitiera seleccionar solo las formas que contienen información, lo que optimiza el resultado de este tipo de estudios (Manni et al., 2005, Boattini et al., 2011).

El trazado de esta segunda barrera no coincide en absoluto con la distribución territorial de los lenguajes cooficiales (Figura 1) o la distribución peninsular de lenguajes y dialectos (García, 2007).

Comparación con datos lingüísticos

Dada la doble naturaleza de los apellidos (son marcadores poblacionales pero también palabras) se han comparado los datos lingüísticos con la matriz de parentesco de Hedrick. En este sentido la fuente más fiable de datos de la que disponemos es el ALPI (Atlas Lingüístico de la Península Ibérica) que, desgraciadamente, tuvo una azarosa historia y jamás fue publicado en su totalidad. Los datos originales se dispersaron durante la guerra civil española y la segunda guerra mundial. Existe actualmente un intento de reunir toda esta información por parte del profesor Heap (University of Western Ontario, Canada - http://westernlinguistics.ca/alpi/). Afortunadamente, se ha realizado y publicado un análisis parcial de esos datos (Goebl, 2010). Se analizó la pronunciación de 142 palabras en 529 localidades. Este análisis nos permite hacer una comparación de nuestros resultados con los datos lingüísticos (Figura 7).

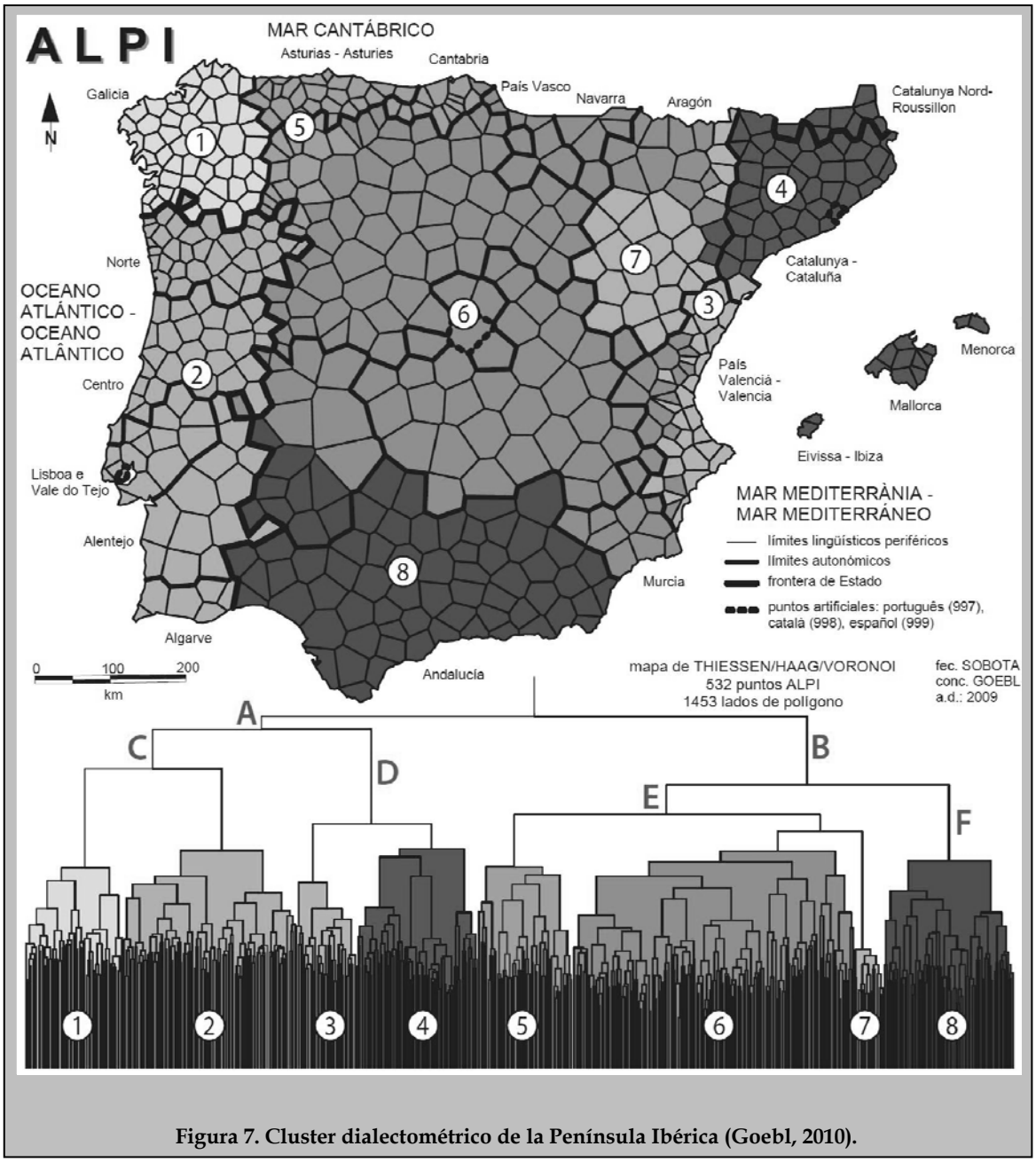


Figura 7. Cluster dialectométrico de la Península Ibérica (Goebl, 2010).

De esta comparación se desprende que la principal división de la población española no se corresponde con las principales divisiones lingüísticas observadas en la península (Figura 7, clusters D, E y F). Sin embargo si podría observarse la influencia de los factores lingüísticos en divisiones secundarias. Por ejemplo si comparamos los grupos encontrados en el MDS (Figura 4) y lo comparamos con los resultados dialectométricos (Figura 7); en el grupo sur-este, observamos cierta organización, las poblaciones catalanoparlantes se muestran más próximas entre sí y lo mismo sucede con las castellanoparlantes. Los apellidos no dejan de ser palabras y por lo tanto cabría esperar una mayor

vinculación con la estructura lingüística y, sin embargo la comparación muestra que, si bien los factores lingüísticos pueden haber tenido cierta influencia en la estructura de la población esta habrá sido en todo caso secundaria. Este hecho, ya se ha observado en otras poblaciones (Boattini et al., 2011; Barrai et al., 2002; Manni et al., 2008) y significa que las poblaciones se han mezclado con poca influencia del idioma hablado, hasta el punto que el efecto del idioma en los apellidos como marcadores lingüísticos no se percibe.

Factores Históricos y geográficos

Cuando comparamos la barrera optimizada con el mapa físico de España (Martínez, 2003) se puede observar que los dos grupos corresponden con las cuencas hidrográficas. El grupo Centro- Noroeste corresponde a las vertientes Cantábrica y Atlántica (Norte y Oeste). El grupo Sureste corresponde a la vertiente Mediterránea (Sureste). Las dos vertientes están separadas por dos cadenas montañosas (el Sistema Ibérico separa el este y el oeste y Sierra Morena separa norte y sur). Podemos hablar entonces de dos grupos principales en la población española, el de influencia cantábrica y atlántica por un lado y por el otro un grupo de influencia mediterránea.

Cuando la estructura poblacional observada se coteja con la cartografía histórica (Martínez, 2003) en busca de factores históricos que hayan podido condicionarla, se observa que esta misma estructura se repite pertinazmente a lo largo de la historia de la península ibérica. Así podemos remontarnos al neolítico, en el que ya se habla de dos arcos de influencia diferenciados, el cantábrico y el mediterráneo. Posteriormente se corresponde a la división entre pueblos celtas e íberos (Siglos I-V a. C.). Se repite también durante la dominación cartaginesa (siglo II a. C.), la dominación romana (Siglos I y II a. C.) o la división hispano-árabe de la península (Siglos VII-X). Más recientemente, idéntica división se observa durante los siglos XII y XIII, precisamente la época de aparición de los apellidos, algo que entra dentro de lo esperable. De nuevo

es la división que se da durante prácticamente toda de la edad media (desde el siglo XVII) y llega hasta los siglos XIX y XVIII.

El hecho de que la estructura observada por isonimia no guarde relación alguna con ninguna división de tipo etno-lingüístico o socioeconómico, unido a de que refleja fielmente la estructura geográfica peninsular y que esta misma división se encuentra repetidamente a lo largo de la historia, hacen pensar los factores que han condicionado la actual estructura de la población española, no son solo los que se pueden observar desde que se utilizan los apellidos (siglo XIII), como permitiría asegurar la profundidad de los apellidos como marcadores; sino que en realidad es el reflejo de la acción de factores mucho más antiguos que ya estarían influenciando los primeros asentamientos peninsulares (Martínez, 2003).

Esta antigüedad de la acción de los factores que condicionan la estructura de la población española concuerda también con la impresión desprendida de los estudios genéticos de que es una población altamente conservada (Adams et al, 2008). Algo que por otra parte, refrenda la relación existente entre lo observado en el estudio de los apellidos y el resultado del análisis de marcadores genéticos (Adams et al, 2008).

La larga permanencia de estos factores y su distribución espacial, invita a pensar en los factores de tipo geográfico (Barrai et al., 2000; Esparza et al., 2006; Rodríguez-Díaz y Blanco-Villegas, 2010). La conformación del territorio (cuencas hidrográficas) determina el trazado de las barreras que condicionan la movilidad de los individuos y en definitiva, la estructura biológica de la población. Mientras que pese a lo observado en algunas poblaciones (Rodríguez-Larralde et al., 1998; Barrai et al., 2004) y de acuerdo con lo constatado en otras (Manni et al., 2004; Boattini et al., 2011) los factores lingüísticos no parecen ser un factor de aislamiento por sí mismos.

Conclusión

Por primera vez en un estudio de isonimia a nivel nacional se han utilizado todos los apellidos contenidos en el padrón como base de datos. Esta fuente de información presenta una importante ventaja frente a las clásicas al incluir a todos los individuos de una población, evitando así los sesgos a los que están sujetas las bases de datos tradicionales. Por otro lado, con esta base de datos se trabaja con el total de la población, evitando así posibles desviaciones al usar estimaciones.

El estudio previo de la distribución de los apellidos, permite controlar la calidad de la información de la base de datos y utilizar los apellidos (monofiléticos) como marcadores poblacionales seguros.

En el caso de España, la estructura de la población obedece a la influencia conjunta de una serie de factores diferentes. En primer lugar, la población española obedece a un modelo de aislamiento por distancia. El ajuste obtenido a este modelo utilizando los datos procedentes del padrón es mayor que el observado con anterioridad.

Poblacionalmente, el estudio de la isonimia en España muestra la península dividida en dos mitades bien diferenciadas, una Nor-Oeste y otra Sur-este. En la división Nor-Oeste, el país vasco aparece también como un grupo diferenciado. Esta estructura poblacional no aparece relacionada con ningún factor etnolingüistico, sin embargo parece ser una estructura que se repite a lo largo de toda la historia peninsular. Este hecho junto a la correlación de nuestros resultados con los obtenidos en estudios genéticos previos, hacen pensar: primero, que esta estructura está reflejando una historia biológica que va mucho más allá de lo que la antigüedad de los apellidos como marcadores (siglo XIII) permitiría afirmar. Y en segundo lugar, que los factores que han determinado la estructura genética de la población española son de índole geográfico; más concretamente la conformación del territorio peninsular en dos cuencas hidrográficas (arco cantábrico - arco mediterráneo).

Bibliografía

Adams SM, Bosch E, Balaresque PL, Ballereau S J, Lee A C, et al. (2008) The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. Am J Hum Genet 83: 725-736.

Alvarez L, Santos C, Ramos A, Pratdesaba R, Francalacci P, et al. (2010) Mitochondrial DNA patterns in the Iberian Northern plateau: Population dynamics and substructure of the Zamora province. Am J Phys Anthropol 142: 531-539.

Balanovskaia EV, Romanov A G, Balanovskii OP (2011) Namesakes or relatives? Approaches to investigating the relationship between Y chromosomal haplogroups and surnames. Mol Biol 45: 473-485.

Barbujani G, Stenico M, Excoffier L, Nigro L (1996) Mitochondrial DNA sequence variation across linguistic and geographic boundaries in Italy. Hum Biol 68: 201-215.

Barrai I, Rodríguez-Larralde A, Mamolini E, Manni F, Scapoli C (2000) Elements of the surname structure of Austria. Ann Hum Biol 27: 607-622.

Barrai I, Rodríguez-Larralde A, Manni F, Ruggiero V, Tartari D, et al. (2004) Isolation by language and distance in Belgium. Ann Hum Genet 68: 1-16.

Barrai I, Rodriguez-Larralde A, Manni F, Scapoli C (2002) Isonymy and isolation by distance in the Netherlands. Hum biol 7 42: 263-283.

Belle EM, Landry PA, Barbujani G (2006) Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. Proc Biol Sci 273: 1595-1602.

Bertranpetit J, Sala J, Calafell F, Underhill PA, Moral P, et al. (1995) Human mitochondrial DNA variation and the origin of Basques. Ann Hum Genet 59: 63-81.

Blanco-Villegas MJ, Boattini A, Otero HR, Pettener D (2004) Inbreeding patterns in La Cabrera, Spain: Dispensations, multiple consanguinity analysis, and isonymy. Hum Biol 76: 191–210.

Boattini A, Blanco-Villegas MJ, Pettener D (2007) Genetic structure of La Cabrera, Spain, from surnames and migration matrices. Hum Biol 79: 649-666.

Boattini A, Calboli FC, Blanco-Villegas MJ, Gueresi P, Franceschi MG, et al. (2006) Migration matrices and surnames in populations with different

isolation patterns: Val di Lima (Italian Apennines), Val di Sole (Italian Alps), and La Cabrera (Spain). Am J Hum Biol 18: 676-690.

Boattini A, Griso C, Pettener D (2011) Are ethnic minorities synonymous for genetic isolates? Comparing Walser and Romance populations in the Upper Lys Valley (Western Alps). J Anthropol Sci 89: 161-173.

Boattini A, Lisa A, Fiorani O, Zei G, Pettener D, et al. (2012) General Method to Unravel Ancient Population Structures through Surnames, Final Validation on Italian Data. Hum Biol 84: 235-270.

Caravello GU, Tasso M (1999) An analysis of the spatial distribution of surnames in the Lecco area (Lombardy, Italy) Am J Hum Biol 11: 305-315.

Caravello GU, Tasso M (2002) Use of surnames for a demo-ecological analysis: a study in southwest Sardinia. Am J Hum Biol 14: 391-397.

Chakraborty R, Schwartz RJ (1990) Selective neutrality of surname, distribution in an immigrant indian community of Houston, Texas. Am J Hum Biol 2: 1-15.

Cheshire J, Mateos P, Longley PA (2011) Delineating Europe's Cultural Regions: Population Structure and Surname Clustering. Hum Biol 83: 573-598.

Colantonio SE, Lasker GW, Kaplan BA, Fuster V (2003) Use of surname models in human population biology: a review of recent developments. Hum Biol 75:785-807.

Crow JF, Mange AP (1965) Measurement of inbreeding from the frequency of marriages between persons of the same surname. Eugen Q 12: 199-203.

Dipierri J, Rodríguez-Larralde A, Alfaro E, Scapoli C, Mamolini E, et al. (2011) A Study of the Population of Paraguay through Isonymy. Ann Hum Genet 75: 678-687.

Dipierri JE, Alfaro EL, Scapoli C, Mamolini E, Rodríguez-Larralde A, et al. (2005) Surnames in Argentina: a population study through isonymy. Am J Phys Anthropol 128: 199-209.

Esparza M, García-Moro C, Hernández M (2006) Genetic relationships between parishes in the Ebro delta region (Spain) as estimated by migration matrix and surnames. Hum Biol 78: 647-662.

Faure R, Ribes MA, García A (2001) Diccionario de apellidos españoles. Madrid: Espasa-Calpe.

Fiorini S, Tagarelli G, Boattini A, Luiselli D, Piro A, et al. (2007) Ethnicity and Evolution of the Biodemographic Structure of Arbëreshe and Italian

Populations of the Pollino Area, southern Italy (1820-1984). A. Anthropol 109: 735-746.

Gagnon A, Heyer E (2001) Intergenerational correlation of effective family size in early Quebec (Canada). Am J Hum Biol 13: 645-659.

García P (2007) Lenguas y dialectos de España. Madrid: Arco Libros.

Goebl H (2010) La dialectometrización del ALPI: Rápida presentación de los resultados. 26th CILFR. Valencia.

Graf OM, Zlojutro M, Rubicz R, Crawford MH (2010) Surname distributions and their association with Y-chromosome markers in the Aleutian Islands. Hum Biol 82: 745-757.

Hedrick PW (1971) A new approach to measuring genetic similarity. Evolution 25: 276-280.

Jombart T (2008) Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics, 24: 1403-1405.

King TE, Ballereau SJ, Schurer KE, Jobling MA (2006) Genetic signatures of coancestry within surnames. Curr Biol 16: 384-388.

King TE, Jobling MA (2009) Founders, drift and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. Mol Biol Evol 26:1093-102.

King TE, Jobling MA (2009) What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. Trends Genet 25: 351-360.

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cyber 43: 59-69.

Kohonen T (1984) Self-organization and associative memory. Berlin: Springer.

Lasker GW (1977) A coefficient of relationship by isonymy: A method for estimating the genetic relationship between populations. Hum Biol 49: 489-493.

Lasker GW, Kaplan BA (1985) Surnames and genetic structure: repetition of the same pairs of names of married couples, a measure of subdivision of the population. Hum biol 57: 431-440.

Lisa A, De Silvestri A, Mascaretti L, Degiuli A, Guglielmino CR (2007) HLA genes and surnames show a similar genetic structure in Lombardy: does this reflect part of the history of the region?. Am J Hum Biol 19: 311-318.

Malecot G (1955) Decrease of relationship with distance. Cold Spring Harbor Syrup. Quant Biol 20: 52-53.

Manni F, Barrai I (2001) Genetic structures and linguistic boundaries in Italy: a microregional approach. Hum Biol 73: 335-347.

Manni F, Guerard E, Heyer E (2004) Geographic patterns of (genetic, morphologic, linguistic) variation: how barriers can be detected by using Monmonier's algorithm. Hum Biol 76: 173-190.

Manni F, Heeringa W, Toupance B, Nerbonne J (2008) Do surname differences mirror dialect variation? Hum biol 80: 41-64.

Manni F, Toupance B, Sabbagh A, Heyer E (2005) New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Y-chromosome sampling. Am J Phys Anthropol 126: 214-228.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27: 209-220.

Mardia KV, Kent JT, Bibby JM (1979) Multivariate Analysis. London: Academic Press.

Martínez E (2003) Atlas histórico de España. Madrid: Istmo.

Mateos P, Tucker DK (2008) Forenames and Surnames in Spain in 2004. Names: A Journal of Onomastics, 56(3), 165–184.

Monmonier MS (1973) Maximum-difference barriers: An alternative numerical regionalization method. Geographical Analysis 5: 245-261.

Morton NE (1973) Kinship bioassay. In: Crow JF, Denniston C, editors. Genetic distance. NewYork: Plenum Press. p 97-104.

Morton NE, Yee S, Harris DE, Lew R (1971) Bioassay of Kinship. Theor Popul Biol 2: 507-524.

Nicholas J, Lewin-Koh RB, contributions by Pebesma EJ, Archer E, Baddeley A, et al. (2012) Maptools: Tools for reading and handling spatial objects. R package version 0.8-14. http://CRAN.R-project.org/package=maptools.

Palme AE, Su Q, Rautenberg A, Manni F, Lascoux M (2003) Postglacial recolonization and cpDNA variation of silver birch, Betula pendula. Molec Ecol 12: 201-212.

Pettener D, Pastor S, Tarazona-Santos E (1998) Surnames and genetic structure of a high-altitude quechua community from the Ichu river valley peruvian central Andes 1825-1914. Hum Biol 70: 865-87.

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL http://www.R-project.org/.

Relethford JH (1988) Estimation of kinship and genetic distance from surnames. Hum Biol 60: 475-492.

Rodríguez-Díaz R, Blanco-Villegas MJ (2010) Genetic structure of a rural region in Spain: distribution of surnames and gene flow. Hum Biol 82: 301-314.

Rodríguez-Larralde A, Dipierri J, Gómez EA, Scapoli C, Mamolini E, et al. (2011) Surnames in Bolivia: A study of the population of Bolivia through isonymy. Am J Phys Anthropol 144: 177-184.

Rodríguez-Larralde A, Gonzales-Martin A, Scapoli C, Barrai I. 2003. The names of Spain: a study of the isonymy structure of Spain. Am J Phys Anthropol 121: 280-292.

Rodríguez-Larralde A, Morales J, Barrai I (2000) Surname frequency and the isonymy structure of Venezuela. Am J Hum Biol 12: 352-362.

Rodríguez-Larralde A, Scapoli C, Beretta M, Nesti C, Mamolini E, et al. (1998) Isonymy and the genetic structure of Switzerland. II. Isolation by distance. Ann Hum Biol 25: 533-540.

Rogers AR (1991) Doubts about isonymy. Hum Biol 63: 663-668.

Scapoli C, Mamolini E, Carrieri A, Rodríguez-Larralde A, Barrai I (2007) Surnames in Western Europe: a comparison of the subcontinental populations through isonymy. Theor Popul Biol 71: 37-48.

Solís JA (2002) El gran libro de los apellidos. La Coruña: El arca de papel.

Sturroch K, Rocha J (2000) A multidimensional scaling stress evaluation table. Field Methods 12: 49-60.

Sykes B, Irven C (2000) Surnames and the Y Chromosome. Am J Hum Genet 66: 1417-1419.

Thioulouse J, Chessel D, Dole´dec S, Olivier JM (1997) ADE-4: a multivariate analysis and graphical display software. Stat Comput 7: 75-83.

Venables WN, Ripley BD (2002) Modern Applied Statistics with S. Fourth Edition. New York: Springer.

Wehrens R, Buydens LMC (2007) Self- and Super-organising Maps in R: the Kohonen package. J Stat Softw, 21.

Weiss V (1980) Inbreeding and genetic distance between hierarchically structured populations measured by surname frequencies. Mankind Quarterly 21: 135-149.

Wilcoxon F (1945) Individual Comparisons by Ranking Methods. Biometrics Bulletin 6: 80-83.

2. *MOVIMIENTOS MIGRATORIOS DE LA POBLACIÓN ESPAÑOLA*

MOVIMIENTOS MIGRATORIOS DE LA POBLACIÓN ESPAÑOLA

Resumen

Introducción: Se ha utilizado una técnica de minería de datos auto-organizados (SOM) con el objetivo de representar sus procesos demográficos pasados. El análisis permite la comprensión de los movimientos históricos que se han producido en la población española y cómo y en qué medida estos movimientos han gobernado la estructura genética actual del país.

Material y métodos: La técnica identifica los grupos de apellidos con el mismo origen. Una vez establecido el origen de cada, podemos suponer que cada sujeto que porta un apellido encontrado fuera de su origen, en algún momento en el pasado, se ha trasladado. Este concepto permite el estudio de los movimientos históricos.

Resultados: Varios tipos de movimientos han tenido lugar (aislamiento por distancia, movimientos a corta distancia, movimientos de media distancia y movimientos de larga distancia). Los de corta y media distancia han sido los más frecuentes y los que en mayor medida han determinado la estructura actual.

Los movimientos observados revelan la existencia de dos arcos migratorios principales. Ambos se han movido a lo largo de la costa; el primero ha seguido el mar Mediterráneo y el segundo el mar Cantábrico. Parece que estos dos arcos son los que han constituido la columna vertebral de la población española, dividiéndola en dos mitades que alcanzarían hasta donde llegan las zonas de influencia de estos arcos.

Introducción

Recientemente, el desarrollo de las tecnologías informáticas ha aumentado la disponibilidad de bases de datos poblacionales que permiten analizar grandes grupos poblacionales, a nivel de país (Rodríguez-Larralde et al., 1998; Barrai et al., 2000; Rodríguez-Larralde et al., 2000; Rodríguez-Larralde et al., 2003; Barrai et al., 2004; Dipierri et al., 2005; Manni et al., 2005; Dipierri et al., 2011; Rodríguez-Larralde et al., 2011) o incluso de continente (Scapoli et al., 2007; Cheshire et al., 2011). Tradicionalmente estas amplias bases de datos han sido registros de votantes y listines telefónicos (Rodríguez-Larralde et al., 2003; Cheshire et al., 2011) y más recientemente censos poblacionales, representando una magnífica fuente de información para estudios poblacionales. En ese tipo de trabajos, se espera que sea el amplio tamaño de la muestra el que minimice las posibles desviaciones del uso de los apellidos como estimador (Barrai et al., 2000; Rodríguez-Larralde et al., 2003).

En el año 2005 (Manni et al., 2005) se propone la aplicación de una técnica de minería de datos (Self Organized Maps) a grandes bases de datos biodemográficas. El propósito era poder analizarlas sin necesidad de acudir a registros genealógicos. La técnica permitió la identificación de grupos de apellidos con un mismo origen, en otras palabras, se logró distinguir los apellidos autóctonos de cada zona, y de esta forma poder utilizar los apellidos con origen conocido como marcadores fiables. Los trabajos realizados hasta el momento avalan la fiabilidad del método (Manni et al., 2005; Boattini et al., 2010; Rodríguez-Díaz y Blanco-Villegas, 2010; Boattini et al., 2012). Sin embargo las bases de las bases de datos utilizadas, tenían limitaciones que reducían el alcance de las conclusiones extraídas. En algunos casos el tamaño de la base de datos era limitado como consecuencia del reducido ámbito geográfico (Boattini et al., 2010; Rodríguez-Díaz y Blanco-Villegas, 2010) en otros los apellidos están muy recientemente asentados (Manni et al., 2005). La validación de esta metodología se realizó aún más recientemente sobre el caso particular de los apellidos italianos (Boattini et al., 2012). En ese trabajo se comprobó la validez

del método comparando los orígenes identificados para cada apellido con bases de datos preexistentes, obteniendo magníficos resultados (Boattini et al., 2012). Para completar el recorrido de la metodología faltaría testar su funcionamiento y resultados sobre una población diferente.

La población seleccionada para tal fin es la española, una población idónea para este tipo de estudio. Durante siglos la influencia española sobre otros países ha sido muy importante. Esta influencia se ha traducido en una amplísima presencia del sistema español de apellidos principalmente en Sudamérica (Rodríguez-Larralde et al, 2000; Dipierri et al, 2005, 2011), aunque no sólo (Scapoli et al, 2007; Cheshire et al, 2011) por lo que su estudio resulta de especial interés.

Por otro lado la española es una población que ha estado sometida a unas condiciones especiales. Geográficamente está situada en el extremo del continente europeo y separada de él por la cordillera pirenaica. A este aislamiento se suma un relieve complicado en comparación al resto de países europeos. Esta alta diversidad de factores se observa también en el ámbito lingüístico, dentro de la población española coexisten también diferentes lenguas oficiales y variedades de las mismas (García, 2007; Goebl, 2010). Todas estas condiciones han hecho que la población española esté sometida a la influencia constante de una amplia variedad de factores.

Estas condiciones (gran variedad de factores, alto grado de aislamiento y conservación; y la distribución global de su sistema de apellidos) son características que hacen de la población española la idónea para la aplicación definitiva de estas nuevas metodologías que permitirán obtener unas conclusiones amplísimamente extrapolables a otras poblaciones. Este es el segundo objetivo de este trabajo, conocer en profundidad la estructura genética actual de la población española y los procesos que han desembocado en ella.

Material y métodos

Objetivo

Este capítulo persigue la comprensión de los movimientos históricos que se han producido en el interior de la población española y cómo y en qué medida han determinado la estructura genética actual. Para ello se han analizado los apellidos de toda la población española e identificado el origen de cada grupo de apellidos utilizando por primera vez una metodología reciente, que fue propuesta sobre el caso particular de la población Holandesa (Manni et al, 2005) y testada sobre la población italiana (Boattini et al, 2012).

Zona de estudio

España es un país situado en la península ibérica en el extremo sur-oeste del continente europeo, tiene una superficie de 504.645 Km^2 y está rodeado por el mar por Norte, sur y Este. Por tierra, tiene fronteras con Portugal por el Oeste y Francia por el Noreste.

Geográficamente, hablamos de un territorio situado a bastante altitud sobre el nivel del mar, con un promedio de 660 m y de marcado carácter montañoso en comparación con el promedio de los países europeos.

Tiene una población de 47 millones de habitantes desigualmente repartidos, que se concentran principalmente en los entornos costeros, dejando el interior del país con baja densidad poblacional (excepción hecha de Madrid, capital administrativa).

Administrativamente, la España peninsular está organizada en 15 comunidades autónomas y 47 Provincias. El idioma oficial de España es el español o castellano, pero comparte espacio (Figura 1) con otras lenguas cooficiales en algunos territorios (catalán, gallego y euskera).

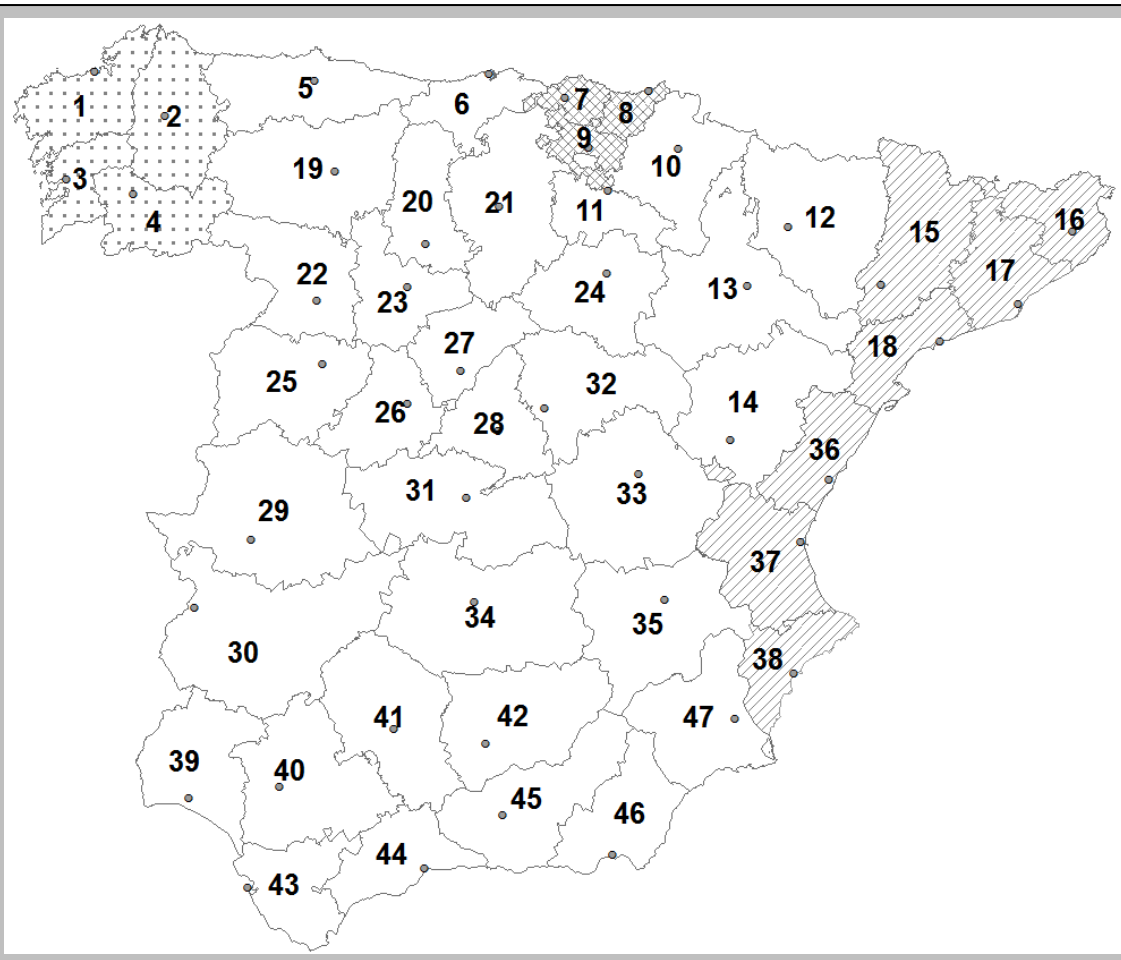


Figura 1. Mapa provincial y extensión geográfica de los idiomas cooficiales.

Base de datos Dada la doble naturaleza de los apellidos (son marcadores poblacionales pero también palabras) se han comparado los datos lingüísticos con la matriz de parentesco de Hedrick. En este sentido la fuente más fiable de datos de la que disponemos es el ALPI (Atlas Lingüístico de la Península Ibérica) que, desgraciadamente, tuvo una azarosa historia y jamás fue publicado en su totalidad. Los datos originales se dispersaron durante la guerra civil española y la segunda guerra mundial. Existe actualmente un intento de reunir toda esta información por parte del profesor Heap (University of Western Ontario, Canada - http://westernlinguistics.ca/alpi/) . Afortunadamente, se ha realizado y publicado un análisis parcial de esos datos (Goebl, 2010). Se analizó la pronunciación de 142 palabras en 529 localidades. Este análisis nos permite hacer una comparación de nuestros resultados con los datos lingüísticos (Figura 7).

En el sistema de transmisión de apellidos español, se heredan dos apellidos. Cada sujeto hereda el primer apellido del padre (será su primer apellido) y el primero de la madre (será su segundo apellido). Dada la disponibilidad de los dos apellidos de cada individuo, la base de datos se construyó utilizando tanto el primero como el segundo apellido algo que según diversos autores (Pettener et al, 1998; Colantonio et al., 2003, Dipierri et al., 2011), duplica la cantidad de información y contribuye a la robustez del análisis.

Los datos sobre los apellidos fueron facilitados por el INE y proceden del padrón de población de 2008. En la base de datos están incluidos todos los apellidos de cada municipio siempre que aparezcan un mínimo de 5 veces. La base de datos inicial incluía 56976706 registros, correspondientes a 87.148 apellidos diferentes.

Corrección y Depurado de los datos

La base de datos inicial fue minuciosamente revisada. Se encontraron apellidos repetidos, diferentes grafías del mismo apellido, espacios entre las palabras, errores ortográficos y formas compuestas. Todas estas incidencias representan un problema a la hora del procesado estadístico. Para evitar estos problemas, todos los apellidos fueron revisados con apoyo bibliográfico (Faure et al, 2001; Solís, 2002) y cartográfico; y corregidas cuando fue necesario.

En el siguiente paso se eliminaron todos aquellos apellidos que no aparecían un mínimo de 20 veces en la base de datos a fin de evitar el exceso de ruido en los procedimientos estadísticos (Manni et al, 2005; Boattini et al, 2012). En total, una vez finalizado el depurado de los datos se dispone de un total de 51.419.788 datos (33.753 apellidos diferentes).

Procesado de los datos

Terminado el depurado se ha construido una matriz de doble entrada, en la que las filas (i) corresponden a cada apellido y las columnas (j) a cada provincia. De

esta forma cada casilla (ij) corresponde a la frecuencia que representa cada apellido en la población total de cada población.

Posteriormente se ha realizado una transformación de las frecuencias en dos pasos (Boattini et al, 2012):

- En un primer paso, se pretende evitar que las poblaciones más pequeñas tengan un peso excesivo. Para ello se aplica la siguiente expresión:

$$f_i = \frac{fabs_{ij}}{\log(pop_j)}$$

  Siendo $fabs_{ij}$ la frecuencia absoluta del apellido "i" en la provincia "j" y pop_j es la población total de la provincia "j"

- En un segundo paso, se pretende evitar que los apellidos se agrupen en función de lo numerosos que sean, empleando a la expresión:

$$wf_i = \frac{f_i}{\Sigma f_i}$$

  Siendo f_i el resultado de la expresión anterior.

Agrupamiento de los apellidos

Los apellidos se han agrupado en función de su distribución geográfica utilizando un procedimiento estadístico tipo Cluster de minería de datos, los mapas autoorganizados de Kohonen o SOM (Kohonen, 1982; 1984).

Los SOM son redes neuronales de aprendizaje no supervisado que permite obtener un reconocimiento estadístico de patrones. En este estudio se ha utilizado para el reconocimiento de patrones en la distribución geográfica de los apellidos. Este es un procedimiento que permite agrupar los apellidos en función de su distribución y analizar esta para identificar sus orígenes. Su aplicación en el campo de la Biodemografía es reciente (Manni et al, 2005) pero es una metodológia testada (Boattini et al, 2012) y que ha demostrado arrojar

buenos resultados (Boattini et al, 2010; Rodríguez-Díaz y Blanco-Villegas, 2010).

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	
1	379 109.913 38	347 328.049 38	249 226.727 42	328 225.552 47	340 93.148 47	128 47.302 36	370 96.345 36	96 70.696 36	561 129.934 37	268 557.946 37	366 81.971 16	792 87.851 16	Nº de apellidos Nº de datos Origen
2	330 51.482 38	130 118.756 38	111 26.249 35	403 604.140 47	250 152.672 46	367 102.035 ??	267 128.261 36	144 313.877 37	333 242.288 37	44 73.291 ??	103 28.804 15	903 29.331.181 ??	Nº de apellidos Nº de datos Origen
3	188 21.933 ??	251 455.480 44	182 33.015 25	129 10.691 17	146 22.587 46	227 92.956 31	228 38.158 31	164 27.253 34	186 26.493 42	123 67.768 47	204 133.137 16	58 7.433 16	Nº de apellidos Nº de datos Origen
4	273 29.890 44	246 84.016 44	179 29.022 45	124 64.576 18	326 87.355 18	236 381.156 43	283 23.088 28	158 32.552 33	83 9.750 27	167 20.590 ??	349 408.258 13	802 111.898 13	Nº de apellidos Nº de datos Origen
5	80 69.273 44	199 142.155 45	101 36.528 39	109 48.741 15	422 60.216 18	178 102.726 27	197 12.266 28	338 100.614 28	160 21.775 32	167 31.273 14	84 27.163 13	566 148.503 13	Nº de apellidos Nº de datos Origen
6	478 56.573 40	320 68.054 40	163 22.763 39	116 60.473 39	230 144.560 18	218 142.997 34	203 94.418 23	92 19.478 26	525 8.632.184 ??	76 13.787 24	241 83.500 13	290 79.596 12	Nº de apellidos Nº de datos Origen
7	110 49.251 40	274 556.362 40	25 14.520 ??	15 14.190 ??	206 131.876 19	147 19.111 23	164 35.566 19	110 12.916 21	223 102.909 11	162 31.141 11	92 20.294 12	189 17.238 12	Nº de apellidos Nº de datos Origen
8	447 59.822 43	259 67.081 43	237 162.893 5	423 118.263 5	150 375.211 5	63 6.980 20	196 46.725 22	164 18.160 29	223 100.187 29	194 563.756 41	139 197.834 35	254 46.221 15	Nº de apellidos Nº de datos Origen
9	17 2.858 ??	230 1.235.934 3	232 562.171 1	40 11.582 2	48 37.836 5	265 45.963 6	106 140.030 30	233 22.481 30	213 225.085 30	174 16.408 41	8 567 41	474 42.414 15	Nº de apellidos Nº de datos Origen
10	460 150.192 3	343 214.038 3	145 17.041 4	226 29.594 2	214 102.994 2	191 124.801 6	151 55.807 20	267 72.341 30	220 71.307 41	195 116.422 7	328 523.706 38	199 67.488 15	Nº de apellidos Nº de datos Origen
11	184 371.366 1	158 101.690 4	69 26.806 ??	0 0 ??	88 142.773 10	119 45.253 8	344 257.614 8	170 147.591 8	150 33.320 9	210 90.025 21	135 81.247 7	401 137.154 7	Nº de apellidos Nº de datos Origen
12	596 186.413 1	366 296.365 1	214 78.682 10	485 53.841 10	446 154.402 10	162 73.258 10	408 79.776 8	138 52.789 8	129 10.962 9	135 43.097 7	83 43.915 6	644 93.542 7	Nº de apellidos Nº de datos Origen

Tabla 1. Tabla resumen del SOM. Muestra el número de apellidos y el número de datos agrupados en cada celda; y el origen de cada grupo de apellidos.

En nuestro caso el software utilizado ha sido el paquete de R-project "Kohonen" (Wehrens y Buydens, 2007). Utilizando este software hemos clasificado los apellidos en una matriz rectangular. El tamaño de dicha matriz debe establecerse. El criterio debe ser elegir un tamaño que no sea tan grande que imposibilite su interpretación ni tan pequeño que los resultados no resulten representativos. Para conseguir esto, el criterio que hemos adoptado tras probar diferentes tamaños, es utilizar la matriz más pequeña en la que aparezca una celda vacía (Boattini et al, 2011), de esta manera utilizamos el tamaño más pequeño para el que todos los grupos ya son representativos (por eso empiezan a aparecer celdas vacías), que en nuestro caso es de 12 celdas de ancho por 12 de alto, con 1.000 repeticiones (Tabla 1).

En definitiva, el SOM consta de una capa de entrada de 33.753 vectores (un vector por cada apellido) y una capa de salida de 144 celdas (grupos de apellidos con similar distribución geográfica).

Origen de los apellidos

Finalmente, cada grupo de apellidos se ha representado gráficamente mediante mapas de gradientes utilizando el software Arcgis 10.0 lo que permite observar su distribución geográfica (Figura 2).

Observando estos mapas de gradientes, podemos identificar el origen de cada grupo poblacional en función del apellido. El método se basa en la asunción de que un grupo poblacional será tanto más numeroso cuanto más nos acerquemos a su origen (Manni et al, 2005, Boattini et al, 2012). De esta manera, observando el mapa de gradientes de la distribución de cada apellido (Figura 2) podemos asumir que el grupo poblacional que porta ese apellido tiene su origen en el lugar donde este apellido es más numeroso (Figura 1).

Matrices de migración

Una vez identificado el origen de cada apellido, el siguiente paso natural es asumir que cada apellido que encontremos fuera de su origen salió en algún momento del pasado de él (Boattini et al, 2012). Partiendo de esta base, se ha construido una matriz de migración (Bodmer y Cavalli-Sforza, 1968), con tantas filas ("i") y columnas ("j") como provincias incluye el estudio (47x47). De manera que cada celda ("ij") recoge el número de veces que un apellido con origen en la población "i" a aparecido en la población "j".

Esta metodología permite el estudio de los movimientos poblacionales históricos que han tenido lugar en el interior de una población y cuál ha sido su contribución a la estructura biológica de dicha población (Boattini et al, 2012)

Censos históricos

Los resultados obtenidos mediante la isonimia, han sido comparados con la población histórica. Concretamente, se han realizado regresiones entre los censos históricos de cada provincia y la población autóctona (nº de portadores de apellidos autóctonos) de cada provincia.

La información sobre los censos históricos hasta 1.857 está disponible en los censos históricos del Instituto Nacional de Estadística (www.ine.es). Para remontarnos aún más, hemos consultado el censo de Floridablanca de 1.787 disponible en la biblioteca digital de la Real Academia de la Historia (www.bibliotecadigital.rah.es).

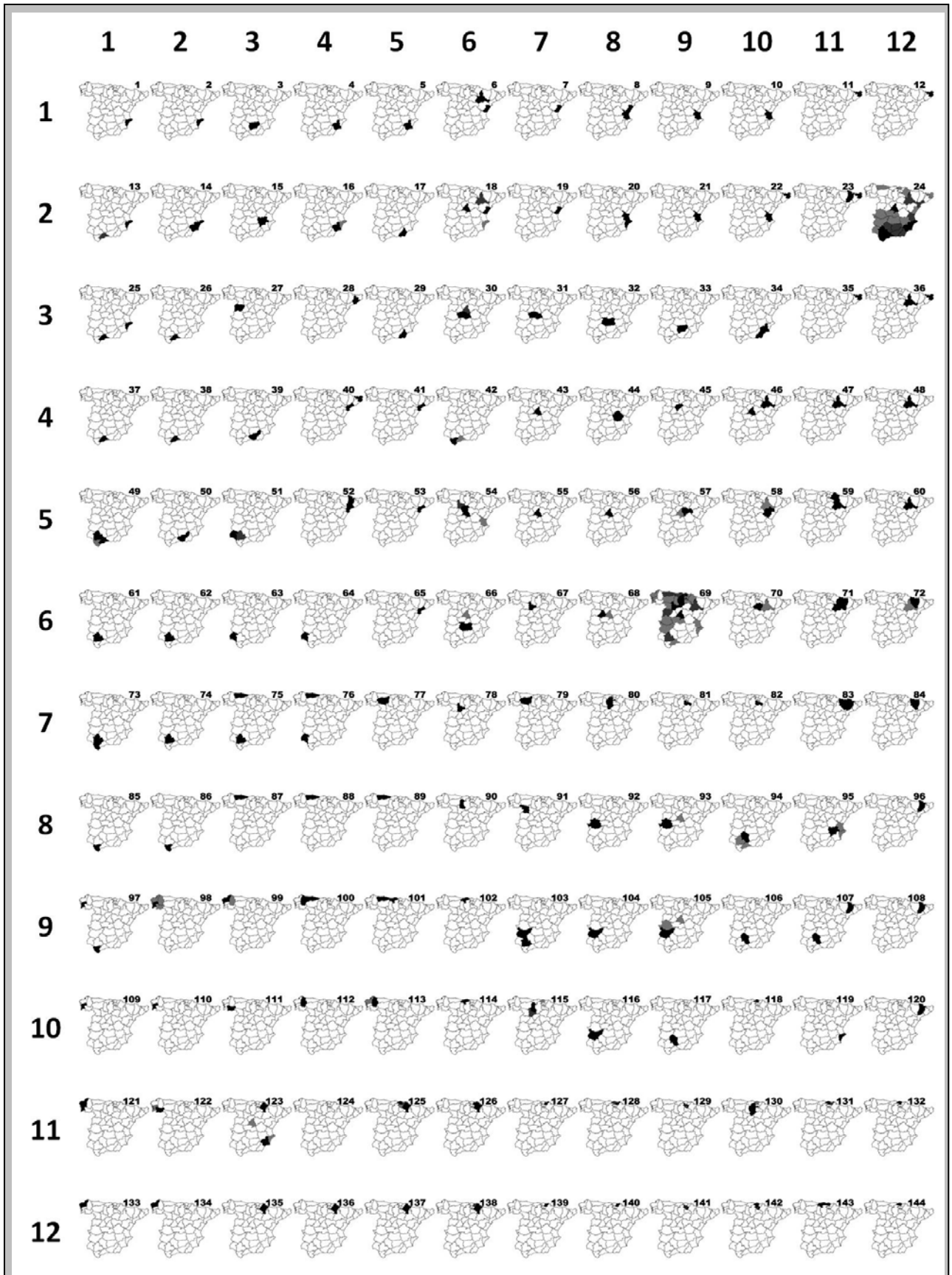


Figura 2. Matriz resultante de realizar los SOMs. Cada celda es un mapa de gradiente del territorio peninsular español que representa la distribución geográfica del correspondiente grupo de apellidos.

Resultados

SOM

Para analizar los movimientos internos de la población española, primero resulta indispensable identificar el origen de cada apellido y el patrón de dispersión que han tenido.

	España	*Holanda*	*Italia*
Nº de apellidos	33.753	9.929	77.451
Población	51.419.788	1.642.354	17.579.891
Nº de apellidos / Población	0,00065642	0,00604559	0,00440566
Nº de ap. polifiléticos	1.428	145	16.307
Población con ap. polifiléticos	37.963.365	398.880	9.488.993
% de apellidos polifiléticos	4,23	1,46%	21,05%
% Población con ap. polifiléticos	73,83%	24,29%	53,98%
Nº de ap. ambiguos	892	2.095	2.238
Población con ap. ambiguos	276.223	-	377.871
% de apellidos ambiguos	2,64%	21,10%	2,89%
% Población con ap. ambiguos	0,54%	-	2,15%
Nº de ap. monofiléticos	31.433	7.834	58.906
Población con ap. monofiléticos	16.687.001	-	7.713.027
% de apellidos monofiléticos	93,13%	78,90%	76,06%
% Población con ap. monofiléticos	32,45%	-	43,87%

Tabla 2. Tabla comparativa de los principales resultados obtenidos en España (presente trabajo), Holanda (Manni et al, 2005) e Italia (Boattini et al, 2012).

Estos orígenes se han identificado organizando los apellidos en función de su patrón de distribución mediante el uso de redes neuronales, de esta manera no es necesario identificar el origen de cada apellido y estudiar su movimiento, sino que se puede estudiar el de cada grupo. Los apellidos españoles (Figura 2, Tabla 1) se han organizado en 144 grupos de los cuales 2 se identificaron como grupos de apellidos polifiléticos, 8 de distribución dudosa, 1 quedó en blanco y de los restantes 133 grupos se identificó el origen. En otras palabras, gracias a los SOM hemos conseguido conocer el origen de 31.433 de los 33.753 (16.687.001 datos). Todas y cada una de las provincias tienen, al menos un grupo de

apellidos con origen en ella (todas las provincias están representadas en la muestra de apellidos con origen conocido).

Comparativamente (Tabla 2), España tiene pocos apellidos teniendo en cuenta el tamaño poblacional, lo que provoca que unos pocos apellidos polifiléticos representen a una mayor parte de la población.

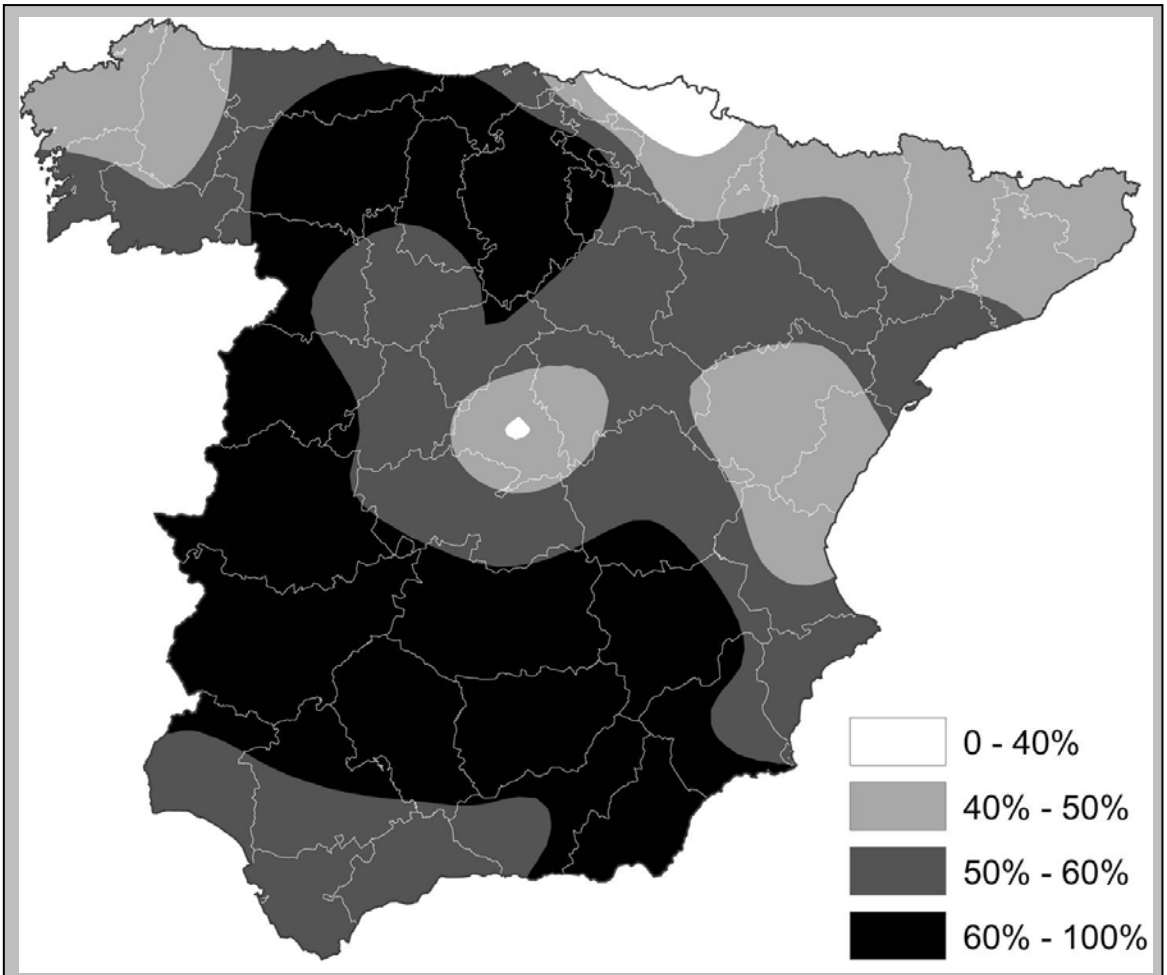


Figura 3. Porcentaje de sujetos con apellidos originales de cada provincia encontrados fuera de ella.

Caracterización de los movimientos

Con el origen de cada apellido identificado, se han podido construir matrices de migración. El análisis de estas inmigración ha permitido caracterizar los procesos migratorios de cada provincia. Así (Figura 3), sabemos que la franja Oeste y Sur de España son las zonas desde las que proporcionalmente más

población ha salido. Por el contrario, en el Norte y el Este peninsular la salida de población ha sido mucho más escasa.

Por otro lado, parece que cuatro poblaciones destacan como destinos favoritos de los movimientos poblacionales, son Vizcaya en el Norte, Madrid en el Centro (dos de los principales centros económicos del país), Valencia-Alicante en el Este y Sevilla-Málaga en el Sur. Serían los receptores de los movimientos poblacionales. Parece que puede considerarse que la franja Sur y Oeste es una fuente de población y el Noreste es el sumidero.

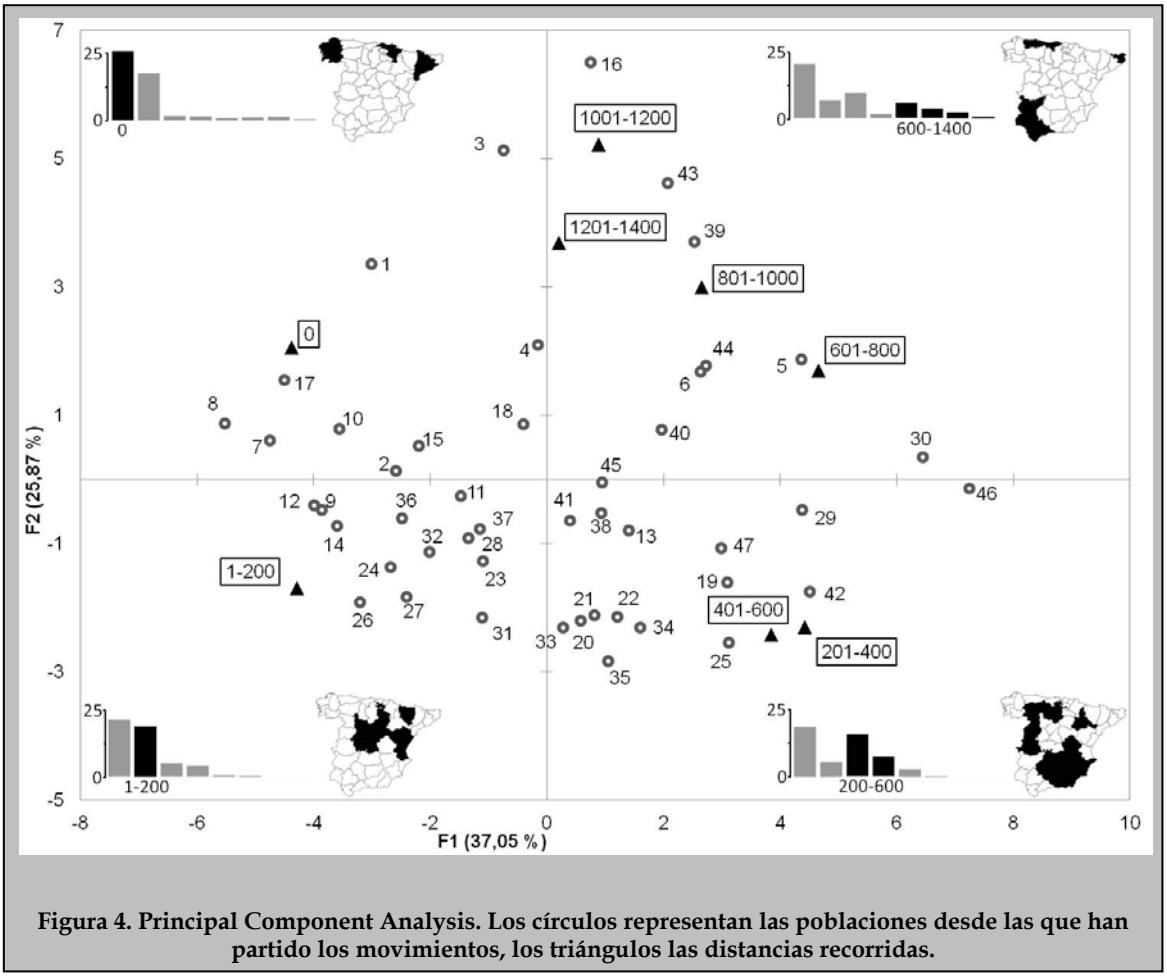


Figura 4. Principal Component Analysis. Los círculos representan las poblaciones desde las que han partido los movimientos, los triángulos las distancias recorridas.

Distancias de migración

Conociendo las características generales de los movimientos internos de la población española, se han analizado las distancias que recorren mediante análisis PCA. Lo primero que revela este análisis es que el modelo de aislamiento por distancia no es homogéneo en toda la población española.

Existen tendencias que lo deforman (Figura 4). Concretamente cuatro tendencias diferentes. La primera se ajusta perfectamente a lo esperable de un modelo de aislamiento por distancia y corresponde a poblaciones ubicadas en el Norte-Oeste, Norte y Norte-Este.

En un segundo grupo, que se localiza principalmente alrededor del centro peninsular, el modelo de aislamiento por distancia se ve deformado por una alta frecuencia de los movimientos a corta distancia (1-200).

En un tercer grupo, integrado por poblaciones del Sur y Oeste peninsular son los movimientos a media distancia (200-600) los que desvían el modelo de aislamiento por distancia.

El cuarto grupo y último grupo, correspondiente a la periferia de la península, está caracterizado por los movimientos de larga distancia (más de 600).

Dirección y sentido de los movimientos

Conociendo que existen varios tipos diferentes de movimientos (Figura 4) en función de la distancia, se han analizado por separado (Figura 5) mediante un PCA. En él se han separado las poblaciones (Figura 5, círculos) en función del destino (Figura 5, triángulos) hacia el que ha emigrado su población.

En primer lugar (Figura 5 A) se han estudiado los movimientos migratorios de corto alcance (menos de 200 km, representan el 18,67 % de todos los movimientos). Tres parecen los destinos más característicos de estos movimientos (Figura 1 y Figura 5 A, 7, 23 y 28, todos ellos situados en la mitad Norte-Oeste.

En segundo lugar (Figura 5 B) se han analizado los movimientos a media distancia (200-600 km, representan el 23,65 % del total). En esta ocasión 7 parecen los destinos más importantes, pero se pueden agrupar en dos, 4 destinos correspondientes al Sur y Este de la península y 3 situados en el centro y el Norte. Cada grupo de destinos recibe población principalmente de la mitad

peninsular en la que está situado (Los centros del Sur-Este reciben población de esa parte de España, los del Norte-Oeste la reciben de esa mitad).

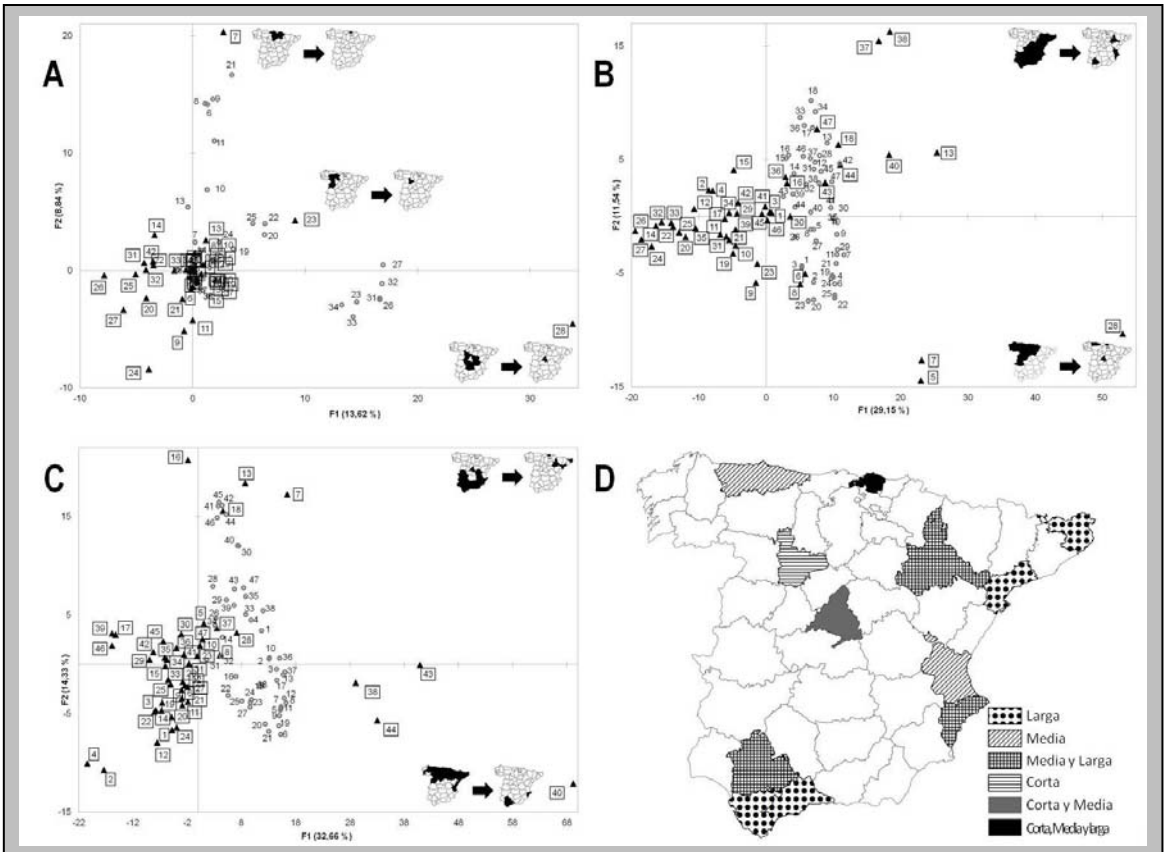


Figura 5. A), B) y C)Principal Component Analysis de los movimientos de corto alcance (menos de 200 km). Los círculos representan los orígenes de los movimientos, los triángulos los destinos (A - Movimientos de corta distancia - menos de 200 km-; B - Movimientos de media distancia - entre 200 y 600 km; C - Movimientos de más larga distancia - más de 600 km). D) Mapa resumen de los principales destinos de cada tipo de movimiento.

Por último (Figura 5 C) se han analizado los movimientos a larga distancia (por encima de 600 km, representan el 13,36 % del total). Antes de continuar, debe de tenerse en cuenta que las provincias del centro del país apenas tienen unos pocos destinos a distancias superiores a 600 km. Por lo tanto estos movimientos migratorios se realizan mayoritariamente entre la periferia del país. Destacan dos grupos de destinos, los situados el Norte-Este de España y los destinos situados en la parte Sur-Este del país.

En el caso de los movimientos de larga distancia si existe transferencia de población entre la parte Norte-Oeste de España y la parte Sur-Este.

En los tres PCAs, existe un grupo de destinos situados en dirección opuesta al resto de destinos y a todos los orígenes (Figura 5 A). En los movimientos a corta y media distancia este grupo está formado precisamente por las poblaciones que rodean a los tres destinos más importantes de esta categoría de movimientos. En cambio, en los de larga distancia, son las poblaciones ubicadas en la franja Oeste. En los tres casos representan a las poblaciones que menos movimientos inmigratorios reciben.

Centros de recepción

Atendiendo a los movimientos que reciben (Figura 5 D), los centros de recepción se pueden clasificar en tres categorías:

- Centros de importancia nacional. Son los que reciben al menos 2 tipos diferentes de movimientos. Esto significa que el alcance de su "campo gravitatorio" alcanza todo el país. Madrid es el único de estos destinos que pierde importancia en los movimientos de larga distancia. Algo razonable si consideramos que está situado en el centro de la península.
- Centros regionales. Destinos cuyo alcance es regional, cuya importancia se manifiesta sólo a media y corta distancia. Dos de estos centros están situados en el Norte-Oeste y la población que reciben es mayoritariamente de esa mitad del territorio nacional. El tercero está en la costa Sur-Este y los movimientos que recibe son precisamente de esa región.
- Centros sólo de largo alcance. Situados en la costa Sur-Este. sólo reciben movimientos de larga distancia, de menor relevancia que los anteriores.

Principales movimientos

Los principales movimientos en el interior de la población española han sido representados, concretamente los dos movimientos emigratorios mayoritarios en cada provincia (Figura 6) y los dos movimientos inmigratorios mayoritarios (Figura 7).

El análisis (Figura 6) permite detectar los mismos destinos relevantes que en el análisis de los movimientos por distancias (Figura 5) y tienen los mismos campos de atracción. De la misma manera (Figura 7) se puede observar la existencia de dos focos emisores principales; uno la zona Norte-Oeste de España, el otro en la Sur-Este y que los movimientos originados en cada uno de estos focos se quedan en su misma mitad del país.

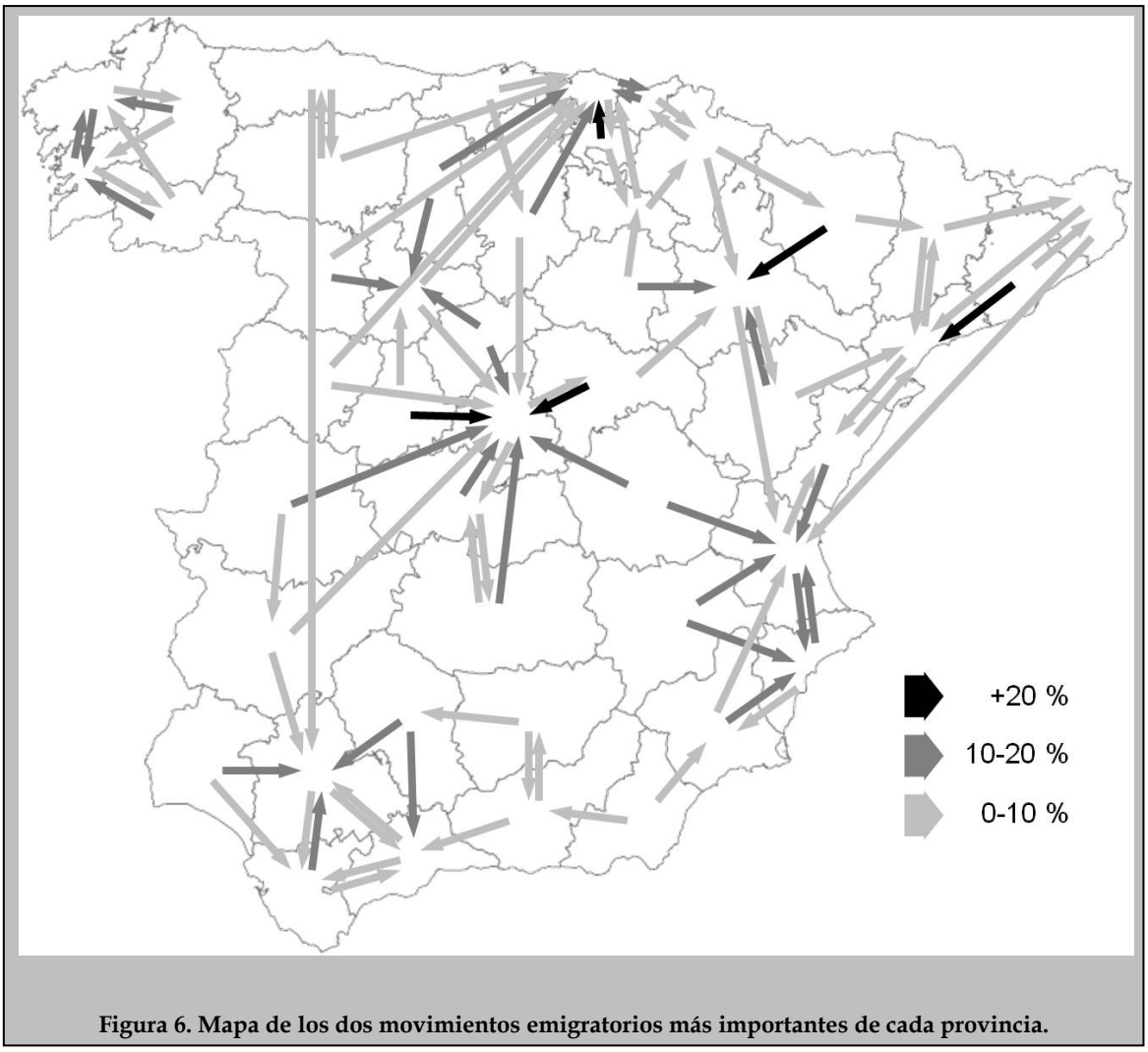


Figura 6. Mapa de los dos movimientos emigratorios más importantes de cada provincia.

Estas representaciones aportan también información sobre las "corrientes" mayoritarias (Figura 6). La mitad Norte-Oeste se mueve mayoritariamente hacia el Norte o el Centro de la península y la mitad Sur-Este se mueve mayoritariamente por la costa y, en menor medida, hacia el centro. Por otro lado, parece que estos movimientos se han verificado siguiendo lo que podría asimilarse a "corredores" poblacionales. Los dos principales son costeros uno

siguiendo la Costa Cantábrica en el Norte y el otro siguiendo la costa mediterránea en el Sur-Este (Figuras 6 y 7). Aunque menos evidente y probablemente menos relevante, se observa también otro corredor en el Oeste del país.

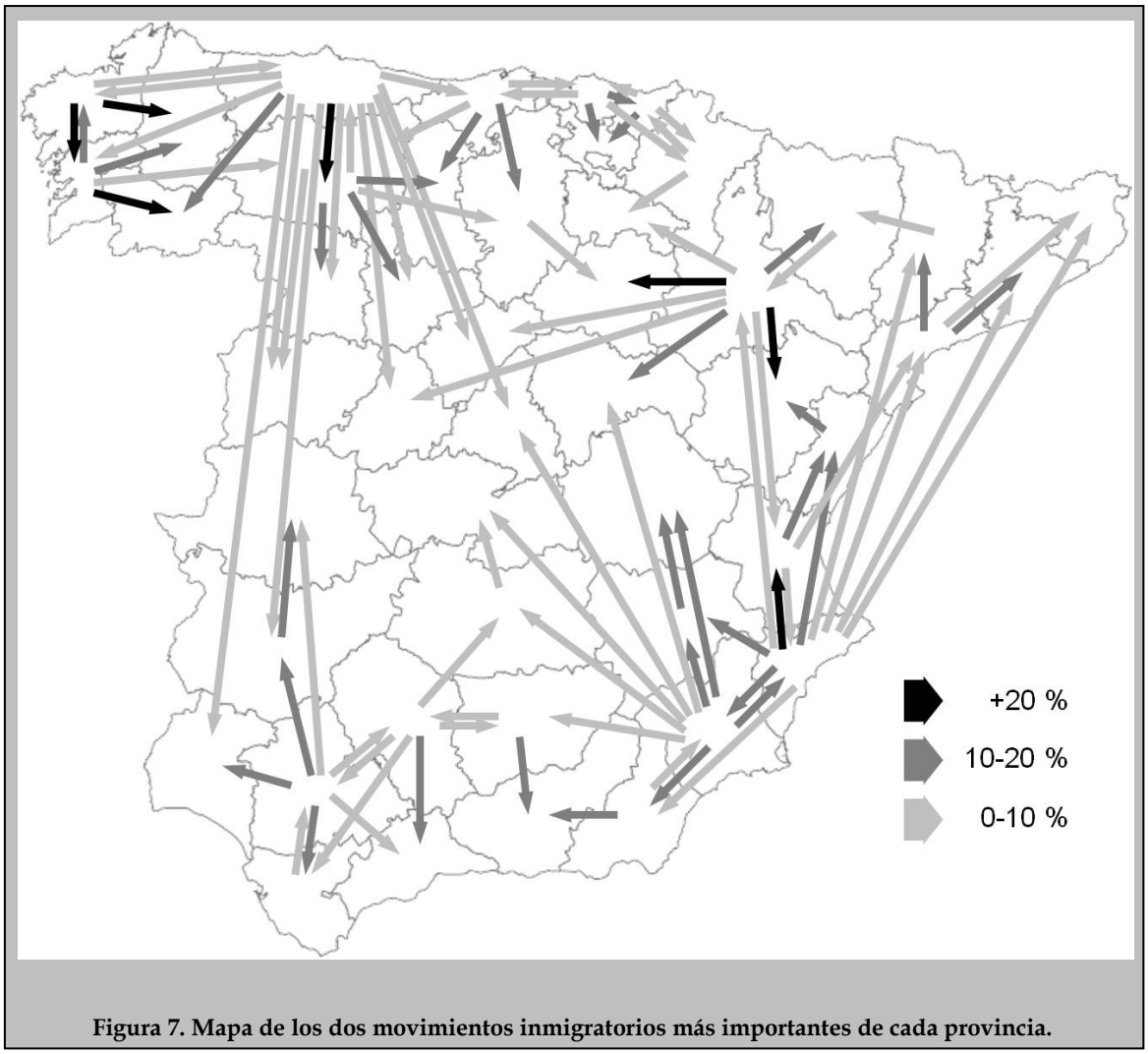


Figura 7. Mapa de los dos movimientos inmigratorios más importantes de cada provincia.

Autoctonía

Los movimientos entre las poblaciones alteran su composición, por ello, uno de los parámetros más interesantes a analizar es la autoctonía de las poblaciones o la proporción de los apellidos presentes en una población que tienen su origen en la misma (Figura 8) y su relación con los movimientos poblacionales.

En España encontramos dos zonas donde la proporción de apellidos autóctonos es especialmente alta. Ambas zonas están ubicadas en la costa. La zona más

autóctona de España resulta ser toda la costa Cantábrica en el Norte (que en casos supera el 60 % de apellidos autóctonos). La segunda zona más autóctona está en la costa mediterránea, en el Sureste.

Anteriormente se han identificado tres corredores por donde se verifican gran parte de los movimientos. Las dos zonas más autóctonas de la península, se corresponden precisamente con los dos corredores costeros. Por el contrario, el corredor del Oeste transita por una zona mucho menos autóctona.

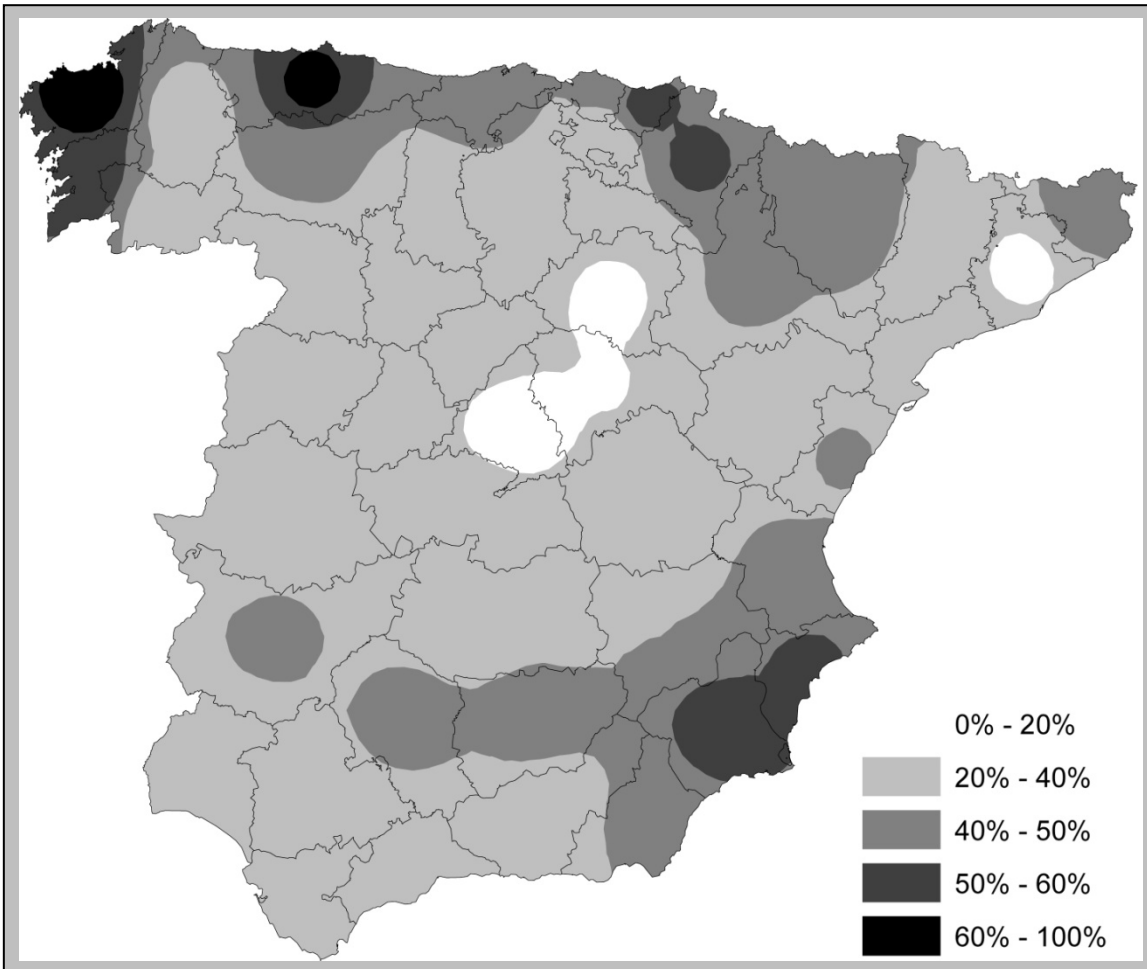


Figura 8. Porcentaje de la población que es portadora de apellidos autóctonos en cada provincia.

El resto de España muestra unos valores que oscilan entre el 20 % y el 40 % con la sola excepción de Madrid y sus inmediaciones en el centro y Barcelona en el Noreste que muestran valores que no llegan al 20%.

Historia

Obtenidos los resultados que describen la estructura de la población española, sus relaciones y movimientos internos, resulta interesante abordar la cuestión de qué profundidad histórica tienen estos resultados. A tal efecto (Figura 9) se ha comparado la población autóctona de cada provincia con el tamaño histórico de estas poblaciones.

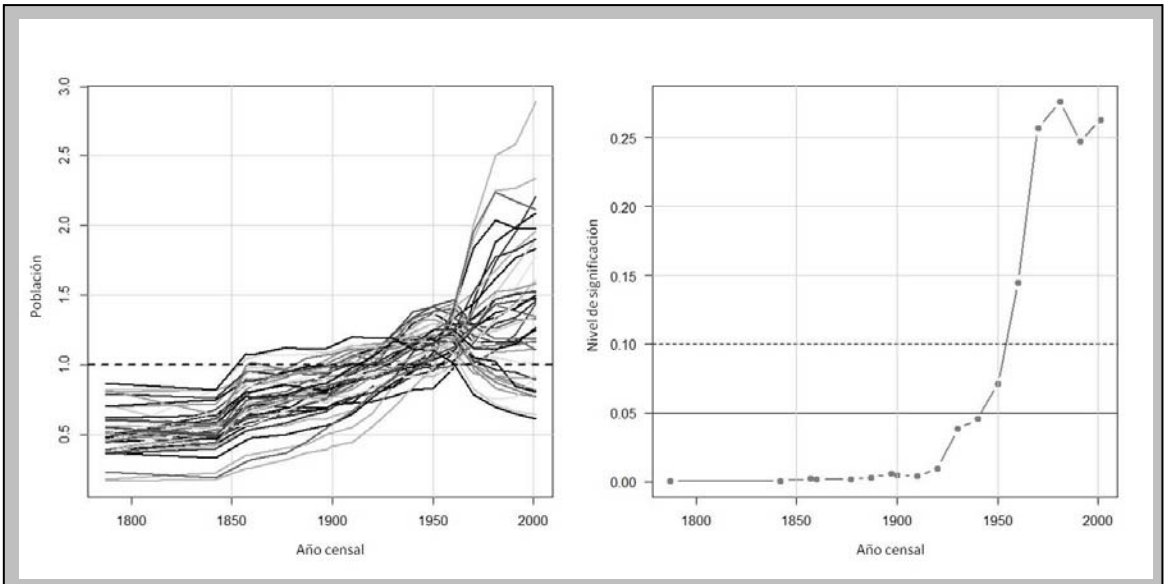


Figura 9. Izq) Gráfica de variación de la población en cada provincia española alrededor de la media (1.0) de población para todo el periodo (1787-2000). Dcha) Nivel de significación entre la población actualmente portadora de apellido autóctono y el tamaño poblacional histórico de cada provincia.

Por un lado la evolución del tamaño poblacional provincial aumenta ligeramente de manera estable hasta 1.950, año a partir del cual los tamaños poblacionales experimentan cambios bruscos. Por otro lado las correlaciones entre la población autóctona con los tamaños poblacionales históricos, se aprecian dos cosas muestran primero que la significación es más alta cuanto más antiguo es el censo y segundo que es precisamente a partir del año 1.950 cuando esta correlación deja de ser significativa.

Discusión

SOM

La metodología que se emplea en este capítulo fue propuesta, sobre el caso particular de la población holandesa (Manni et al, 2005) y testada sobre el caso de Italia más recientemente (Boattini et al, 2012). Por primera vez, esta metodología se aplica para analizar los procesos internos que han desembocado en la estructura actual de una población, en el caso que nos ocupa, la población española

La base metodológica es simple, los SOM agrupan los apellidos en función de su distribución geográfica de manera que esta se puede estudiar por grupos. Cada grupo de apellidos será tanto más frecuente cuanto más cerca se encuentre de su origen. De esta forma se puede estudiar el origen de cada grupo de apellidos. Siguiendo estos parámetros pueden distinguirse tres tipos básicos de apellidos (Manni et al, 2005):

a) Apellidos cuya distribución se extiende por el área de estudio sin obedecer a ningún patrón aparente. Son considerados apellidos polifiléticos y suelen identificar a mucha población.
b) Apellidos cuya distribución muestra un patrón ambiguo que no permite establecer un origen claro. Son apellidos ambiguos, que no pueden considerarse como monofiléticos para el resto de los casos.
c) Apellidos cuya distribución muestra un patrón claro y es posible establecer su origen. Son los apellidos monofiléticos, los que contienen información valiosa para el estudio de la población.

Los dos primeros tipos de apellidos no contienen información sobre el origen de sus portadores y por lo tanto no pueden ser utilizados en el estudio.

Resulta ilustrativo sobre algunos aspectos poblacionales la comparación de los resultados en bruto en las tres poblaciones en las que se ha utilizado esta

metodología, Holanda (Manni et al, 2005) e Italia (Boattini et al, 2012). Para empezar destaca la escasa diversidad de apellidos que tiene la población española comparativamente (España = 0,656 apellidos por cada 1.000 habitantes; Holanda = 6,046 apellidos por cada 1000 habitantes; Italia = 4.406 apellidos por cada 1.000 habitantes). Este dato nos estaría hablando de una diversidad genética más baja en España. Algo ya apuntado por estudios (Rodríguez-Larralde et al, 2003; Scapoli et al, 2007; Adams et al, 2008; Cheshire et al, 2011) y que entraría dentro de lo esperable dada su aislada situación geográfica (una península en el extremo del continente y separada de este por los Pirineos) y unas condiciones orográficas que la convierten en una amalgama de aislados.

Siguiendo con la comparación la proporción de polifiletismo (4,23 %) resulta más parecido al dato de Holanda, donde el 1,46 % de los apellidos son polifiléticos y difiere de la de Italia donde lo son el 21,05 % de los apellidos. El hecho de que haya tan pocos apellidos polifiléticos en España habla de una población muy asentada y regionalizada, o al menos con una estructura bien marcada, lo que podría ser fruto de un proceso muy antiguo de asentamiento de la población (Adams et al, 2008).

Sin embargo, estos pocos apellidos polifiléticos representa una porción enorme de la población española (el 73,83 %). Por contra en Holanda, un porcentaje parecido de apellidos representa a una porción de la población notablemente menor (24,29 %) y en Italia muchos más apellidos representa a un porcentaje de la población similar al caso español. La comparación muestra a la española como una población menos diversa que la holandesa.

En definitiva, en esta primera impresión, parece que la española es una población con baja diversidad genética, fruto del mayor aislamiento y de la mayor regionalización.

Movimientos migratorios

Una primera aproximación muestra que los movimientos migratorios no son un fenómeno homogéneo en la población española (Figura 3). Existe dos grandes zonas con comportamientos muy diferentes una zona donde la mayoría de la población original (más del 60 %) ha salido y que está ubicada en toda la franja Oeste y Sur; y otra en la que hay menos población original (menos del 50 %) que foránea y que se sitúa en la zona centro y Nor-Este.

Esta distribución geográfica de la migración evidencia que los movimientos poblacionales no han sido en absoluto homogéneos. A priori parece que una parte de España ha emitido población (Emisora) que ha recibido otra parte de España (Receptora). La naturaleza de estos movimientos, por lo tanto, habrán determinado la estructura genética de la población española y merece un análisis detenido.

Distancia, sentido y dirección de los movimientos

Por lo general se considera que las poblaciones obedecen a un modelo de aislamiento por distancia (Malecot, 1955). De hecho sabemos que esto es lo que sucede en líneas generales en el caso de España (Rodríguez-Larralde et al, 2003). Sin embargo, en España, de la misma forma que sucede en Italia (Boattini et al, 2012), los movimientos distan mucho de ser homogéneos, ni se van reduciendo homogéneamente a medida que aumenta la distancia, ni obedecen a un modelo único para toda la geografía (Figura 4). Muy al contrario, se pueden clasificar los movimientos en cuatro grupos y analizar por separado para ver cómo, cuánto y en qué modo han contribuido a la conformación de la estructura poblacional española.

- Aislamiento por distancia: Es interesante observar (Figura 4) que las poblaciones que mejor se ajustan al aislamiento por distancia coinciden con aquellas en las que está presente un idioma oficial diferente al castellano (Figura 1). Parece que si bien los idiomas no han

jugado un papel relevante en la estructura global de la población española, si podrían haber jugado un papel secundario a menor nivel geográfico.

- Movimientos de corta distancia: Representan un porcentaje de la población muy bajo, el 18,87 % de todos los movimientos. Esto sumado a que representan a movimientos de poco alcance hace que tengan una importancia menor en la estructura de la población. Son más representativos en las poblaciones situadas en torno a centros importantes (Figura 5 A y D) en la mitad Norte-Oeste. La atracción de estos centros es tan importante que ha alterado el aislamiento por distancia.
- Movimientos de media distancia: Son los más importantes (representan el 23,65 % de los movimientos). La mayoría de las provincias españolas están separadas por este rango de distancias, por lo que este grupo es el más representativo de las relaciones interpoblacionales. Muestra a la población española partida en dos mitades. La parte Norte-Oeste y la Sur-Este no se relacionan entre ellas, los movimientos se producen desde las provincias hacia centros situados en esa misma mitad.
- Movimientos de larga distancia: Son los movimientos menos representativos (el 13,36 %) no en vano sólo las poblaciones de la periferia están separadas por tanta distancia. Es precisamente por la periferia por donde se producen estos movimientos.

Principales Movimientos

Un análisis detenido de los principales movimientos nos permite observar cómo han sido las relaciones entre las poblaciones españolas, representando los dos principales destinos (Figura 6) y los dos principales orígenes (Figura 7) de los movimientos de cada provincia.

Tanto los destinos (Figura 6) como los orígenes (Figura 7) muestran dos arcos migratorios principales, ambos recorren la costa. El primero siguiendo el mediterráneo, el segundo siguiendo el Cantábrico. Parece que estos dos arcos son los que han vertebrado la población española dividiéndola en dos mitades que alcanzarían hasta donde llegan sus áreas de influencia. Secundariamente se observan la presencia de un tercer arco (menos relevante) en el Oeste (Figuras 6 y 7) que se corresponde a la perfección con la Ruta de la Plata, una antigua vía de comunicación que retomaron los romanos y sigue representando una importante vía Norte-Sur en la actualidad (Martínez, 2003).

Parece que la existencia de estos grandes movimientos poblacionales es la que ha desembocado en la estructura poblacional descrita anteriormente.

Autoctonía

El grado de autoctonía varía considerablemente de unas zonas a otras como resultado de la influencia de factores geográficos o históricos que dan como resultado patrones diferentes de migración (Manni et al, 2005). En el caso de España parece que la autoctonía se concentra en las costas coincidiendo con los arcos Mediterráneo y Cantábrico.

Parece que existen dos tipos de corredores. A lo largo de los dos corredores costeros la autoctonía es muy elevada de manera semejante a lo que sucede en la Toscana italiana (Boattini et al, 2012), mientras que el corredor del Oeste es muy poco autóctono. Una más que posible explicación de este fenómeno es que cada corredor costero articula su mitad de la población española. La población española está dividida en dos mitades y cada una se extiende a partir de un arco costero, por ellos se producen movimientos dentro de una misma parte de la población (mitad Norte-Oeste, mitad Sur-Oeste) mientras que el corredor del Oeste supone una ruta de intercambio genético entre estas dos poblaciones diferenciadas y de ahí que tenga más población alóctona.

Historia

Una vez conocida la estructura de la población y cómo se ha conformado, la pregunta obvia es cuando se ha producido este proceso. En primer lugar los resultados son consistentes con lo observado por el Instituto Geográfico Nacional y el Instituto Nacional de Estadística (www.ign.es, www.ine.es) tanto más consistentes cuanto más antiguos son los movimientos migratorios registrados (los más antiguos corresponden a la década 1.960-1.970). Algo esperable de la metodología de la isonimia, que refleja el resultado de un proceso histórico.

En idéntica dirección apunta la comparación entre apellidos autóctonos y los censos históricos de las provincias (Figura 9). El hecho de que el número de portadores de apellidos autóctonos correlacione mejor con el tamaño de la población cuanto más antiguo sea el censo empleado, ratifica que los apellidos autóctonos son un fiel reflejo de la población original de cada provincia y es un indicio de la precisión en la identificación del origen de cada apellido.

Hasta 1.950 (Figura 9) esta correlación es significativa. En torno a ese año comienza la emigración rural en España y la población pierde la estabilidad (Fuster y Colantonio, 2002). De hecho, si se observa la evolución de la población española en cada provincia (Figura 9) se ve claramente que todas las provincias mantienen una población estable en suave crecimiento hasta el año 1.950. A partir de ese año y de manera brusca, unas provincias comienzan a perder población a favor de otras y los tamaños poblacionales cambian bruscamente, la estabilidad poblacional ha desaparecido.

Este fenómeno hace pensar de nuevo en los movimientos de larga distancia como un fenómeno reciente y es coherente con la impresión de que la población española está muy conservada (Adams et al, 2008). En este escenario parece que las relaciones intrapoblacionales españolas han sido de un alcance limitado tanto en intensidad como en distancia y que se han verificado hasta épocas muy recientes en el interior de unas zonas claramente delimitadas por

condicionantes geográficos, lo que explica que la población española se muestre claramente dividida en dos partes diferenciadas. Tradicionalmente los movimientos poblacionales se han verificado en el interior de estas zonas y apenas ha existido intercambio entre ellas hasta épocas muy recientes.

Conclusión

En el interior de la población española se han producido varios tipos de movimientos. Los de corta y media distancia han sido los más frecuentes y los más determinantes en la actual estructura. Dada la gran estabilidad de la población hasta épocas recientes (1950) y la menor importancia que tienen, los movimientos de larga distancia parecen un fenómeno más reciente, con una contribución menor a la estructura poblacional.

Los movimientos internos en el seno de la población española son el factor principal que ha desembocado en la actual estructura. Esta movilidad se ha verificado principalmente en el interior de dos regiones geográficamente diferenciadas. En la Nor-Oeste estos movimientos se han producido a lo largo del arco Cantábrico y hasta donde ha llegado su influencia. Simétricamente, en la Sur-Este han seguido el arco Cantábrico y han llegado hasta donde llega su área de influencia.

El intercambio entre ambas áreas ha sido relativamente escaso y se ha verificado principalmente por el Oeste siguiendo un corredor con el recorrido de la antigua "ruta de la plata".

Metodológicamente, una vez más el trabajo con la isonimia demuestra que; siempre que se dispongas las medidas de control necesarias, es una aproximación potente fiable y barata a la estructura genética de una población. La eliminación de los apellidos polifiléticos permite una aproximación más segura. Pero, yendo más allá, en el trabajo que nos ocupa hemos identificado el origen de cada grupo de apellidos y trabajado con ellos. La coherencia con los resultados preexistentes, con análisis anteriores de los movimientos migratorios y la correlación entre la autoctonía con los censos más antiguos demuestran la precisión (teniendo en cuenta el nivel geográfico elegido) de la técnica a la hora de identificar los orígenes de los apellidos.

Bibliografía

Adams S M, Bosch E , Balaresque P L, Ballereau S J, Lee A C, Arroyo E, Lopez-Parra A M, Aler M, Grifo M S, Brion M, Carracedo A, Lavinha J, Martinez-Jarreta B, Quintana-Murci L, Picornell A, Ramon M, Skorecki K, Behar D M, Calafell F, Jobling M A. 2008. The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. Am J Hum Genet 83: 725-736.

Alvarez,L.; Santos,C.; Ramos,A.; Pratdesaba,R.; Francalacci,P.; Aluja,M.P. Mitochondrial DNA patterns in the Iberian Northern plateau: Population dynamics and substructure of the Zamora province. Am.J.Phys.Anthropol., 2010.

Balanovskaia,E.V.; Romanov,A.G.; Balanovskii,O.P. Namesakes or relatives? Approaches to investigating the relationship between Y chromosomal haplogroups and surnames. Mol.Biol.(Mosk), 2011, 45, 3, 473-485, Russia (Federation).

Barrai,I.; Rodríguez-Larralde,A.; Mamolini,E.; Manni,F.; Scapoli,C. Elements of the surname structure of Austria. Ann.Hum.Biol., 2000, 27, 6, 607-622, ENGLAND.

Barrai,I.; Rodríguez-Larralde,A.; Manni,F.; Ruggiero,V.; Tartari,D.; Scapoli,C.. Isolation by language and distance in Belgium. Ann.Hum.Genet., 2004, 68, Pt 1, 1-16, England.

Boattini A, Griso C, Pettener D. 2011. Are ethnic minorities synonymous for genetic isolates? Comparing Walser and Romance populations in the Upper Lys Valley (Western Alps). J Anthropol Sci 89: 161-173.

Boattini A, Lisa A, Fiorani O, Zei G, Pettener D, Manni F. 2012. General Method to Unravel Ancient Population Structures through Surnames, Final Validation on Italian Data. Hum Biol 84: 235-270.

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

Boattini, A.; Blanco-Villegas, M.J.; Pettener,D. Genetic structure of La Cabrera, Spain, from surnames and migration matrices. Hum.Biol., 2007, 79, 6, 649-666, United States.

Bodmer W F, Cavalli-Sforza L L. 1968. A migration matrix model for the study of random genetic drift. Genetics 59: 565-592.

Cheshire J, Mateos P, Longley PA. 2011. Delineating Europe's Cultural Regions: Population Structure and Surname Clustering. Hum Biol 83: 573-598.

Colantonio SE, Lasker GW, Kaplan BA, Fuster V. 2003. Use of surname models in human population biology: a review of recent developments. Hum Biol 75:785-807.

Dipierri J, Rodríguez-Larralde A, Alfaro E, Scapoli C, Mamolini E, Salvatorelli G, Caramori G, De Lorenzi S, Sandri M, Carrieri A, Barrai I. 2011. A Study of the Population of Paraguay through Isonymy. Ann Hum Genet 75: 678-687.

Dipierri,J.E.; Alfaro,E.L.; Scapoli,C.; Mamolini,E.; Rodríguez-Larralde,A.; Barrai,I. Surnames in Argentina: a population study through isonymy. Am.J.Phys.Anthropol., 2005, 128, 1, 199-209, Wiley-Liss, Inc, United States.

Esparza,M.; García-Moro, C.; Hernández, M. Genetic relationships between parishes in the Ebro delta region (Spain) as estimated by migration matrix and surnames. Hum.Biol., 2006, 78, 6, 647-662, United States.

Faure R, Ribes MA, García A. 2001. Diccionario de apellidos españoles. Madrid: Espasa-Calpe.

Fuster V, Colantonio SE. 2002. Consanguinity in Spain: socioeconomic, demographic, and geographic influences. Hum.Biol 74: 301-315.

Gagnon,A.; Heyer,E. Intergenerational correlation of effective family size in early Quebec (Canada). Am.J.Hum.Biol., 2001, 13, 5, 645-659, United States.

García P. 2007. Lenguas y dialectos de España. Madrid: Arco Libros.

Goebl H. 2010. La dialectometrización del ALPI: Rápida presentación de los resultados. 26th CILFR. Valencia.

King,T. E.; Ballereau, S. J.; Schurer, K. E.; Jobling, M. A. Genetic signatures of coancestry within surnames. Curr.Biol., 2006, 16, 4, 384-388, England.

King,T. E.; Jobling, M. A. Founders, drift and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. Mol.Biol.Evol., 2009.

King,T. E.; Jobling, M. A. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. Trends Genet., 2009, 25, 8, 351-360, England.

Kohonen T. 1982. Self-organized formation of topologically correct feature maps. Biol Cyber 43: 59-69.

Kohonen T. 1984. Self-organization and associative memory. Berlin: Springer.

Lisa, A.; De Silvestri, A.; Mascaretti, L.; Degiuli, A.; Guglielmino, C. R. HLA genes and surnames show a similar genetic structure in Lombardy: does this reflect part of the history of the region? Am.J.Hum.Biol., 2007, 19, 3, 311-318, Wiley-Liss, Inc, United States.

Malecot G. 1955. Decrease of relationship with distance. Cold Spring Harbor Syrup. Quant Biol 20: 52-53.

Manni F, Toupance B, Sabbagh A, Heyer E. 2005. New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Y-chromosome sampling. Am J Phys Anthropol 126: 214-228.

Martínez E. 2003. Atlas histórico de España. Madrid: Istmo.

Pettener D, Pastor S, Tarazona-Santos E. 1998. Surnames and genetic structure of a high-altitude quechua community from the Ichu river valley peruvian central Andes 1825-1914. Hum Biol 70: 865-87.

Rodríguez-Díaz R, Blanco-Villegas MJ. 2010. Genetic structure of a rural region in Spain: distribution of surnames and gene flow. Hum Biol 82: 301-314.

Rodríguez-Díaz R, Manni F, Blanco-Villegas M J. XXXX. Usefulness of isonymy methods for describing the genetic structure of large populations. The case of Spain.XXXX

Rodríguez-Larralde A, Gonzales-Martin A, Scapoli C, Barrai I. 2003. The names of Spain: a study of the isonymy structure of Spain. Am J Phys Anthropol 121: 280-292.

Rodríguez-Larralde,A.; Morales,J.; Barrai,I. Surname frequency and the isonymy structure of Venezuela. Am.J.Hum.Biol., 2000, 12, 3, 352-362.

Rodríguez-Larralde,A.; Scapoli,C.; Beretta,M.; Nesti,C.; Mamolini,E.; Barrai,I. Isonymy and the genetic structure of Switzerland. II. Isolation by distance. Ann.Hum.Biol., 1998, 25, 6, 533-540, ENGLAND.

Rodríguez-Larralde,Alvaro; Dipierri,José; Gomez,Emma Alfaro; Scapoli,Chiara; Mamolini,Elisabetta; Salvatorelli,Germano; De Lorenzi,Sonia; Carrieri,Alberto; Barrai,Italo. Surnames in Bolivia: A study of the population of Bolivia through isonymy. Am.J.Phys.Anthropol., 2011, 144, 2, 177-184, Wiley Subscription Services, Inc., A Wiley Company.

Rogers, K. B. (1991). *The relationship of grouping practices to the education of the gifted and talented learner* (RBDM 9102). Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.

Scapoli C, Mamolini E, Carrieri A, Rodríguez-Larralde A, Barrai I. 2007. Surnames in Western Europe: a comparison of the subcontinental populations through isonymy. Theor Popul Biol 71: 37-48.

Solís JA. 2002. El gran libro de los apellidos. La Coruña: El arca de papel.

Sykes, B.; Irven, C. Surnames and the Y Chromosome. Am. J. Hum. Genet., 2000, 66, 4, 1417-1419.

Wehrens R, Buydens LMC. 2007. Self- and Super-organising Maps in R: the Kohonen package. J Stat Softw, 21.

3. DIVERSIDAD GENÉTICA-BIODIVERSIDAD

DIVERSIDAD GENÉTICA-BIODIVERSIDAD

Resumen

Introducción: La relación entre diversidad genética y diversidad biológica es un tema objeto de reciente estudio tanto en el ámbito de la biogeografía como en el de la antropología. En este último contexto el estudio de su imbricación puede resultar especialmente fructífero en lo que a conocer los procesos que han determinado las poblaciones se refiere.

Material y métodos: Para estudiar la población humana se han utilizado los apellidos contenidos en la base de datos del padrón continuo del INE de 2008, que incluye unos 57 millones de datos. Para la fauna de vertebrados se ha empleado la base de datos del Inventario Español de Especies Terrestres que engloba la información de los diferentes Atlas y Libros Rojos.

En un primer paso, se han calculado las distancias genéticas entre las provincias españolas y la diversidad genética en base a su composición de apellidos. Por otro lado, se ha calculado las distancias entre provincias y su diversidad específica en base a las especies animales.

Resultados: La estructura genética de la población española y la estructura específica presentan resultados muy similares. Ambas estructuras muestran el territorio peninsular español dividido en dos partes principales (Norte-Oeste y Sur-Este). La comparación entre diversidad genética y diversidad específica muestra unas similitudes más limitadas. Solo en el Este peninsular correlacionan.

Introducción

Tradicionalmente, la estructura actual de las poblaciones humanas se ha considerado como el resultado de la influencia de factores lingüísticos (Barrai et al., 2002, 2004; Manni et al., 2008), étnicos (Fiorini et al., 2007), socio económicos (Caravello y Tasso, 1999, 2002) o físicos (Esparza et al., 2006; Boattini et al., 2007; Boattini et al., 2006).

Uno de los campos que más interés ha suscitado dentro del estudio biológico de poblaciones, es su estructura genética. Pero este campo no resulta sólo interesante en la biología de poblaciones humanas, sino en toda la biogeografía (Baselga, 2013). Recientemente y gracias a la disponibilidad de grandes bases de datos y a los medios para analizarlas (Vellend y Geber, 2005; Hubbell, 2001), la biogeografía está fijando la vista en la relación que existe entre los procesos experimentados por la diversidad genética y los experimentados por la biodiversidad (Baselga, 2013).

A nivel de biodiversidad, entendida como un concepto jerárquico que abarca multitud de niveles diferentes biológicos, desde comunidades a especies o genes (Noss, 1990), la influencia de los factores geográficos y ambientales es algo conocido (Lomolino et al., 2010). Por lo general se asume que esta misma influencia se produce a nivel genético. Desde un punto de vista racional, parece obvio por lo tanto que, para la mayoría de las especies, aquellos factores físicos que condicionan la biodiversidad tendrán también una influencia trascendental en la diversidad genética. Sin embargo esta asociación ni es tan inmediata como podría parecer ni ha sido sujeto de estudio hasta hace muy poco (Vellend y Geber, 2005; Hubbell, 2001).

Estos acercamientos, en lo que respecta a las poblaciones modernas de la especie humana se han quedado en estudios someros de la relación existente entre biodiversidad y diversidad cultural o lingüística (Sutherland, 2003; Stepp et al. 2004), sin llegar en ninguno de los casos a analizar las implicaciones

genéticas para el hombre pese a la estrecha relación existente (Cavalli-Sforza, 2000).

Con el propósito de analizar esta relación entre biodiversidad y diversidad genética humana y los factores que determinan ambas, pretendemos comparar la estructura espacial de la diversidad de fauna vertebrada del territorio español (en forma de diversidad de especies animales) con la estructura genética de la población española.

Anteriormente al analizar la estructura genética de la población española los resultados mostraron una fuerte coincidencia con la conformación geográfica de la península Ibérica. La población española se segregaba en dos grupos principales articulados uno por la costa cantábrica y el otro por la costa mediterránea. Posteriormente se han estudiado los movimientos poblacionales interiores. El análisis reveló que, efectivamente, los movimientos dentro de la población española están condicionados por los factores geográficos.

Todos estos análisis muestran una fuerte vinculación de la estructura genética de la población española con los factores físicos. Merece la pena detenerse en esta relación y valorar su profundidad, si la relación es exclusiva con factores de tipo físico o si esta relación se extiende a factores biogeográficos. El análisis de esta posible relación es el objeto del presente trabajo. Para ello se ha analizado la diversidad animal del territorio peninsular español. Esta diversidad está completamente marcada por los factores geográfico-ambientales. La relación entre ambas estructuras nos permitirá comprobar la si estos factores han determinado la población humana y en qué medida ha sido así.

Materiales y métodos

Zona de Estudio

La Península Ibérica se emplaza en el Suroeste del continente europeo, a caballo entre el Océano Atlántico y el Mar Mediterráneo, y separada de África por el Estrecho de Gibraltar. Su ubicación le dota de un carácter de lugar de convergencia de diversas influencias, constituyendo una encrucijada tanto desde el punto de vista biogeográfico como humano. España es el país que ocupa la mayor parte de la Península Ibérica con una superficie de 504,645 km2 (de los cuales 492,175 km2 corresponden a la Península). Por tierra tiene fronteras con Portugal por el Oeste y Francia por el Noreste.

Dentro del espacio geográfico europeo, su posición periférica y el cierto grado de aislamiento producido por su carácter peninsular y la barrera geográfica de los Pirineos, ha hecho que históricamente España esté relativamente alejada de los centros de toma de decisiones central europeos. Sin embargo forma parte del Arco Atlántico que iría desde Portugal hasta el Sur del Reino Unido, y del Arco Mediterráneo, que conectaría el Mediterráneo español con el Sur de Francia e Italia.

Poblaciones humanas

La población española es de 47 millones de habitantes (43 millones en la Península Ibérica), organizados administrativamente en 17 comunidades autónomas y 52 provincias, (15 y 47 de ellas en la Península). Su distribución geográfica es muy desigual, la mayor densidad se concentra en las zonas costeras dejando el centro (excepto Madrid, la capital administrativa) mucho menos poblado.

Por último, esta población habla cuatro idiomas diferentes; el castellano que es el lenguaje oficial en todo el territorio; el catalán, hablado en la zona costera del Este peninsular; el gallego, en la esquina Noroeste; y el euskera, en el Norte.

Biogeografía y biodiversidad animal.

El emplazamiento geográfico de la Península Ibérica, junto a la variedad de climas, relieve y geología, explican una complejidad biogeográfica mayor de la esperada de acuerdo a su superficie.

Tanto Atlántico como Mediterráneo ejercen su acción reguladora sobre clima de la Península Ibérica pero mientras el Atlántico aporta humedad y temperaturas moderadas (≈ 14° C), la costa mediterránea se caracteriza por la aridez estival y temperaturas más elevadas (15-18° C). Al alejarnos de la costa aumenta la continentalidad con temperaturas más extremas y una mayor amplitud térmica. Así, en el interior las temperaturas medias anuales van desde los 10° C en el Norte hasta los más de 16° C en el Sur. En las zonas montañosas las medias bajan por debajo de los 7.5° C. Las precipitaciones también muestran un gradiente desde el Norte-Norte Oeste donde se superan los 800l/m^2 hasta menos de 300l/m^2 en zonas del SE peninsular. En las zonas montañosas del interior también se superan los 800 l/m^2.

Presenta un promedio de 660 m de altitud sobre el nivel del mar y de marcado carácter montañoso en comparación al resto de países europeos. El centro peninsular lo ocupa una gran meseta central que engloba las cuencas del Duero y del Tajo y en torno a ella se articulan varios sistemas montañosos (Cordillera Cantábrica al Norte, Sistema Ibérico al Este y Sierra Morena al Sur) y las depresiones del Ebro y del Guadalquivir. En el centro de la meseta se encontraría el Sistema Central. En el extremo nororiental destacan los Pirineos. En cuanto a la geología, hay una clara diferencia entre el Oeste silicio (pizarras, cuarcitas y granitos), y el Este calcáreo (predominio de calizas). Los materiales arcillosos serían los dominantes en los valles de los grandes ríos.

En base a estos factores la Península Ibérica tradicionalmente se ha dividido en tres grandes regiones biogeográficas: la eurosiberiana, la mediterránea y la alpina. La región eurosiberiana se extendería por la franja Norte limitando la Cordillera Cantábrica una mayor extensión hacia el Sur. Se caracteriza por un

clima oceánico con veranos suaves y húmedos, favoreciéndose el desarrollo de vegetación con predominio de bosques caducifolios. En la mediterránea los veranos son calurosos y secos por lo que la vegetación se encuentra sometida a estrés hídrico. Predominan los perennifolios con adaptaciones a la aridez estival. La región alpina se limita a las zonas de mayor altitud con coníferas en los pisos bioclimáticos más bajos y matorral y pastizal de alta montaña en las áreas de mayor altitud.

La configuración de la Península y el relieve influyen también en esa diversidad. El relieve extremadamente contrastado permite la constitución de gran número de hábitats de características muy diferentes que permiten la vida a muy variados tipos de vegetación. En la Península se encuentran representadas prácticamente todas las principales formas de relieve: la Meseta, importantes cordilleras, depresiones, llanuras costeras, etc. A la variación climática y morfológica deben añadirse los contrastes líticos y edáficos que contribuyen a hacer más complejo el paisaje vegetal.

Origen de los datos

Apellidos

Para la estimación de la estructura genética de la población española se han utilizado los apellidos. El proceso de herencia de los apellidos emula al proceso de herencia genética, de tal manera que los apellidos pueden asimilarse a formas de un locus de un cromosoma (Chakraborty et al., 1990), por lo que se puede considerar a los apellidos como un estimador fiable de los linajes paternos y que dos poblaciones estarán tanto más emparentadas cuanto mayor es el número de apellidos en común y cuanto más parecida es su distribución (Lasker y Kaplan, 1985). Esta característica unida a su relativamente fácil disponibilidad, los han convertido en un medio muy extendido y útil en el estudio de poblaciones humanas. Los estudios más recientes sustentan la fiabilidad del método (Sykes e Irven, 2000; Gagnon y Heyer, 2001; Esparza et al., 2006; King et al., 2006; Boattini et al., 2007; Lisa et al., 2007; King y Jobling,

2009A; King y Jobling, 2009B; Álvarez, 2010; Rodríguez-Díaz y Blanco-Villegas, 2010; Balanovskaia et al., 2011), haciendo de éste una alternativa rápida, barata y fiable para el estudio de las poblaciones humanas.

En concreto en nuestro caso se ha utilizado la base de datos del Instituto Nacional de Estadistica (INE), procedente del padrón continuo de 2008. En la base de datos están incluidos todos los apellidos de cada municipio siempre que aparezcan un mínimo de 5 veces. La base de datos inicial incluía 56976706 registros, correspondientes a 87148 apellidos diferentes.

En el sistema de transmisión de apellidos español, se heredan dos apellidos. Cada sujeto hereda el primer apellido del padre (será su primer apellido) y el primero de la madre (será su segundo apellido). Dada la disponibilidad de los dos apellidos de cada individuo, la base de datos se construyó utilizando tanto el primero como el segundo apellido algo que según diversos autores, multiplica la cantidad de información y contribuye a la robustez del análisis (Pettener et al., 1998; Colantonio et al., 2003; Dipierri et al., 2011).

La base de datos de los apellidos se revisó y depuró hasta prepararla adecuadamente para su uso. Una vez dispuesta la base de datos, los apellidos han sido agrupados y organizados utilizando los mapas autoorganizados (Kohonen, 1982; 1984). El propósito de este procedimiento es agrupar los apellidos en función de su distribución y analizarla para identificar sus orígenes, en base a eso se pueden eliminar los apellidos polifiléticos y así hacer de los apellidos unos marcadores más fiables (Manni et al., 2005; Boattini et al., 2011).

Fauna

Las distribuciones de la fauna de vertebrados se han obtenido de la base de datos del Inventario Español de Especies Terrestres promovido por el Ministerio Español de Agricultura, Alimentación y Medio Ambiente (http://www.magrama.gob.es/es/) y que engloba la información de los

diferentes Atlas y Libros Rojos. Para cada especie se han compilado datos de presencia procedentes de citas de voluntarios, publicaciones, revisión de colecciones, muestreos y censos en campo, así como de informes inéditos que incluyen observaciones directas, capturas, rastros, atropellos, o datos de caza y pesca. Esta información se actualiza periódicamente con los nuevos datos aportados por los programas de seguimiento de cada grupo taxonómico. Las distribuciones espaciales resultantes se muestran en una cuadrícula UTM de 10x10 km que cubre todo el territorio español. Se han considerado solo los vertebrados terrestres puesto que sus distribuciones se conocen mejor y presentan un menor sesgo derivado de diferentes intensidades de muestreo entre zonas. Así, en los análisis de los patrones de distribución espacial de la biodiversidad peninsular española se incluyeron 27 especies de anfibios, 256 de aves nidificantes, 87 de mamíferos, 38 de peces de agua dulce, y 42 de reptiles; un total de 450 especies. Se han eliminado de esta base de datos las especies exóticas.

Estructura genética - Diversidad específica

Con el propósito de comparar la estructura genética de la población española y la estructura de la biodiversidad, se han calculado las diferencias entre las 47 provincias peninsulares españolas en función de su composición de apellidos por un lado y en función de su composición específica por el otro.

Genéticamente hablando, se puede considerar que dos poblaciones diferentes serán tanto más próximas cuanto mayor sea la probabilidad de que dos alelos para un mismo locus en dichas poblaciones sean idénticos por descendencia (Morton, 1971). El proceso de herencia de los apellidos emula al del cromosoma Y, por lo que la composición de apellidos de una población puede considerarse una aproximación a la estructura genética de esa población (Lasker y Kaplan, 1985; Chakraborty y Schwartz, 1990). De manera que, para el análisis de los apellidos españoles, se ha empleado el coeficiente de Hedrick (Hedrick, 1971; Weiss, 1980) que viene dado por la expresión:

$$H_{ij} = \frac{\Sigma_k(p_{ki}p_{kj})}{1/2\Sigma_k(p_{ki}^2 + p_{kj}^2)}$$

Donde pki y pkj son las frecuencias del apellido o la procedencia "k" en las poblaciones "i" y "j" respectivamente.

Para realizar estos cálculos, y con el objetivo de no subestimar y distorsionar las distancias entre provincias y emplear solo apellidos que podamos considerar marcadores fiables (Manni et al., 2005), se ha utilizado la base de datos de apellidos con origen conocido calculada anteriormente.

En cuanto a la diversidad específica, se calculó una matriz de distancias euclídeas entre provincias en función de su composición de especies animales.

El resultado de ambos procedimientos son sendas matrices cuadradas de distancias con un número de filas y columnas igual al número de provincias incluidas en el estudio (47x47). En estas matrices, cada casilla indica la similitud o la distancia genética entre las poblaciones de la fila y columna correspondientes.

Relación entre distancias

Para completar la comparación entre las distancias, se ha construido también una matriz cuadrada que recoge la distancia en línea recta que separa a las capitales de provincia usando el paquete "maptools" de R-Project (Nicholas, 2012). El hecho de seleccionar las capitales de provincia como puntos de referencia para calcular las distancias responde a un doble criterio. Por un lado, desde un punto de vista histórico-demográfico, las capitales de provincia son los centros administrativos históricos de cada provincia. Por otro lado, geográficamente, se ha realizado un cálculo exploratorio de las áreas de influencia de las capitales de provincia mediante poligonaciones de Voronoi (Voronoi, 1908) con el software ArcGis 10.0. El resultado muestra un mapa extremadamente similar al mapa de provincias.

Las tres matrices de distancias se compararon realizando un test de Mantel (Mantel, 1967) a tres utilizando el paquete de R-Project "ade4" (Thioulouse, 1997).

La representación gráfica de las diferentes distancias entre las poblaciones, nos permite también profundizar en la comparación entre estructura genética y estructura de la diversidad de fauna vertebrada. Para ello se han utilizado dos caminos. Por un lado ambas estructuras se han representado mediante un gráfico tipo de aglomeración según el método de Ward (Ward, 1963). Por otro lado, estos gráficos de aglomeración se han plasmado sobre un mapa geográfico español utilizando el software GeoPhyloBuilder (Kidd y Liu, 2008).

Clima

Con objeto de explorar las relaciones de la estructura poblacional española y los patrones de distribución de la diversidad faunística con las condiciones ambientales, se han comparado las principales variables climáticas (pluviosidad y temperatura) presentes en los grupos resultantes.

Los datos de temperatura y precipitaciones se obtuvieron del proyecto Worldclim versión 1.4. (Hijmans et al., 2005). Las capas de información climática se han generado por interpolación a partir de los datos puntuales obtenidos de la red de estaciones meteorológicas. La resolución de la malla es de 30" (aproximadamente 1km^2 cada pixel)

Estas variables se han comparado visualmente representándolas en un boxplot. Posteriormente se ha analizado la posible relación de las variables climáticas en los grupos utilizando el contraste no paramétrico de Wilcoxon (Wilcoxon, 1945), se ha utilizado el paquete "MASS" de R-Project (Venables y Ripley, 2002).

Diversidad genética - diversidad faunística

Tradicionalmente, para la estima de la diversidad genética a partir de los apellidos se vienen utilizando los mismos coeficientes de diversidad especifica

que se emplean en ecología (Lewontin, 1972). Generalmente con muestras de tamaños comparables, estos coeficientes resultan muy adecuados, pero no siempre los tamaños muestrales son controlables. En el caso que nos ocupa las provincias españolas tienen tamaños tanto poblacionales como geográficos muy diferentes. En una provincia más poblada será esperable encontrar mayor diversidad genética, de la misma forma que en una provincia más extensa, cabe esperar mayor diversidad especifica. Con el objeto de solventar el sesgo que los diferentes tamaños muestrales pueden provocar se ha utilizado la rarefacción como estimación de la consanguinidad (Hulbert, 1971; Magurran, 1988) que muestra la diversidad para un mismo número de individuos.

Para el cálculo de la diversidad genética se han utilizado la base de datos completa de apellidos por varias razones. En primer lugar la identificación del origen de los apellidos se ha hecho a nivel nacional, la diversidad genética se calcula a nivel provincial y lo que a nivel nacional puede parecernos que ha tenido varios orígenes, a nivel provincial puede no tenerlos. En segundo lugar, el nivel provincial es mucho más limitado y a niveles más pequeños geográficamente, se considera que la posibilidad de que un mismo apellido haya aparecido en más de una ocasión es despreciable (Pettener et al., 1998; Esparza et al., 2006; Boattini et al., 2007; Rodríguez-Díaz y Blanco-Villegas, 2010; Boattini et al., 2011). Por último, la base de datos del INE elimina los apellidos que aparecen en menos de 5 ocasiones, de esta manera la posibilidad de que, por azar, dos apellidos aparezcan en el mismo sitio en diferentes ocasiones se ve reducida.

El índice de rarefacción por provincia se ha calculado tanto para apellidos como para especies empleando el paquete de R-Project Vegan (Oksanen et al, 2011)

Relación diversidad genética - diversidad específica

Una vez calculada así las diversidades específica y genética, se han analizado, para conocer si existe alguna relación y, de ser así, tratar de comprenderla. Para ello se ha representado la relación y se ha analizado la nube de puntos

utilizando el paquete graphics de R-Project (R Development Core Team, 2011). Una vez analizada la nube de puntos, se ha estudiado la relación de ambas diversidades con el paquete stats de R-Project (R Development Core Team, 2011).

Resultados

Relación entre distancias

	P-Value	*R*
Distancias geográficas - Distancias de genéticas	*0,001*	*0,4019882*
Distancias de div. faunística - Distancias de genéticas	*0,001*	*0,3655734*
Distancias de div. faunística - Distancias geográficas	*0,001*	*0,6170604*

Tabla 1. Resultados Test de Mantel a tres (Distancia geográfica, distancia genética y distancia por especies).

Como resultado de los cálculos de distancias, se han obtenido tres matrices de distancias; una de distancias genéticas, otra de distancias de diversidad de especies (grado de divergencia en la composición de especies entre provincias) y distancias en línea recta. Para explorar la posible existencia de relación entre estas tres matrices de distancias se ha realizado un test d Mantel a tres (Tabla 1). Los resultados son significativos en los tres casos ($p<0,01$). Lo que apuntaría a la existencia de una relación entre las distancias genéticas y las distancias faunísticas y de ambas con las distancias geográficas.

Estructura genética - Estructura específica

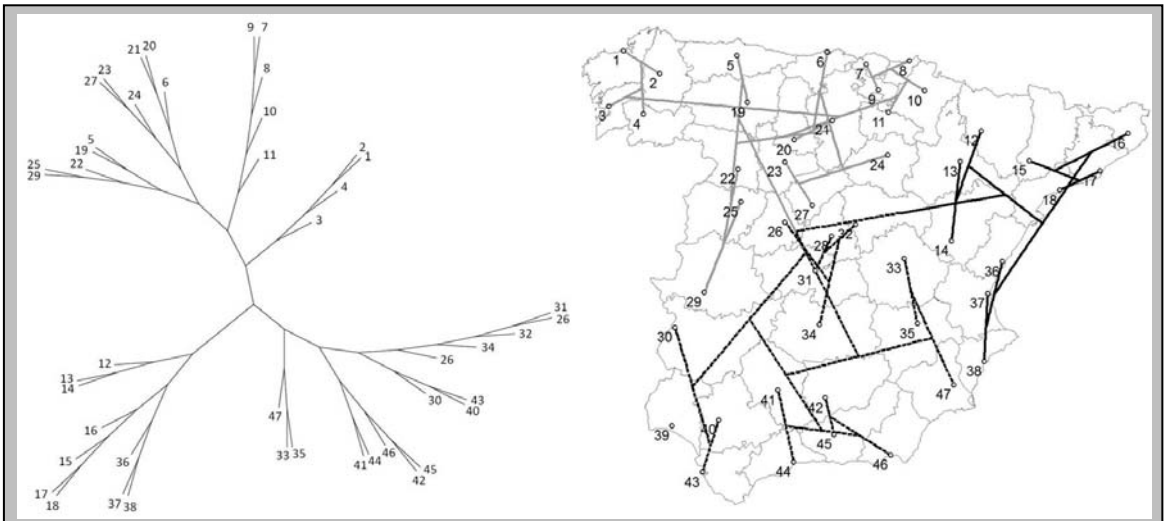


Figura 1. Cluster de las provincias en base a sus Apellidos.

En el caso de las distancias genéticas (Figura 1) resulta evidente la existencia de dos grupos diferenciados. Uno corresponde geográficamente a la zona Norte-

Oeste, el otro a la parte Sur-Este. Este resultado es prácticamente igual al observado al representar las distancias faunísticas (Figura 2).

Además de la coincidencia de los dos grandes grupos resultantes, resulta interesante observar otras similitudes y diferencias de menor entidad.

En la estructura genética (Figura 1), dentro del grupo Norte-Oeste, se observa la existencia de un clado que rompe la continuidad a lo largo de la cornisa Norte y que se extiende hacia el Sur (Figura 1: 7, 8 , 9, 10 y 11). En la estructura faunística (Figura 2), ese grupo no aparece, y la continuidad por la cornisa cantábrica es perfecta.

En cuanto al grupo Sur-Este, en la estructura genética (Figura 1), se encuentra dividido en dos, una zona Este y otra Sur, en la Parte Este se observa una continuidad por la costa mediterránea. Prácticamente lo mismo sucede al observar las especies (Figura 2), igualmente se encuentra dividido en Este y Sur y se observa la misma continuidad por la costa, con la única diferencia de que en esta ocasión se prolonga más hacia el Sur siguiendo el Mediterráneo.

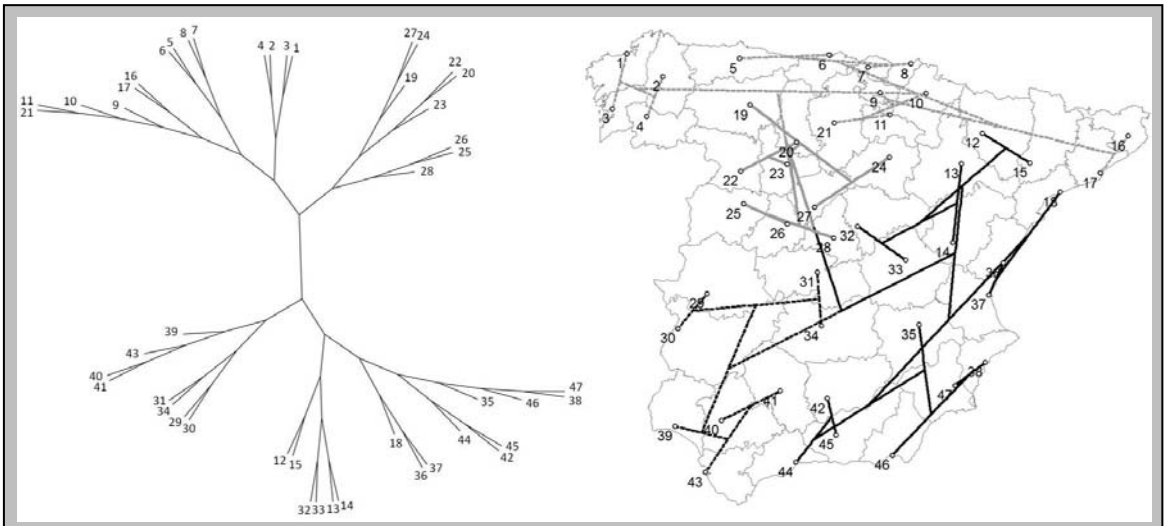


Figura 2. Cluster de las provincias en base a sus Especies animales.

Completando el estudio de la relación entre distancias, se han comparado las variables climáticas presentes en la geografía de cada grupo (Figura 3). Las precipitaciones y la temperatura en el grupo Norte-Oeste por apellidos son prácticamente idénticas a las precipitaciones y la temperatura en ese mismo

grupo por especies. Simétricamente ambas variables son idénticas en el grupo Sur-Este, tanto en el estimado usando los apellidos como en el estimado usando los animales.

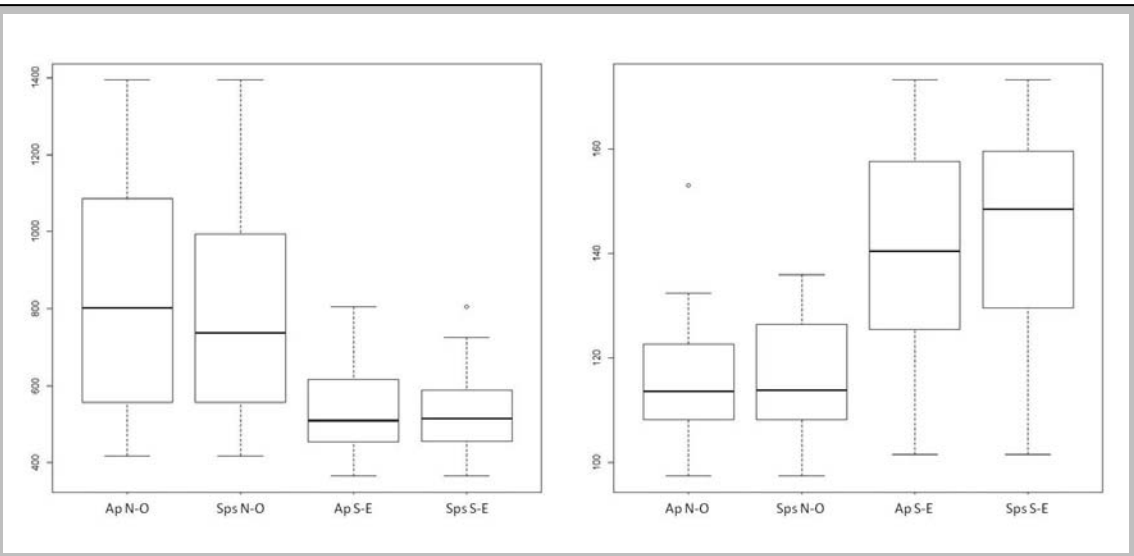


Figura 3. Boxplot comparando las variables climáticas (Precipitaciones Izquierda, Temperatura derecha) entre los grupos según apellidos y según especies.

La pluviosidad en el grupo del Norte-Oeste y en el grupo Sur-Este (Tabla 2), resultan ser significativamente diferentes, tanto en la división genética (p-value = 0,0002) como en la división específica (p-value = 0,0003). Idéntica situación se refiere en lo tocante a la temperatura (p-value = 0,0007 en la división genética; p-value = 0,0001; en la división específica).

		P-Value
Pluviosidad	*Entre grupos de estructura por apellidos*	*0,0002*
	Entre grupos de estructura por especies	*0,0003*
Temperatura	*Entre grupos de estructura por apellidos*	*0,0007*
	Entre grupos de estructura por especies	*0,0001*

Tabla 2. Resultados del test de Wilcoxon comparando las variables climáticas entre grupos.

Diversidad genética - diversidad específica

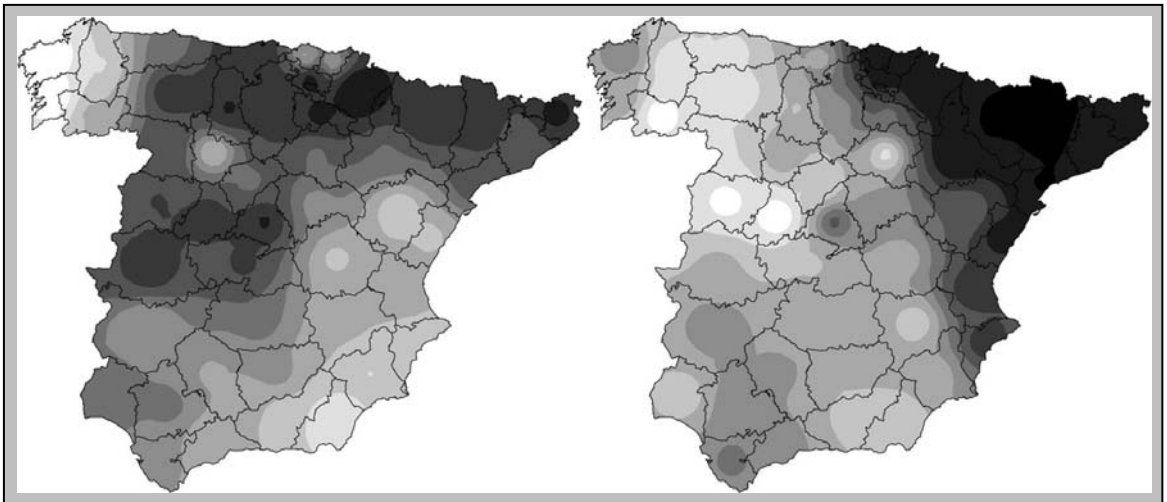


Figura 4. Distribución geográfica de la diversidad (rarefacción) específica (Izquierda) y genética (Derecha).

Conocidas las estructuras genética y faunística, se han observado también las diversidades y sus posibles relaciones. Para empezar se ha representado el índice de rarefacción de cada provincia (Figura 4) tanto para la fauna de vertebrados como genético. En líneas generales, la mayor diversidad de especies de vertebrados parece distribuirse en un gradiente Noroeste-Sureste de mayor diversidad a menor. Concentrándose en los entornos montañosos. La distribución de la diversidad genética obedece también a un gradiente, pero en esta ocasión Noreste-Suroeste.

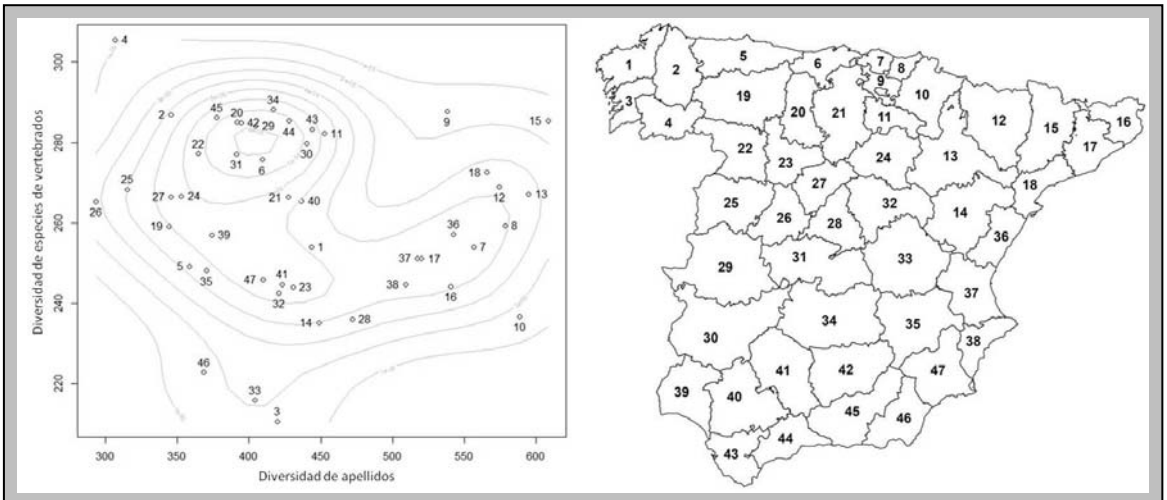


Figura 5. Relación entre la diversidad biológica y la diversidad genética.

Más allá de la comparación visual se ha estudiado la posible correlación entre diversidad de fauna y diversidad genética (Figura 5 y Tabla 3). A nivel nacional no parece existir correlación (Tabla 3). Sin embargo hay que analizar esta relación más cuidadosamente. Con ese propósito se ha hecho un gráfico de dispersión (Figura 5) y se ha analizado la densidad de puntos. Este análisis muestra las aglomeraciones de puntos (provincias en este caso) y si hay grupos diferentes que puedan estar mostrando relaciones diferentes. El análisis muestra dos grupos claros.

	P-Value	*R-squared*
Diversidad α apellidos - Diversidad α especies (Total)	*0,4101*	*-0,006752*
Diversidad α apellidos - Diversidad α especies (Grupo1)	*0,00654*	*0,5772*
Diversidad α apellidos - Diversidad α especies (Grupo2)	*0,3356*	*-0,001372*

Tabla 3. Correlación entre Biodiversidad y diversidad genética.

En el primer grupo (Figura 6 y Tabla 3), que coincide con las poblaciones del Norte-Este español (Figura 5), si existe correlación positiva entre la diversidad genética y la diversidad faunística (p-value = 0,0065; R-squared = 0,5772). El segundo grupo (Figura 6 y Tabla 3), que coincide con poblaciones del Centro, Sur y Norte-Oeste, no muestra correlación (p-value = 0,3356; R-squared = -0,0014).

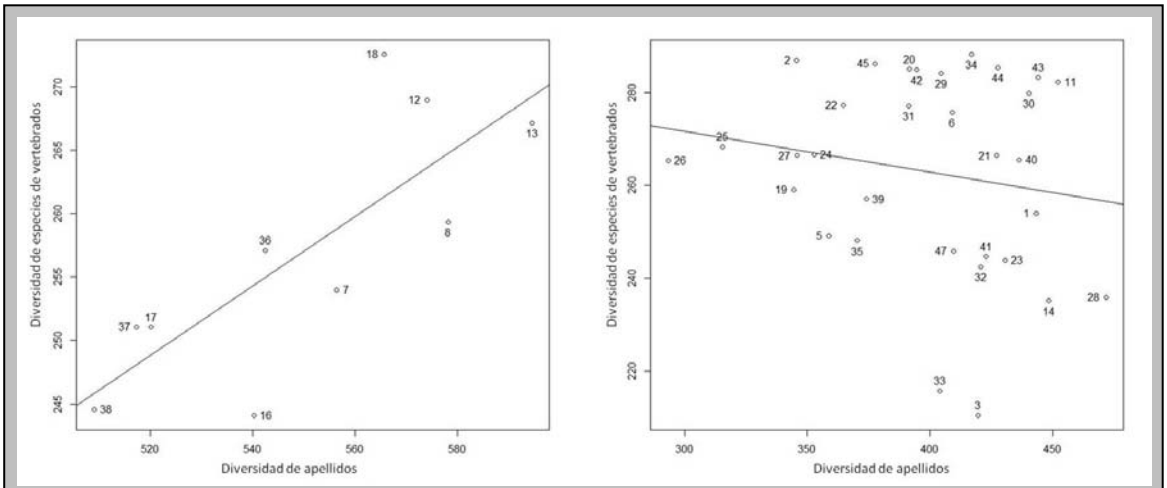


Figura 6. Izquierda relación entre la diversidad específica y la diversidad genética en Este peninsular (Grupo 1); derecha relación entre la diversidad específica y la diversidad genética en el Centro, Sur y Norte-Oeste peninsular (Grupo2).

Para intentar conocer bien la razón por la que la relación entre diversidad genética y diversidad de fauna es diferente en cada grupo hemos estudiado la relación de la diversidad genética con la densidad poblacional (Tabla 4). al hablar de la población peninsular española en general, la diversidad genética correlaciona claramente con la densidad poblacional (p-value = 0,01098, R-squared = 0,116). En el grupo 1 sin embargo (Tabla 4) no existe esta correlación (p-value = 0,2102, R-squared = 0,08681) que si existe en el resto del territorio español (p-value = 0,01107, R-squared = 0,1696).

	P-Value	*R-squared*
Diversidad a apellidos - Densidad poblacional	*0,01098*	*0,116*
Diversidad a apellidos - Densidad poblacional (Grupo1)	*0,2102*	*0,08681*
Diversidad a apellidos - Densidad poblacional (Grupo2)	*0,01107*	*0,1696*

Tabla 4. Correlación entre diversidad genética y otros parámetros.

Discusión

Por lo que sabemos de anteriores resultados, la estructura genética española parece estar muy vinculada con los factores físicos, en este caso con la estructura geográfica de la península (Martínez, 2003), de forma similar a lo que sucede en otras poblaciones (Boattini et al., 2006; Esparza et al., 2006; Boattini et al., 2007). Es el resultado de las relaciones entre poblaciones en forma de intercambios poblacionales que, en este caso, se producen mayoritariamente en función de la conformación geográfica.

La estructura de la diversidad de especies viene determinada por factores biogeográficos ya sean bióticos o abióticos. Por ello, la comparación de la diversidad faunística con la genética puede ser una interesante fuente de información. De hecho, la vinculación entre estructura genética de diferentes especies y biodiversidad y los factores que condicionan a ambas es un campo de estudio candente en la biogeografía (Baselga, 2013) y que no ha alcanzado aún a la antropología.

Relación entre distancias

Las matrices de distancia que recogen la estructura genética y la estructura de la diversidad específica (Tabla 1), mantienen una clara vinculación entre sí (p-value<0,001). Sin embargo, esta similitud debe analizarse minuciosamente, ya que ambas matrices de distancias también guardan relación con las distancias geográficas (Tabla 1; p-value<0,001). En las poblaciones humanas, esta relación es la esperable de un modelo de aislamiento por distancia (Malecot, 1955), que también se verifica en la diversidad de especies animales (Bell, 2001; Rosindell et al., 2011) y que, a priori, podría significar sencillamente que las provincias se parecen más cuanto más cerca están, tanto en lo que a composición de especies de animales se refiere, como a los seres humanos que las pueblan.

Por lo tanto, antes de realizar cualquier afirmación sobre el alcance de la similitud entre ambas estructuras, es imprescindible profundizar más en la comparación.

Estructura genética - Estructura específica

Para ver qué similitudes guardan realmente ambas estructuras, se han representado en un cluster y este sobre un mapa del territorio español (Figuras 1 y 2). La comparación de ambas estructuras nos permitirá ver hasta qué punto llega su relación y extraer conclusiones de la misma.

La primera división en ambas estructuras muestra dos grandes grupos uno en el Sur-Este y otro en el Norte-Oeste. El hecho de que las dos estructuras coincidan a este nivel hace pensar que los procesos que han desembocado en las estructuras genética y específica de la península, han estado desencadenados por los mismos factores (factores biogeográficos). Esta similitud entre distancias intraespecíficas y diversidad de especies, no es algo extraño y ya se ha observado en otras especies (Vellend y Ripley, 2005).

Con el objeto de comprender bien las variables implicadas en la estructura genética y la diversidad específica españolas (Figura 2), se han analizado las variables climáticas en cada uno de los dos grandes grupos observados (Figura 3). La importancia de estas variables climáticas reside en que están estrechamente relacionadas con los otros factores biogeográficos (altitud, latitud, longitud, orografía,...) y por lo tanto suponen un excelente resumen.

Lo más interesante es que estas diferencias son significativas tanto entre los dos grupos encontrados a partir de la diversidad faunística como en los encontrados a partir de los apellidos (Figura 3). Parece que a este nivel los factores climáticos también guardan una relación significativa con la población española, probablemente condicionando todo el resto de factores que han desembocado en la actual estructura. Ambas variables determinan las

características del entorno que parece muy relacionado tanto con la estructura genética como con la estructura específica (Tabla 2).

Por los resultados anteriores, sabemos también que los movimientos migratorios que han conformado la estructura actual de la población española se han producido principalmente en el interior de estos dos grupos. Parece que estos movimientos podrían estar relacionados con el entorno dentro del que se producen.

Sin embargo la comparación entre ambas estructuras puede resultar aún más fructífera. Si analizamos lo que sucede dentro de cada uno de los dos grandes grupos (una escala menor), se observa que existen ciertas diferencias:

- Dentro del grupo Norte-Oeste (Figuras 1 y 2), en la estructura genética, observamos una fragmentación en grupos menores; las cuatro provincias de Galicia (Figura 1; 1, 2, 3 y 4) y las tres del País Vasco (Figura 1; 7, 8 y 9). Ambos grupos poseen un idioma diferente al hablado en el resto del territorio nacional (García, 2007; Goebl, 2010). En el mismo grupo Norte-Oeste aparece otro grupo menor (Figura 1; 5, 19, 22, 25 y 29) que se corresponde con las provincias de la histórica Ruta de la Plata (Martínez, 2003). Ninguno de estos grupos aparece en el cluster de especies, que muestra un continuo que sigue el trazado de las cordilleras cantábrica y pirenaica.
- Dentro del grupo Sur-Este (Figuras 1 y 2), en la estructura genética, aparece un grupo diferenciado del resto que tampoco aparece en el cluster de especies y que incluye las provincias catalanoparlantes (García, 2007; Goebl, 2010) relacionadas también con las provincias aragonesas (Figura 1; 12, 13, 14, 15, 16, 17, 18, 36, 37 y 38).

Si asumimos que las similitudes pueden ser consecuencia de la influencia de factores similares, las diferencias deberían ser consecuencia de factores diferentes. Así parece que en el interior de cada grupo, a una escala menor, los factores históricos y etno-lingüísticos, si han jugado el papel que se observa en

otras poblaciones. La presencia de los grupos correspondientes a Galicia, País Vasco y las provincias catalanoparlantes podrían deberse a la influencia de factores etnolingüisticos (García, 2007; Goebl, 2010) y la presencia del grupo correspondiente a la ruta de la plata a factores históricos (Martínez, 2003). Estas diferencias se observan a niveles menores por lo que parece que la influencia de unos u otros factores pueden estar muy ligadas al nivel poblacional, viéndose la influencia de los factores biogeográficos en los niveles más amplios.

La herramienta que hemos utilizado para plasmar los cluster en el mapa, fue diseñada por sus autores para identificar los corredores biológicos empleados por una especie (Kidd y Liu, 2008), por lo que puede aportarnos información en este sentido. En cuanto a la diversidad zoológica, la interpretación de estos gráficos, al incluir una gran variedad de especies, sería más complicada. Pero en lo que a la estructura genética se refiere, los continuos se pueden interpretar como auténticos corredores biológicos (Figura 1). Como hemos dicho estos continuos serían tres. Dos de esos corredores (el cantábrico y el mediterráneo) coinciden con la estructura zoológica y por lo tanto podría considerarse que están relacionados con factores biogeográficos, el tercero (la ruta de la plata) no coincide y estaría relacionado con factores históricos.

Diversidad genética - diversidad específica

Siguiendo con la comparación se han tomado esta vez diversidad genética y diversidad de fauna vertebrada. En nuestro caso y para evitar la influencia de la diferencia de tamaño en las muestras (Hulbert, 1971; Magurran, 1988), hemos utilizado la rarefacción (Figura 4).

Si analizamos la relación entre ambas diversidades en todo el territorio peninsular español, no existe relación entre diversidad genética y diversidad específica (Tabla 3). Sin embargo se ha analizado el diagrama de dispersión con diagramas de densidad. El resultado es que existen dos grupos de provincias muy diferenciados (Figura 6, Tablas 3 y 4).

El grupo 1 está situado al Este de la península y corresponde con las provincias catalanas, valencianas y casi todas las aragonesas (Figuras 5 y 6, Tabla 4). Hablamos de un grupo de provincias peculiares, con más densidad poblacional, diferencias lingüísticas (García, 2007; Goebl, 2010) y situadas entre irregularidades orográficas, los Pirineos, el Sistema Ibérico, el río Ebro y el mar mediterráneo. La distribución de la diversidad genética y de la diversidad específica en este grupo, siguen un mismo patrón (Figura 4). Es máxima en la zona de los pirineos y se va reduciendo a medida que nos acercamos al mediterráneo. De hecho diversidad genética y diversidad específica correlacionan (Tabla 3). El grupo 2, situado en el centro y Oeste peninsular (Figuras 5 y 6, Tabla 4). En esa zona, la distribución de la biodiversidad y la de la diversidad genética (Figura 4) no parecen guardar el mismo patrón. De hecho, en esta zona menos poblada del territorio nacional, diversidad de especies y diversidad genética no guardan relación (Tabla 3).

Sabiendo que la diversidad de especies viene determinada por factores biogeográficos, las similitudes que guarda con la diversidad genética podrían ser atribuidas estos mismos factores. Aunque reducir el tamaño dividiendo la muestra en dos grupos, complica la extracción de conclusiones.

Intentando comprender mejor lo que sucede se ha comparado la diversidad genética con la densidad poblacional (Tabla 4). A nivel nacional si existe una relación que después sólo se da en el grupo 2 (Tabla 4). Parece por lo observado que, aunque existe una vinculación entre diversidad faunística y diversidad genética y este hecho apunta a una influencia de factores biogeográficos sobre la diversidad genética humana, esta influencia no es constante ni homogénea y, en ocasiones, parece que podría estar más relacionada con factores socioeconómicos como la densidad poblacional.

Conclusión

La estructura genética de la población española y la de la biodiversidad coinciden. Una evidencia de la importancia de los factores biogeográficos en las poblaciones humanas modernas que parecen haber influenciado la población española en el mismo grado que la diversidad animal (no necesariamente siguiendo los mismos procesos).

A una menor escala, se evidencia en la población española la influencia de otros factores como lingüísticos (Galicia, País Vasco, Cataluña y Valencia) o históricos (Ruta de la Plata).

La diferencia de comportamiento entre los dos grupos puede deberse a las peculiaridades del grupo dos, una población en la zona de contacto entre la península y el continente y, por lo que sabemos de anteriores trabajos, situada entre barreras (Pirineos y Sistema Ibérico), que no está especialmente poblada y no es objeto de grandes movimientos poblacionales.

Pese a la complejidad del sistema de factores que se relacionan con la diversidad genética humana, a la luz de los resultados, parece que los factores biogeográficos podrían jugar un papel que hasta la fecha no ha sido suficientemente tenido en cuenta.

Bibliografía

Adams S M, Bosch E , Balaresque P L, Ballereau S J, Lee A C, Arroyo E, Lopez-Parra A M, Aler M, Grifo M S, Brion M, Carracedo A, Lavinha J, Martinez-Jarreta B, Quintana-Murci L, Picornell A, Ramon M, Skorecki K, Behar D M, Calafell F, Jobling M A. 2008. The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. Am J Hum Genet 83: 725-736.

Alvarez,L.; Santos,C.; Ramos,A.; Pratdesaba,R.; Francalacci,P.; Aluja,M.P. Mitochondrial DNA patterns in the Iberian Northern plateau: Population dynamics and substructure of the Zamora province. Am.J.Phys.Anthropol., 2010.

Balanovskaia EV, Romanov A G, Balanovskii OP (2011) Namesakes or relatives? Approaches to investigating the relationship between Y chromosomal haplogroups and surnames. Mol Biol 45: 473-485.

Barrai I, Rodríguez-Larralde A, Manni F, Ruggiero V, Tartari D, et al. (2004) Isolation by language and distance in Belgium. Ann Hum Genet 68: 1-16.

Barrai I, Rodriguez-Larralde A, Manni F, Scapoli C (2002) Isonymy and isolation by distance in the Netherlands. Hum biol 7 42: 263-283.

Baselga A, Fujisawa T, Crampton-Platt A, Bergsten J, Foster P G, Monaghan M T, Vogler A P (2013) Whole-community DNA barcoding reveals a spatio-temporal continuum of biodiversity at species and genetic levels. Nat Commun 4:1892.

Bell, G (2001) Neutral macroecology. Science 293: 2413–2418.

Boattini A, Blanco-Villegas MJ, Pettener D (2007) Genetic structure of La Cabrera, Spain, from surnames and migration matrices. Hum Biol 79: 649-666.

Boattini A, Calboli FC, Blanco-Villegas MJ, Gueresi P, Franceschi MG, et al. (2006) Migration matrices and surnames in populations with different isolation patterns: Val di Lima (Italian Apennines), Val di Sole (Italian Alps), and La Cabrera (Spain). Am J Hum Biol 18: 676-690.

Boattini A, Griso C, Pettener D (2011) Are ethnic minorities synonymous for genetic isolates? Comparing Walser and Romance populations in the Upper Lys Valley (Western Alps). J Anthropol Sci 89: 161-173.

Caravello GU, Tasso M (1999) An analysis of the spatial distribution of surnames in the Lecco area (Lombardy, Italy) Am J Hum Biol 11: 305-315.

Caravello GU, Tasso M (2002) Use of surnames for a demo-ecological analysis: a study in southwest Sardinia. Am J Hum Biol 14: 391-397.

Cavalli-Sforza L L (2000) Genes, Peoples, and Languages. North Point Press, New York.

Chakraborty R, Schwartz RJ (1990) Selective neutrality of surname, distribution in an immigrant indian community of Houston, Texas. Am J Hum Biol 2: 1-15.

Colantonio SE, Lasker GW, Kaplan BA, Fuster V (2003) Use of surname models in human population biology: a review of recent developments. Hum Biol 75:785-807.

Dipierri J, Rodríguez-Larralde A, Alfaro E, Scapoli C, Mamolini E, et al. (2011) A Study of the Population of Paraguay through Isonymy. Ann Hum Genet 75: 678-687.

Esparza M, García-Moro C, Hernández M (2006) Genetic relationships between parishes in the Ebro delta region (Spain) as estimated by migration matrix and surnames. Hum Biol 78: 647-662.

Fiorini S, Tagarelli G, Boattini A, Luiselli D, Piro A, et al. (2007) Ethnicity and Evolution of the Biodemographic Structure of Arbëreshe and Italian Populations of the Pollino Area, southern Italy (1820-1984). A. Anthropol 109: 735-746.

Gagnon A, Heyer E (2001) Intergenerational correlation of effective family size in early Quebec (Canada). Am J Hum Biol 13: 645-659.

García P (2007) Lenguas y dialectos de España. Madrid: Arco Libros.

Goebl H (2010) La dialectometrización del ALPI: Rápida presentación de los resultados. 26th CILFR. Valencia.

Hedrick PW (1971) A new approach to measuring genetic similarity. Evolution 25: 276-280.

Hijmans R J, Cameron S E, Parra J L, Jones P G, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. Intern J Climat 25: 1965-1978.

Hubbell S P (2001) The Unified Neutral Theory of Biodiversity and Biogeography. Princeton. Princeton University Press.

Hurlbert S H (1971) The nonconcept of species diversity: a critique and alternative parameters. Ecology 52:577-86.

Kidd D M, Liu X (2008) GEOPHYLOBUILDER 1.0: an ARCGIS extension for creating 'geophylogenies'. Mol Ecol Resour 8: 88–91.

King TE, Ballereau SJ, Schurer KE, Jobling MA (2006) Genetic signatures of coancestry within surnames. Curr Biol 16: 384-388.

King TE, Jobling MA (2009) Founders, drift and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. Mol Biol Evol 26:1093-102.

King TE, Jobling MA (2009) What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. Trends Genet 25: 351-360.

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cyber 43: 59-69.

Kohonen T (1984) Self-organization and associative memory. Berlin: Springer.

Lasker G W, Kaplan BA (1985) Surnames and genetic structure: repetition of the same pairs of names of married couples, a measure of subdivision of the population. Hum biol 57: 431-440.

Lewontin R C (1972) The apportionment of human diversity. Evol Biol 6:381-398.

Lisa A, De Silvestri A, Mascaretti L, Degiuli A, Guglielmino CR (2007) HLA genes and surnames show a similar genetic structure in Lombardy: does this reflect part of the history of the region?. Am J Hum Biol 19: 311-318.

Lomolino MV, Riddle BR. Whittaker RJ, Brown JH (2010) Biogeography 4th edition. Sinauer. Associates, Inc. 878 pp.

Magurran A E (1988) Ecological Diversity and Its Measurement. Princeton. Princeton University Press.

Malecot G (1955) Decrease of relationship with distance. Cold Spring Harbor Syrup. Quant Biol 20: 52-53.

Manni F, Heeringa W, Toupance B, Nerbonne J (2008) Do surname differences mirror dialect variation? Hum biol 80: 41-64.

Manni F, Toupance B, Sabbagh A, Heyer E (2005) New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Y-chromosome sampling. Am J Phys Anthropol 126: 214-228.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27: 209-220.

Martínez E (2003) Atlas histórico de España. Madrid: Istmo.

Morton NE, Yee S, Harris DE, Lew R (1971) Bioassay of Kinship. Theor Popul Biol 2: 507-524.

Nicholas J, Lewin-Koh RB, contributions by Pebesma EJ, Archer E, Baddeley A, et al. (2012) Maptools: Tools for reading and handling spatial objects. R package version 0.8-14. http://CRAN.R-project.org/package=maptools.

Noss R F (1990) Indicators for monitoring biodiversity - a hierarchical approach. Conserv Biol 4: 355–364.

Oksanen J, Blanchet F G,Kindt R, Legendre P, Minchin P R, O'Hara R B, Simpson G L, Solymos P, Henry M, Stevens H, Wagner H (2011) vegan: Community Ecology Package. R package version 2.0-2. http://CRAN.R-project.org/package=vegan.

Pettener D, Pastor S, Tarazona-Santos E (1998) Surnames and genetic structure of a high-altitude quechua community from the Ichu river valley peruvian central Andes 1825-1914. Hum Biol 70: 865-87.

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL http://www.R-project.org/.

Rodríguez-Díaz R, Blanco-Villegas MJ (2010) Genetic structure of a rural region in Spain: distribution of surnames and gene flow. Hum Biol 82: 301-314.

Rosindell J, Hubbell S P, Etienne R S (2011) The unified neutral theory of biodiversity and biogeography at age ten. Trends Ecol Evol 26: 340–348.

Stepp J R, Cervone S, Castaneda H, Lasseter A, Stocks G, Gichon Y (2004) Development of a GIS for global biocultural diversity. Policy Matters 13 (special issue): 267–270.

Sutherland WJ (2003) Parallel extinction risk and global distribution of languages and species. Nature 423: 276–279.

Sykes B, Irven C (2000) Surnames and the Y Chromosome. Am J Hum Genet 66: 1417-1419.

Thioulouse J, Chessel D, Dole´dec S, Olivier JM (1997) ADE-4: a multivariate analysis and graphical display software. Stat Comput 7: 75-83.

Vellend M, Geber M A (2005) Connections between species diversity and genetic diversity. Ecol Lett 8: 767–781.

Venables WN, Ripley BD (2002) Modern Applied Statistics with S. Fourth Edition. New York: Springer.

Voronoi G F (1908) Nouvelles applications des paramètres continus à la théorie de formes quadratiques. J Reine Ang Mat 134: 198–287.

Ward J H (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58: 236-244.

Weiss V (1980) Inbreeding and genetic distance between hierarchically structured populations measured by surname frequencies. Mankind Quarterly 21: 135-149.

Wilcoxon F (1945) Individual Comparisons by Ranking Methods. Biometrics Bulletin 6: 80-83.

4. CONSANGUINIDAD EN LA POBLACIÓN ESPAÑOLA

CONSANGUINIDAD EN LA POBLACIÓN ESPAÑOLA

Resumen

Introducción: Los estudios de consanguinidad en amplias regiones geográficas son relativamente poco frecuentes o someros. En este contexto, la población española con su sistema de dos apellidos (el primero heredado del padre y el segundo de la madre) supone una oportunidad especial que permite estimar el nivel de consanguinidad. Pero además, la homogeneidad cultural y religiosa de esta población permite valorar la influencia de otros factores. El objetivo es analizar los niveles de consanguinidad en las diferentes provincias españolas y valorar la influencia de factores culturales, económicos, sociodemográficos o ambientales.

Material y métodos: Se ha utilizado la base de datos del padrón continuo de 2008 correspondiente a las 47 provincias peninsulares españolas.

Se ha estimado el nivel de consanguinidad por isonimia total (Ft) y cada una de sus componentes, casual (Fr) y no casual (Fn) en cada una de las provincias y se ha estudiado su distribución. Posteriormente se ha analizado la relación de cada una de las componentes con diversos factores (aislamiento, ruralidad, economía, educación, demografía y relieve) y se han elegido los modelos más representativos (según los criterios de información, AIC y BIC).

Resultados: La consanguinidad muestra una relación directa con la ruralidad, la orografía y el relieve e indirecta con el nivel económico poblacional. Dichos factores condicionan la consanguinidad en diferente grado, siendo el nivel económico y la ruralidad los más determinantes

Gracias a todo ello, se ha observado no sólo que el nivel de consanguinidad de una población es el resultado de la influencia conjunta del ambiente rural, el nivel económico, el nivel educativo, la edad al matrimonio y la orografía. De la misma forma se han podido identificar en qué medida cada uno de estos factores condiciona la consanguinidad y hasta qué punto afectan a la percepción social de la consanguinidad.

Introducción

Dentro de los parámetros que caracterizan a una población humana, la consanguinidad probablemente sea uno de los que más tinta ha hecho correr. Su relación con problemas diferentes problemas de salud (Bittles and Black, 2010) en forma de enfermedades hereditarias (Bundey y Alam, 1993; Rittler et al., 2001; Shami et al., 2001), mentales (Vezina et al., 1999; Mansour et al., 2009, 2010; Bener et al., 2012), problemas de mortalidad (Bittles y Neel, 1994; Cavalli-Sforza et al., 2004) o de fertilidad (Helgason et al, 2008) por ejemplo, es algo ampliamente demostrado. Por esta vía, la consanguinidad es un parámetro muy importante en la historia evolutiva de una población.

Es por estas razones por las que se han intentado determinar los factores que condicionan el nivel de la consanguinidad en las poblaciones. Factores de índole religioso y cultural (Bittles and Black, 2010) han demostrado ser determinantes. Pero junto a ellos conocemos también la importancia de factores de otro tipo como demográficos, geográficos, orográficos y económicos (Calderón, 1989; Jorde y Pitkanen, 1991; Fuster et al., 1996; Fuster y Colantonio, 2002; Blanco-Villegas et al., 2004; Rodríguez-Díaz y Blanco-Villegas, 2011).

Dos son las vías que se han utilizado para estudiar la consanguinidad. Por un lado los registros históricos y familiares (Fuster y Colantonio, 2002; Blanco-Villegas et al., 2004; Bittles and Black, 2010; Rodríguez-Díaz y Blanco-Villegas, 2011), que permiten una reconstrucción muy exhaustiva de los parentescos y estudiar tanto los factores que determinan la consanguinidad como la forma en la que esta pueda estar incidiendo en la población. Esta metodología tiene una limitación, y es que se centra en pequeños grupos poblacionales (Calderón, 1989; Jorde y Pitkanen, 1991; Fuster et al., 1996; Blanco-Villegas et al., 2004; Rodríguez-Díaz y Blanco-Villegas, 2011) y sólo permite abarcar grandes grupos mediante recopilaciones (Fuster y Colantonio, 2002; Bittles and Black, 2010).

La otra metodología utilizada es la isonimia (Crow y Mange, 1965; Crow, 1980). Esta metodología si permite abarcar grandes poblaciones (Rodríguez-Larralde,

2003; Scapoli et al., 2007; Dipierri et al., 2011; Rodríguez-Larralde, 2011) pero, por lo general sólo permite realizar una estima somera del nivel de consanguinidad de una población (Scapoli et al., 2007). Sin embargo, el sistema español, en el que se utilizan dos apellidos (Pettener et al., 1998; Fauré et al., 2001; Colantonio et al., 2003; Blanco-Villegas et al., 2004; Mateos y Tucker, 2008; Dipierri et al., 2011) si permite estudiar el nivel de consanguinidad de una población y sus componentes (Crow y Mange, 1965; Crow, 1980). Este sistema ya ha sido aprovechado en algunas ocasiones (Rodríguez-Larralde, 2003, 2011; Dipierri et al., 2011), pero se han realizado sólo descripciones someras del nivel de consanguinidad.

Por otro lado, la población española, además de la particularidad del sistema de apellidos, presenta unas características que la hacen especialmente interesante para el estudio de la consanguinidad. Por ejemplo es uno de los países con mayor consanguinidad de Europa (una sociedad muy avanzada) y tiene las regiones con mayor consanguinidad del continente (Scapoli et al., 2007). Además, la española es una población bastante homogénea en cuanto a cuestiones religiosas y culturales. Lo que ofrece la oportunidad de estudiar la influencia de otros factores de diferente naturaleza en el nivel de consanguinidad de las diferentes provincias.

En el presente estudio, se pretende aprovechar la oportunidad que representan estas características de la población española para estudiar los factores que pueden condicionar el nivel de consanguinidad de una población más allá de factores religiosos y culturales.

Material y métodos

Base de datos

La base de datos de apellidos, se ha obtenido del Instituto Nacional de Estadística (INE), e incluye los datos correspondientes al censo nacional de población del año 2008 en las 47 provincias peninsulares. Los datos del censo nacional son públicos, sin embargo no están accesibles y la información debe solicitarse al INE mediante el procedimiento establecido. Un resumen de esta información está disponible en la página del INE (www.ine.es).

La base de datos incluye los apellidos de toda la población censada en cada municipio en 2008, siempre y cuando aparezcan más de 5 veces en un mismo municipio. Además del primer y el segundo apellido, la base de datos incluye también el número de veces que cada apellido aparece en ambos lugares, es decir aparece a la vez como primer y segundo apellido. Esta información nos permite calcular la consanguinidad de la población española.

Para los cálculos de los anteriores capítulos la base de datos fue depurada, se corrigieron (apellidos repetidos, diferentes grafías, espacios entre palabras, errores ortográficos,...). En cambio en esta ocasión se ha preferido trabajar con la base de datos en bruto, sin depurar. La razón es que la base de datos fue construida por los técnicos del INE y no tenemos forma de saber si nuestras correcciones afectarán a las frecuencias de los apellidos que aparecen a la vez como primero y segundo.

Isonimia

La isonimia en la población española se ha calculado utilizando el método propuesto por Crow y Mange (Crow y Mange, 1965) y modificado posteriormente por Crow (Crow, 1980). Este método permite calcular la consanguinidad total en el seno de una población según la expresión:

$$F_T = F_n + F_r(1 - F_n)$$

Donde F_T es la consanguinidad total, F_n La consanguinidad no casual (consecuencia de la actitud de la población hacia las uniones consanguíneas) y F_r la consanguinidad casual (consecuencia de la presión ambiental).

La consanguinidad casual (F_r) es sólo consecuencia de la composición de apellidos de la población y por lo tanto se puede calcular a partir de cualquier registro de apellidos sin necesidad del sistema español ni de la frecuencia de apellidos dobles:

$$F_r = \frac{\sum p_i q_i}{4}$$

Donde p_i es la frecuencia con la que el apellido i aparece en los varones y q_i es la frecuencia con la que el apellido i aparece en las mujeres. En este caso tenemos las frecuencias en las que cada apellido aparece como primer apellido o como segundo apellido. Por eso asimilamos p_i a la frecuencia con la que el apellido i aparece como primer apellido (apellido paterno o apellido del varón en la generación anterior) y q_i es la frecuencia con la que el apellido i aparece como segundo apellido (Apellido materno o apellido de la mujer en la generación anterior).

Pero el sistema de dos apellidos español, y el hecho de que nuestra base de datos recoja el número de veces que un apellido aparece como primero y segundo, nos permiten calcular también la consanguinidad no casual, dada por la expresión:

$$F_n = \frac{(P - \sum p_i q_i)}{4(1 - \sum p_i q_i)}$$

Donde P es la frecuencia de uniones isonímicas. En el caso que nos ocupa, asimilamos P a la frecuencia de individuos isonímicos (los dos apellidos que porta el sujeto son idénticos).

Siguiendo este procedimiento no estaremos calculando la isonimia de las uniones, sino la isonimia presente en los individuos de la población.

Los resultados de estos cálculos se han representado sobre un mapa del territorio peninsular español utilizando el software ArcGIS 10.0 (Figuras 1, 2 y 3).

Análisis de la isonimia

Por lo general, en los estudios de consanguinidad, se considera que la consanguinidad total es una consecuencia de la presión ambiental y de la actitud social frente a ella (Crow y Mange, 1965; Crow, 1980). Por otro lado, parece que la actitud social frente a la consanguinidad podría ser un equilibrio entre los beneficios que pueden reportar las uniones entre parientes (Pettener, 1985; Calderón, 1989) y los perjuicios (Bittles y Black, 2010). La población española, con el sistema de doble apellido supone una ocasión única para analizar el fenómeno de la consanguinidad en una población tan amplia y con una base de datos tan grande.

El objetivo de este análisis de la consanguinidad es determinar qué factores la condicionan en la población española. Concretamente tratamos de averiguar si el aislamiento, el carácter rural, nivel económico, nivel cultural, demografía y orografía influyen en el nivel de consanguinidad de las diferentes provincias españolas. Para ello, como paso previo se han analizado 31 variables agrupadas en esos 6 bloques para buscar las variables que mejor los puedan representar (Tabla 1).

La consanguinidad que estamos estimando en la población española es la presencia de individuos consanguíneos. Esta presencia será fruto de las uniones entre parientes pasadas. De ahí que a la hora de identificar los factores que influencian la consanguinidad hayamos seleccionado los factores presentes en el año 1995. El año más antiguo al que hemos podido remontarnos para todas las variables consideradas.

Se han realizado regresiones simples entre cada variable y cada una de las componentes de la consanguinidad (Tabla 1), usando el paquete "stats" de R-

Project (R Development Core Team, 2011) para conocer mejor el comportamiento de cada variable. Sin embargo, los diferentes factores y la consanguinidad no están aislados, sino que forman parte de un sistema. Por eso a partir de ahí se han seleccionado las variables que mejor pueden representar a los factores que pueden haber influido en la consanguinidad y se ha estudiado esta influencia de forma conjunta.

Con las variables seleccionadas, se han realizado todos los modelos (combinaciones posibles) mediante regresión múltiple (stepwise) usando el paquete de R-project "stepwise" (Graham et al., 2005) y se ha estimado el peso de cada variable usando el paquete "relaimpo" (Grömping, 2006). Por último, se ha seleccionado el mejor modelo siguiendo el criterio de información de Akaike (Sakamoto et al., 1986), el criterio de información bayesiano (Sakamoto et al., 1986) y la cantidad de variabilidad explicada. El resultado es que podemos seleccionar el mejor modelo para explicar la consanguinidad y en este se habrán seleccionado los factores que la determinan (Tabla 2), conociendo el peso de cada variable, conoceremos también el grado de influencia de cada factor (Figuras 4, 5 y 6).

Resultados

Isonimia

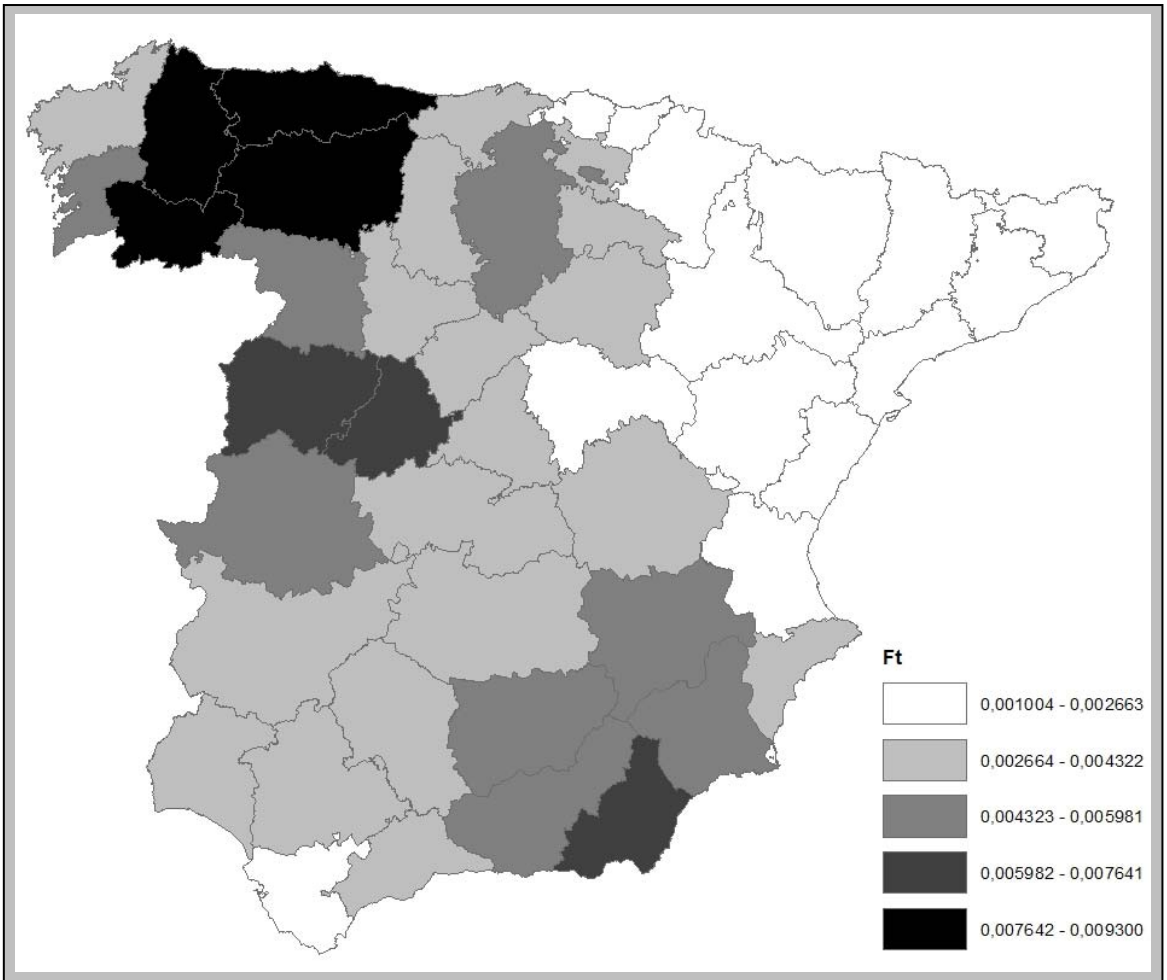


Figura 1. Distribución geográfica de la consanguinidad total (Ft).

La distribución de la consanguinidad total (Ft) muestra una evidente regionalización (Figura 1). La zona de mayor consanguinidad es el Noroeste peninsular. Franja Oeste y Franja Sur también muestran alta consanguinidad. Mientras que la zona Noreste es la de menor consanguinidad total. Parece que la distribución muestra un gradiente horizontal, mayor en el Oeste y menor en el Este.

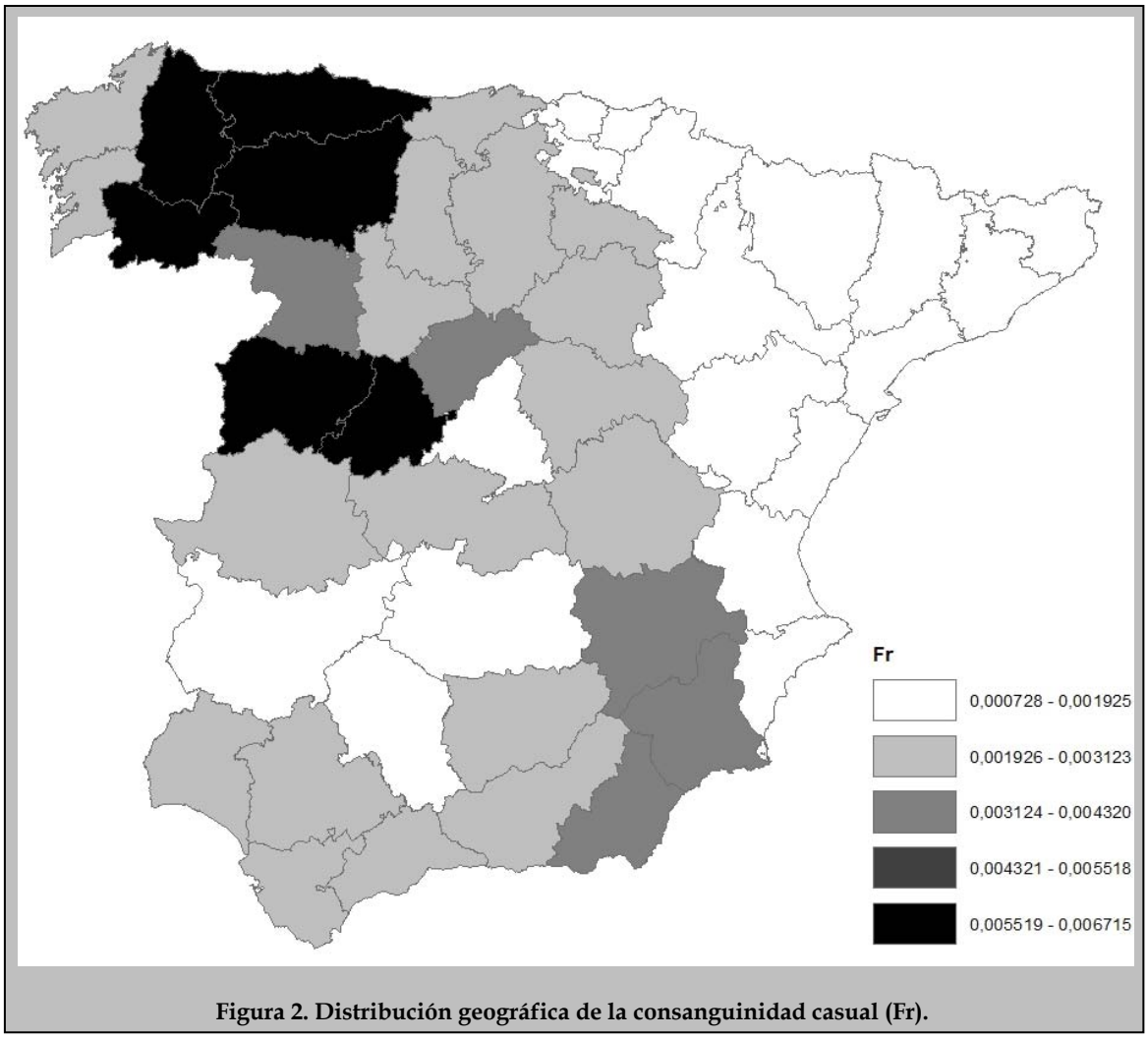


Figura 2. Distribución geográfica de la consanguinidad casual (Fr).

La Figura 2 muestra la consanguinidad casual (Fr), la debida a la influencia del ambiente. La distribución recuerda a la de la consanguinidad total, pero se muestra más concentrada en el Noroeste peninsular. Igual que sucede con la consanguinidad total, el Noreste peninsular muestra una consanguinidad casual muy baja. El posible gradiente Oeste - Este, no es tan claro como en la consanguinidad total.

La consanguinidad no casual (Fn) es la que muestra una distribución más diferenciada, aparece concentrada en el Noroeste y en el Sureste. Por otro lado su distribución es más homogénea, no existe una zona concreta de baja consanguinidad no casual como sucedía con la consanguinidad total y la casual. No se aprecia un patrón como se veía en las consanguinidades total (Ft) y casual (Fr).

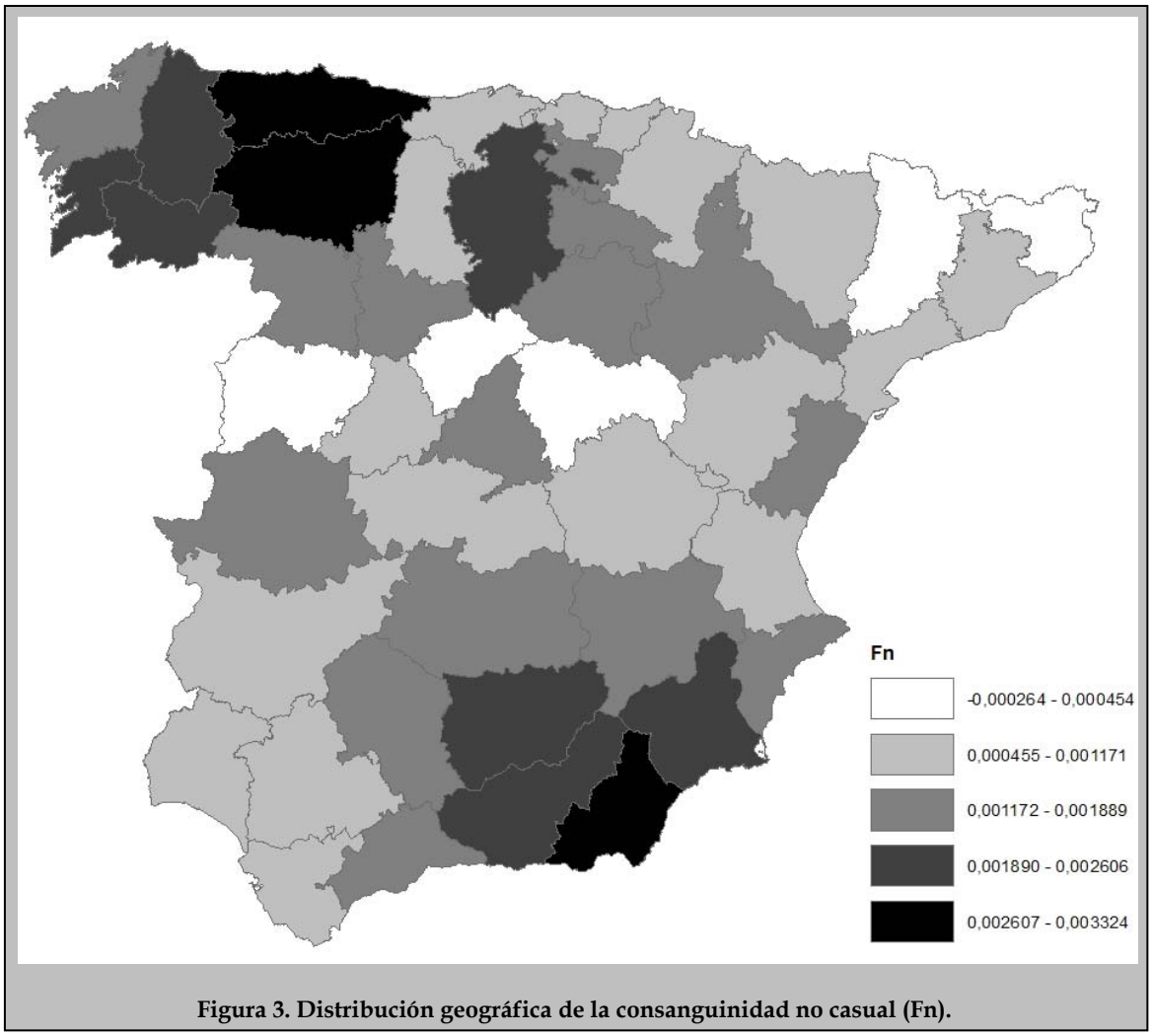


Figura 3. Distribución geográfica de la consanguinidad no casual (Fn).

Análisis de la isonimia

Se han analizado 31 variables diferentes para tratar de dilucidar como influyen el aislamiento, el carácter rural, el nivel económico, el nivel cultural, la demografía y la orografía en la consanguinidad española (Tabla 1).

Lo primero que se ha hecho es estudiar la relación individual de cada una de las variables con cada tipo de consanguinidad (Tabla 1). Es interesante señalar que la consanguinidad total (Ft) y la consanguinidad casual (Fr) correlacionan con alguna variable de todos los factores excepto con el nivel cultural y la demografía. Por el contrario la consanguinidad no casual (Fn), sólo aparece correlacionada con el nivel económico y el cultural.

Tipo	Variable		P - value		
			Ft	Fr	Fn
Aislamiento	Vehículos por 1.000 habitantes	V1	*0,0176*	*0,0428*	*0,0551*
	Vehículos por km^2		0,1098	*0,0566*	0,7727
	Kilómetros de carreteras por 1.000 km^2		0,5237	0,6209	0,5290
	Kilómetros de autopistas y autovías por 1.000 km2	V2	*0,0071*	*0,0016*	0,5326
Ruralidad	% de la superficie para tierras de cultivo		0,1335	0,1543	0,3363
	% de la superficie para prados y pastizales	V3	*0,0207*	*0,0028*	0,9630
	% de la superficie para terreno forestal		0,9125	0,8745	0,9697
	% de la superficie no agrícola		0,3945	0,8360	*0,0802*
	% empleados en agricultura, ganadería, silvicultura y pesca	V4	*0,0005*	*0,0003*	0,1042
	% del PIB de agricultura, ganadería, silvicultura y pesca		*0,0524*	*0,0316*	0,5123
Economía	PIB en millones de €		0,2362	0,1076	0,8713
	% del PIB Nacional		0,2364	0,1077	0,8709
	PIB en € por persona	V5	*0,0003*	*0,0008*	*0,0243*
	% parados		0,6088	0,7878	0,4521
Cultura	Proporción de población de 16 y más sin estudios		0,8088	0,6539	0,7844
	Proporción de población de 16 y más con estudios		0,3801	0,3774	0,6405
	Proporción de población de 25 a 34 con estudios		0,5110	0,5336	0,6706
	Prop. de mujeres de 16 y más años analfabeta/sin estudios		0,8557	0,6941	0,7516
	Prop. de varones de 16 y más años analfabeta/sin estudios		0,7252	0,5773	0,8320
	Prop. de mujeres de 16 y más años con estudios superiores		0,3452	0,3359	0,6332
	Prop. de varones de 16 y más años con estudios superiores		0,4288	0,4316	0,6609
	Unidades de educación infantil por 1000 habitantes		0,6877	0,5764	0,9322
	Número de profesores por 1000 habitantes	V6	0,1443	0,6571	*0,0041*
Demografía	Edad media al matrimonio	V7	0,3370	0,2730	0,7945
	Edad media primera maternidad		0,8566	0,7770	0,3113
	Tamaño poblacional	V8	0,3057	0,1264	0,6860
	Densidad poblacional		0,9061	0,6907	0,2802
Relieve	Altitud Máxima		0,5260	0,5465	0,6843
	Rango de altitud		0,1768	0,1878	0,4260
	Altitud media		0,4670	0,6293	0,3825
	Pendiente media	V9	*0,0252*	*0,0087*	0,6130

Tabla 1. P-valor de la regresión lineal simple entre cada una de las variables consideradas y la consanguinidad.

Sólo queremos estudiar la influencia (si existe) de seis factores sobre la consanguinidad, por eso trabajar con todas las variables no tendría sentido y no aportaría más que ruido estadístico al análisis. Por esta razón se han seleccionado solo las variables que representen de forma adecuada los factores que se consideran. El baremo que hemos aplicado es usar solo las variables que correlacionan individualmente al menos con algún tipo de consanguinidad e incluir como máximo dos variables para cada factor. De esta manera se ha conseguido representar a todos los factores excepto para la demografía, donde ninguna variable muestra correlación significativa con ningún tipo de consanguinidad. En la cultura, hemos recurrido a la distribución de las variables y hemos seleccionado dos variables que muestran dos aspectos

demográficos diferentes y que podrían guardar cierta relación con la consanguinidad (Tabla 1). El resultado es que se han seleccionado nueve variables que representan todos los factores cuya influencia en la consanguinidad queremos valorar (Tablas 1 y 2).

Variable		Ft (Coef. de correl.)	Fr (Coef. de correl.)	Fn (Coef. de correl.)
V1	Vehículos por cada 1000 habitantes			
V2	Km de autop. y autov. por cada 1000 km2			
V3	% de la superficie dedicado a prado y pastizal	0,28700	0,31770	
V4	% de empl. en el sect. Agr., Ganad., Silvicult. y pesca			
V5	PIB en € por habitante	-1,05160	-1,16360	-0,84350
V6	Profesores por cada 1000 habitantes	-1,62980		-2,95890
V7	Edad al matrimonio		-3,27660	
V8	Población			
V9	Pendiente media	0,17260	0,31050	
	p-value	0,00004	0,00001	0,00088
	R^2	0,44930	0,48140	0,27360
	AIC	42,72131	51,64900	

Tabla 2. Modelos de regresión múltiple para cada tipo de consanguinidad. Variables incluidas y coeficientes.

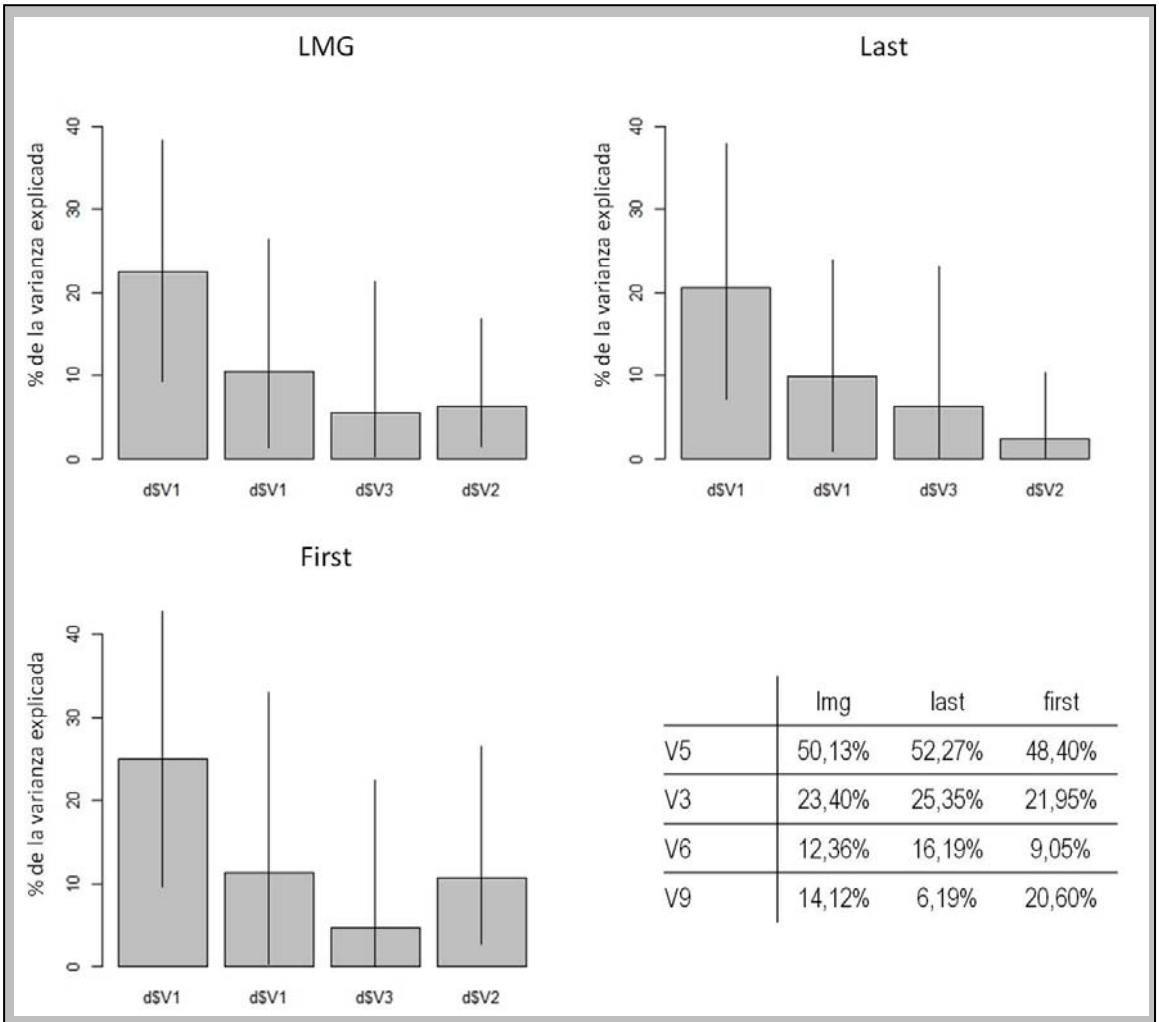
Con las 9 variables seleccionadas se han construido tantos modelos como combinaciones existen usando solo una variable de cada uno de los factores. Los modelos se han realizado mediante una regresión múltiple stepwise. Este método va seleccionando y añadiendo al modelo sólo las variables que son explicativas, de esta forma, sabemos que factores influyen en la consanguinidad y cuáles no. De los modelos construidos se ha seleccionado el mejor para cada tipo de consanguinidad, siguiendo el criterio de información AIC, el BIC y el poder explicativo (Tabla 2).

El resultado puede observarse en la tabla 2. Todos los modelos seleccionados son significativos. La consanguinidad total (Tabla 2, Ft) correlaciona directamente con el porcentaje de la superficie dedicado a prado y pastizal y con la pendiente media; y correlaciona inversamente con el PIB por habitante y el número de profesores que hay por cada 1000 habitantes. La consanguinidad casual (Tabla 2, Fr) correlaciona directamente con la superficie dedicada a

prado y pastizal y con la pendiente media; y negativamente con el PIB por habitante y la edad de los cónyuges en el momento del matrimonio. Por último la consanguinidad no casual (Tabla 2, Fn) correlaciona inversamente con el PIB por habitante y el número de profesores que hay por cada 1000 habitantes.

Pese a que individualmente las variables relacionadas con el aislamiento si correlacionan con la consanguinidad (Tabla 1), no aparecen seleccionadas por ninguno de los modelos.

Consanguinidad total



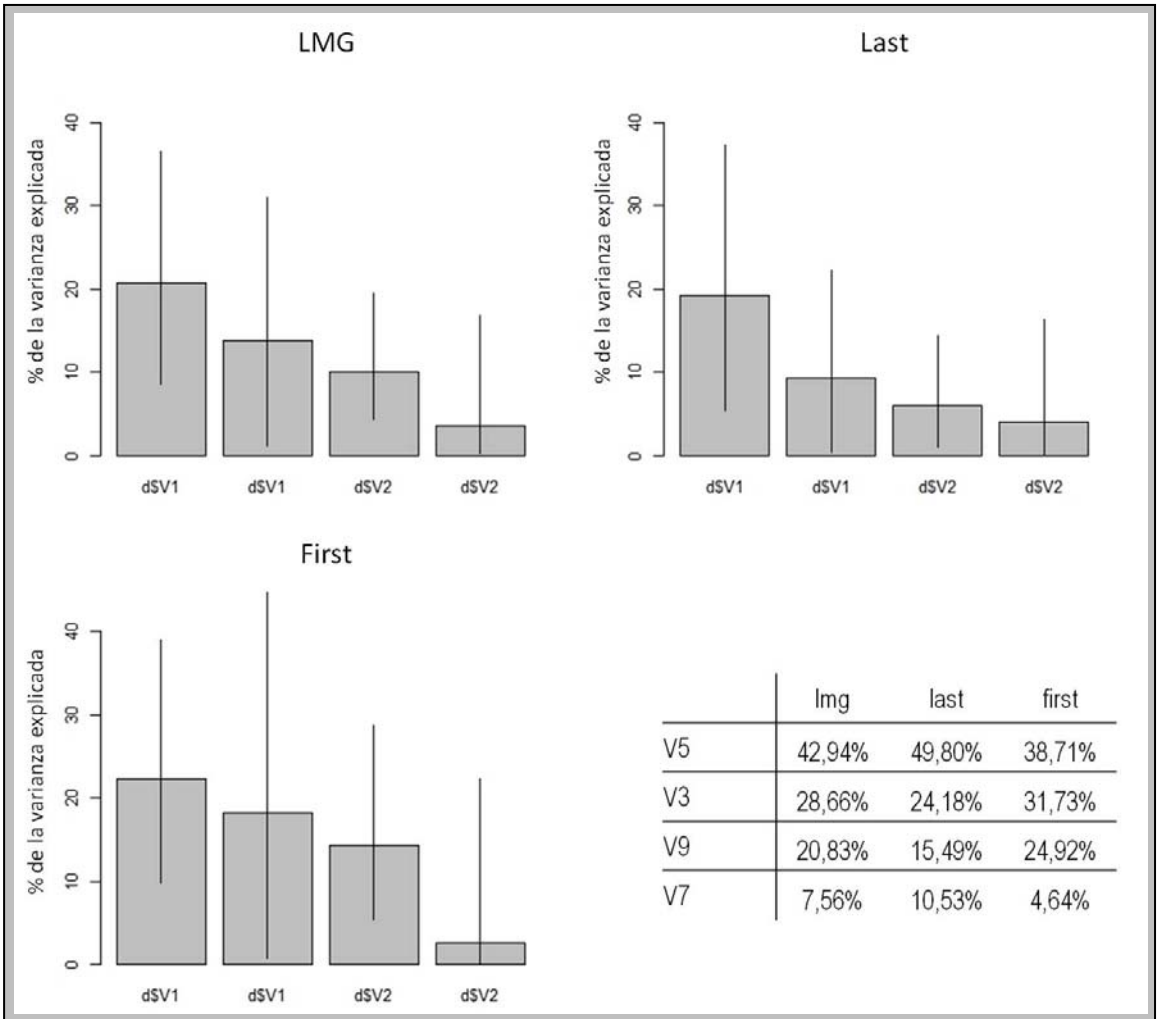
	lmg	last	first
V5	50,13%	52,27%	48,40%
V3	23,40%	25,35%	21,95%
V6	12,36%	16,19%	9,05%
V9	14,12%	6,19%	20,60%

Figura 4. Peso de cada una de las variables explicativas en la varianza explicada para la consanguinidad total (Ft).

En cuanto a la consanguinidad total (Ft) el ambiente rural, nivel económico, nivel cultural y la orografía, explican el 45 % de la variabilidad (Tabla 2). La

variable más explicativa independientemente del método que se use para estimarlo es el PIB en euros por persona (Figura 4), que cubre por sí sola entre el 52 % y el 48 % de la variabilidad explicada. La segunda variable más importante es el porcentaje de superficie que se destina a prado y pastizal que explica entre el 25 % y el 22 % de la variabilidad (Figura 4). Menos peso tienen el número de profesores por 1000 habitantes (entre el 16 % y el 9 %) y la pendiente media (entre el 20 % y el 6 %).

Consanguinidad casual



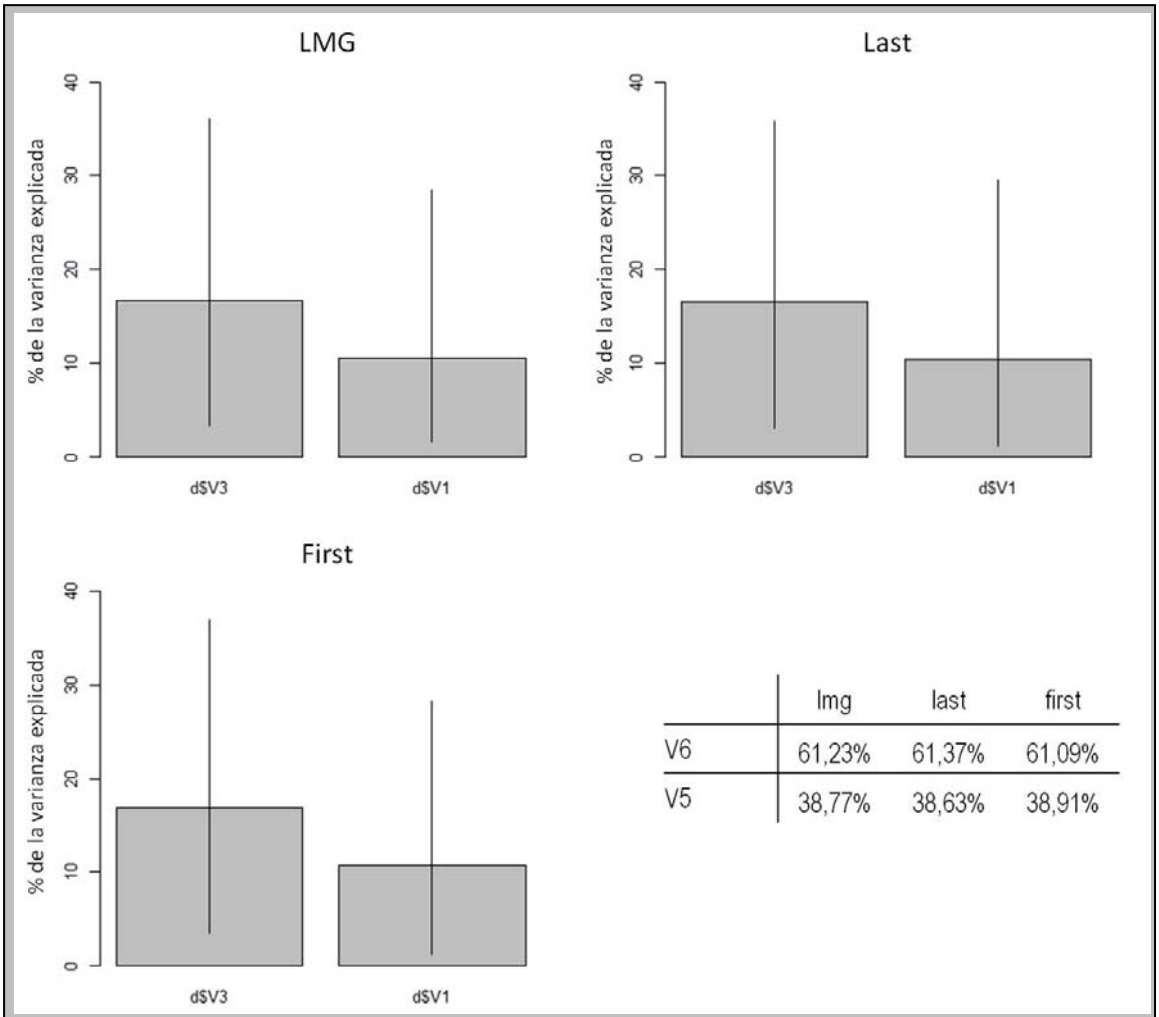
	lmg	last	first
V5	42,94%	49,80%	38,71%
V3	28,66%	24,18%	31,73%
V9	20,83%	15,49%	24,92%
V7	7,56%	10,53%	4,64%

Figura 5. Peso de cada una de las variables explicativas en la varianza explicada para la consanguinidad casual (Fr).

El modelo explicativo para la consanguinidad casual alcanza a explicar el 48 % de la variabilidad (Tabla 2). La principal variable es el PIB en euros por persona (Figura 5), que cubre entre el 50 % y el 39 % de la variabilidad explicada. La

segunda variable es la superficie dedicada a prado y pastizal que explica entre el 32 % y el 25 % (Figura 5). Menos importantes son el relieve (entre el 25 % y el 15%) y la edad de los cónyuges al matrimonio (entre el 11 % y el 8 %).

Consanguinidad no casual



	lmg	last	first
V6	61,23%	61,37%	61,09%
V5	38,77%	38,63%	38,91%

Figura 6. Peso de cada una de las variables explicativas en la varianza explicada para la consanguinidad no casual (Fr).

Por último, la consanguinidad no casual sólo aparece relacionada con dos variables, PIB en euros por persona y profesores por cada mil habitantes. Es el modelo menos explicativo de los tres, alcanza el 27 % de la variabilidad (Tabla 2). La variable más explicativa (Figura 6) es el número de profesores por cada mil habitantes (el 61 %) por delante del PIB (39 %).

Con el objetivo de comprender realmente como funciona todo el sistema de variables, se han analizado las relaciones entre ellas.

Lo más destacable de todas estas interacciones es que no todas las parejas de variables que caracterizan a un mismo factor correlacionan entre sí. Así, mientras las variables de aislamiento (vehículos por cada mil habitantes y kilómetros de autovía por cada mil kilómetros cuadrados) si correlacionan, las de carácter rural (superficie destinada a prado y pastizal y porcentaje de empleados en agricultura, ganadería y pesca) no lo hacen (Figura 7).

El porcentaje de empleados en agricultura (carácter rural) correlaciona positivamente con la pendiente (orografía) y negativamente con los kilómetros de autovía por cada mil kilómetros cuadrados (aislamiento), el PIB en euros por persona (nivel económico) y con el tamaño poblacional (demografía). Por otro lado, el nivel económico (PIB en euros por persona) correlaciona positivamente con las dos variables de aislamiento (vehículos por cada mil habitantes y kilómetros de autovía por cada mil kilómetros cuadrados) y con el tamaño poblacional (Figura 7).

Por último, resulta interesante señalar también que el número de profesores por cada mil habitantes (nivel cultural) correlaciona directamente con la edad de los cónyuges en el momento del matrimonio.

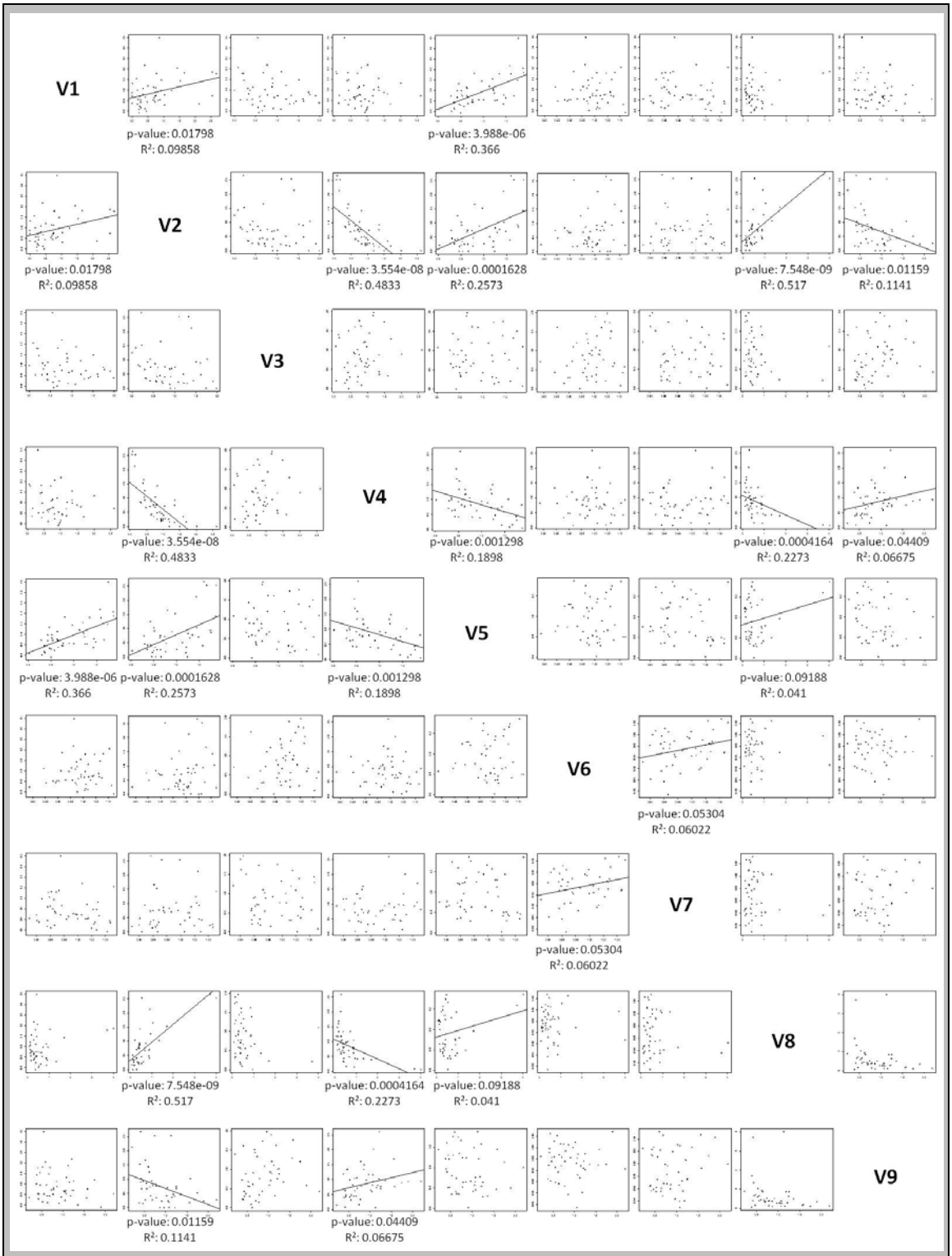


Figura 7. Interacción entre todas las variables utilizadas. En aquellas en las que existe correlación se ha indicado.

Discusión

La población española posee una serie de características muy interesantes para un estudio de consanguinidad. En primer lugar, tradicionalmente, la consanguinidad se ha estudiado en poblaciones limitadas en las que puede controlarse toda la población. Sin embargo, también disponemos en la bibliografía de algunos análisis de la consanguinidad a partir de apellidos a más amplio nivel (Rodríguez-Larralde, 2003; Scapoli et al., 2007; Dipierri et al., 2011; Rodríguez-Larralde, 2011). Pero estos estudios se limitan solo a estimar una de las componentes de la consanguinidad, lo que resulta una estima meramente aproximativa a la consanguinidad de una población (Scapoli et al., 2007).

Como ya hemos mencionado, la población española presenta unas características que la hacen de especial interés en el estudio de la consanguinidad. El sistema español de apellidos (Pettener et al., 1998; Fauré et al., 2001; Colantonio et al., 2003; Blanco-Villegas et al., 2004; Mateos y Tucker, 2008; Dipierri et al., 2011), la homogeneidad cultural y religiosa y la elevada consanguinidad (Scapoli et al., 2007), permiten extraer de esta población unas conclusiones particularmente interesantes y ampliamente extrapolables.

Isonimia

Pese a que la consanguinidad es un fenómeno en declive en las poblaciones europeas (Pettener, 1985; Fuster y Colantonio, 2002; Bittles y Black, 2010), sigue persistiendo. De hecho, la población española presenta una de las consanguinidades más elevadas del continente (Rodríguez-Larralde et al., 2003; Scapoli et al., 2007).

Su distribución es desigual en el territorio nacional y parece estar sujeta a la influencia de ciertos factores que condicionan su distribución geográfica. En líneas generales coincide con lo encontrado anteriormente por Rodríguez-Larralde (Rodríguez-Larralde et al., 2003). En ese trabajo la base de datos está

construida en base a una fuente (registros telefónicos) y una división geográfica diferente a la nuestra. Pese a ello, la concordancia es notable.

Para conocer mejor la distribución se han analizado y representado por separado la distribución geográfica de cada componente de la consanguinidad.

La consanguinidad total (Ft) en el territorio peninsular español (Figura 1), en principio, y como se había observado en trabajos previos (Fuster y Colantonio, 2002) no permite vislumbrar un gradiente o una relación clara con la distribución de otros factores como la densidad poblacional, el nivel económico o los factores geográficos. No resulta sorprendente sin embargo (Rodríguez-Larralde et al., 2003), que las zonas de mayor consanguinidad, coinciden casi exactamente con las zonas de menor diversidad genética identificadas anteriormente.

Respecto a la consanguinidad casual (Fr) la distribución muestra los niveles más altos (Figura 2) en Asturias, Galicia, León y el sur de la meseta central (Salamanca, Ávila, Zamora y Segovia), zonas que sustentan los altos niveles de consanguinidad de la población española en comparación al resto de Europa (Scapoli et al., 2007). La coincidencia con la diversidad genética es aún más clara que en el caso de la consanguinidad total, algo que era de esperar considerando que esta componente depende sólo de la composición de apellidos de la población.

La consanguinidad no casual (Fn) tiene una distribución (Figura 3) más irregular que la consanguinidad total (Ft) o la casual (Fr), parece que depende de la influencia de unos factores diferentes. Que la consanguinidad total (Ft) tenga una distribución más parecida a la consanguinidad casual (Fr) que a la no casual (Fn) obedece al hecho de que esta última contribuye en mucha menor medida al nivel total (Crow y Mange, 1965; Crow, 1980)

Análisis de la isonimia

Para tratar de comprender los factores que condicionan las diferencias en los niveles de consanguinidad, se han explorado 31 variables diferentes (Tabla 1), que tratan de caracterizar a factores que suelen estar relacionados de manera individual con la consanguinidad (Calderón, 1989; Jorde y Pitkanen, 1991; Fuster et al., 1996; Fuster y Colantonio, 2002; Blanco-Villegas et al., 2004; Rodríguez-Diaz y Blanco-Villegas, 2011). El hecho de valorar tantas variables ha demostrado ser muy importante. No todas las variables caracterizan a los factores que queremos estudiar como cabría esperar (Tabla 1). No considerar todas las variables nos podría haber hecho pasar por alto la influencia de un determinado factor sencillamente por no haber seleccionado la variable más adecuada, mientras que seleccionando las más adecuadas podemos estar seguros de haber caracterizado apropiadamente los factores que vamos a considerar.

En una primera aproximación (Tabla 1), parece que la consanguinidad total (Ft) está relacionada con el aislamiento, el ambiente rural, el nivel económico, el nivel cultural y la orografía. Todos esos factores están correlacionados también con la consanguinidad casual (Fr), por lo que parecen condicionar los niveles de consanguinidad total mediante la presión ambiental (Fr) sufrida por la población. Las únicas excepciones son el nivel económico y el nivel cultural, que parecen afectar a la consanguinidad condicionando la actitud social frente a la consanguinidad (Calderón, 1989; Fuster y Colantonio, 2002; Rodríguez-Diaz y Blanco-Villegas, 2011).

Sin embargo las poblaciones no sufren la influencia aislada de un sólo factor, debemos considerar la influencia de todos conjuntamente (Tabla 2).

Consanguinidad total

Los niveles de consanguinidad total (Ft) en la población española obedecen a la influencia conjunta del nivel de ruralidad, el nivel económico, el nivel educacional y la orografía (Tabla 2).

La variable con mayor trascendencia para la consanguinidad es el nivel económico en forma de PIB en euros por habitante (Figura 4), a mayor nivel económico menor consanguinidad de forma similar a lo observado en poblaciones rurales españolas (Fuster y Colantonio, 2002). Es responsable de entorno 50 % de la variabilidad explicada, pesa tanto como los otros tres factores juntos.

En un trabajo previo con poblaciones españolas Fuster y Colantonio (Fuster y Colantonio, 2002) detectaron grandes diferencias en el nivel de consanguinidad entre entornos urbanos y entornos rurales. Sin embargo parece ser el segundo factor (Figura 4) en importancia (entre el 25 y el 23 % de la variabilidad). Por la forma en la que interactúan todas las variables entre sí, sabemos que las zonas más rurales tienen también menor nivel económico (Figura 7), pero el carácter rural aparece en el modelo pese a esa relación con el nivel económico, por lo que más allá de la relación con el nivel económico el carácter rural ejerce su propia influencia.

Al nivel educacional le corresponde entre el 16 y el 9 % de la variabilidad de la consanguinidad (Figura 4). El nivel educacional no parece guardar relación con ninguno de los otros factores, salvo la edad al matrimonio (Figura 7). Parece que cuanto menor es el nivel educacional, menor es también el esfuerzo invertido en la búsqueda de pareja y la edad al matrimonio (Goldstein y Kenney, 2001).

La última variable explicativa es la orografía (entre el 20 y el 6 % de la variabilidad). La orografía está muy relacionada con la ruralidad y aún así ha entrado en el modelo (Figura 7), luego también aporta poder explicativo por sí

misma. Tal y como se señala en otros estudios, la orografía complicada aumenta el nivel de consanguinidad (Fuster y Colantonio, 2002). Probablemente, por lo que sabemos de su relación con el resto de variables, la forma de actuar de la orografía sea dificultar las comunicaciones y el poblamiento, haciendo las poblaciones menos atractivas y rentables (Figura 7).

El aislamiento, que correlaciona individualmente con la consanguinidad (Tabla 1), no aparece como factor en el modelo (Tabla 2). La razón es que es una característica común de casi todos los factores seleccionados (Figura 7), las zonas con menor nivel económico, mayor carácter rural y orografía más complicada son zonas más aisladas. Aunque individualmente el aislamiento si pueda estar relacionado con la consanguinidad, parece que ejerce su influencia conjuntamente con otra serie de variables con las que está muy relacionado.

Consanguinidad casual

La mayor aportación a la consanguinidad total la realiza su componente casual (Fr). La consanguinidad casual se debe a la influencia ambiental (Crow, 1980), por lo que los factores que condicionan la consanguinidad casual (Fr) son el reflejo de la influencia ambiental en la consanguinidad total.

El factor más determinante en la consanguinidad casual (Figura 5) es el nivel económico (entre el 50 y el 39 %). Más allá de condicionar actitudes sociales respecto a la consanguinidad (Pettener, 1985; Calderón, 1989), el nivel económico ejerce una influencia ambiental, probablemente en forma de empobrecimiento poblacional, reduciendo los atractivos poblacionales y con ello los tamaños poblacionales, los flujos y aumentando el aislamiento (Figura 7).

El segundo factor es la orografía (Figura 5), tal y como el carácter rural (entre el 32 y el 24 %). Que tal y como se ha comentado antes, tiene una influencia ambiental sobre la consanguinidad.

El último factor es el relieve (entre el 10 y el 5 %) de la manera en la que ya se ha visto (Figura 5). A esos factores hay que añadir una característica demográfica, la edad al matrimonio (Figura 5), que suele estar muy relacionada con el esfuerzo en la búsqueda de pareja (Goldstein y Kenney, 2001).

Consanguinidad no casual

Menos importancia sobre el nivel total de consanguinidad tiene la componente no casual (Fn). La componente no casual implica una determinada actitud social de la población hacia las uniones entre consanguíneos (Crow y Mange, 1965; Crow, 1980) y, como sería de esperar, los factores de los que depende también son diferentes (Figura 6). En determinadas poblaciones se ha observado una actitud favorable hacia las uniones entre parientes con el objetivo de retener el patrimonio familiar (Pettener, 1985; Calderón, 1989).

En torno al 61 % de la variabilidad explicada se debe al nivel educativo (mayor nivel educativo, menor consanguinidad) y en torno al 39 % del nivel económico (mayor nivel económico, menor consanguinidad). Parece que la actitud social hacia la consanguinidad depende tanto de un equilibrio entre los costes y los beneficios (Bittles y Black, 2010) como de la percepción de los costes de la consanguinidad que aporta a una sociedad una adecuada educación.

El papel del aislamiento en la consanguinidad es difícil de interpretar (Calderón, 1989; Jorde y Pitkanen, 1991; Fuster et al., 1996; Fuster y Colantonio, 2002). Su relación individual y directa con la consanguinidad es evidente (Tabla 1) pero su influencia no aparece cuando se considera el conjunto de factores. A la luz de los resultados, parece que en realidad el aislamiento actúa por medio de otros factores (Figura 7). Así, las comarcas más rurales, las de menor nivel económico y las de orografía más complicada, son también las más aisladas, y estos tres factores si influyen en la consanguinidad.

Conclusión

Parece que un bajo nivel económico es el principal factor ligado a un aumento en la consanguinidad de la población. Esta influencia se produce tanto actuando sobre la propia población, probablemente en forma de una falta de incentivos que ha condicionado la estructura de la población; como sobre la actitud de la población hacia la consanguinidad, viéndola con mejores ojos con el fin de conservar el patrimonio.

El ambiente rural y la complicada orografía ejercen una importante presión ambiental sobre la población. Condicionan su estructura, con menos disponibilidad de no emparentados aumentando así la consanguinidad.

El nivel educacional, es el principal factor que condiciona la actitud social frente a la consanguinidad. Condiciona la percepción que se tiene de los riesgos que supone la consanguinidad.

Bibliografía

Bener A, Dafeeah EE, Samson N (2012) The Impact of Consanguinity on Risk of Schizophrenia. Psychopath 45: 399–400.

Bittles AH (2009) Consanguinity, genetic drift, and genetic diseases in populations with reduced numbers of founders. Hum Genetics: Principles and Approaches. Springer, Heidelberg.

Bittles AH, Black ML (2010) Consanguinity, human evolution, and complex diseases. PNAS 107: 1779-1786.

Bittles AH, Neel JV (1994) The costs of human inbreeding and their implications for variations at the DNA level. Nat Genet 8: 117 - 121.

Blanco-Villegas MJ, Boattini A, Otero HR, Pettener D (2004). Inbreeding patterns in La Cabrera, Spain: dispensations, multiple consanguinity analysis, and isonymy. Hum Biol 76: 191-210.

Bundey S, Alam H (1993) A 5-year prospective study of the health of children in different ethnic groups, with particular reference to the effect on inbreeding. Eur J Hum Genet 1: 206–219.

Calderon R (1989) Consanguinity in the Archbishopric of Toledo, Spain, 1900-79. I. Types of consanguineous mating in relation to premarital migration and its effects on inbreeding levels. J Biosoc Sci 21: 253-66.

Cavalli-Sforza LL, Moroni A, Zei G (2004) Consanguinity, Inbreeding, and Genetic Drift in Italy. Princeton Univ Press, Princeton.

Colantonio SE, Lasker GW, Kaplan BA, Fuster V (2003) Use of surname models in human population biology: a review of recent developments. Hum Biol 75:785-807.

Crow JF (1980) The estimation of inbreeding from isonymy. Hum Biol 52: 1-14.

Crow JF, Mange AP (1965) Measurement of inbreeding from the frequency of marriages between persons of the same surname. Eugen Q 12: 199-203.

Dipierri J, Rodríguez-Larralde A, Alfaro E, Scapoli C, Mamolini E, et al. (2011) A Study of the Population of Paraguay through Isonymy. Ann Hum Genet 75: 678-687.

Dipierri JE, Alfaro EL, Scapoli C, Mamolini E, Rodríguez-Larralde A, et al. (2005) Surnames in Argentina: a population study through isonymy. Am J Phys Anthropol 128: 199-209.

Faure R, Ribes MA, García A (2001) Diccionario de apellidos españoles. Madrid: Espasa-Calpe.

Fuster V, Colantonio SE (2002) Consanguinity in Spain: socioeconomic, demographic, and geographic influences. Hum Biol 74: 301-315.

Fuster V, Morales B, Mesa MS, Martín J (1996) Inbreeding patterns in Gredos mountain range (Spain). Hum Biol 68:75–93.

Goldstein JR, Kenney CT (2001). Marriage delayed ormarriage forgone?New cohort forecasts of first marriage for U.S. Women. Am Sociol Rev 66: 506 – 519.

Graham J, McNeney B and Seillier-Moiseiwitsch F (2005). Stepwise detection of recombination breakpoints in sequence alignments. Bioinformatics, 21: 589-595

Grömping U (2006). Relative Importance for Linear Regression in R: The Package relaimpo. J Stat Softw, 17: 1-27.

Helgason A, Pálsson S, Guöbjartsson DF, Kristjánsson T, Stefánsson K (2008) An Association Between the Kinship and Fertility of Human Couples. Science, 319: 813-6.

Jorde L, Pitkanen K J (1991) Inbreeding in Finland. Amer J Phys Anthropol 84: 127–139.

Mansour H, Fathi W, Klei L, Wood J, Chowdari K, Watson A, Eissa A, Elassy M, Ali I, Salah H, Yassin A, Tobar S, El-Boraie H, Gaafar H, Ibrahim NE, Kandil K, El-Bahaei W, El-Boraie O, Alatrouny M, El-Chennawi F, Devlin B, Nimgaonkar VL (2010) Consanguinity and increased risk for schizophrenia in Egypt. Schizophr Res 120: 108-12.

Mansour H, Klei L, Wood J, Talkowski M, Chowdari K, Fathi W, Eissa A, Yassin A, Salah H, Tobar S, El-Boraie H, Gaafar H, Elassy M, Ibrahim NE, El-Bahaei W, Elsayed M, Shahda M, El Sheshtawy E, El-Boraie O, El-Chennawi F, Devlin B, Nimgaonkar VL (2009) Consanguinity associated with increased risk for bipolar I disorder in Egypt. Am J Med Genet B Neuropsychiatr Genet 150: 879–885.

Pettener D (1985) Consanguineous marriages in the upper Bologna Appenine (1565-1980): microgeographic variation, pedigree structure and correlation of inbreeding secular trend with changes in population size. Hum Biol, 57: 267-288.

Pettener D, Pastor S, Tarazona-Santos E (1998) Surnames and genetic structure of a high-altitude quechua community from the Ichu river valley peruvian central Andes 1825-1914. Hum Biol 70: 865-87.

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL http://www.R-project.org/.

Rittler M, Liascovich R, López-Camelo J, Castilla EE (2001) Parental consanguinity in specific types of congenital anomalies. Am J Med Genet 102: 36–43.

Rodríguez-Larralde A, Dipierri J, Gómez EA, Scapoli C, Mamolini E, et al. (2011) Surnames in Bolivia: A study of the population of Bolivia through isonymy. Am J Phys Anthropol 144: 177-184.

Rodríguez-Larralde A, Gonzales-Martin A, Scapoli C, Barrai I. 2003. The names of Spain: a study of the isonymy structure of Spain. Am J Phys Anthropol 121: 280-292.

Sakamoto Y, Ishiguro M, Kitagawa G (1986). *Akaike Information Criterion Statistics*. D. Reidel Publishing Company.

Scapoli C, Mamolini E, Carrieri A, Rodríguez-Larralde A, Barrai I (2007) Surnames in Western Europe: a comparison of the subcontinental populations through isonymy. Theor Popul Biol 71: 37-48.

Shami SA, Qaisar R, Bittles AH (1991) Consanguinity and adult morbidity in Pakistan. Lancet 338: 954–955.

Ulrike Grömping (2006). Relative Importance for Linear Regression in R: The Package relaimpo. J Stat Softw, 17: 1-27.

Vézina H, Heyer É, Fortier I, Ouellette G, Robitaille Y, Gauvreau, D (1999) A genealogical study of Alzheimer disease in the Saguenay region of Québec. Genet Epidemiol 16:412–425.

"El ser humano es maravilloso,

... pero para un ratito."

(Javier Cansado)

5. *CONCLUSIONES*

CONCLUSIONES

1. La población española está estructurada en dos grandes grupos, las provincias del Norte-Oeste y las del Sur-Este. Esta estructura se repite recurrentemente a lo largo de la historia de España, lo que induce a pensar que son los factores geográficos, siempre presentes, los que han determinado esta organización. Otros factores, como los etno-lingüísticos, que diferencian al País Vasco, parecen tener su influencia a un nivel más local, en la organización interna de cada uno de los grandes grupos.
2. En el territorio español peninsular se pueden distinguir tres tipos de movimientos, de corta, de media y de larga distancia. Los de corta y media distancia son los movimientos de mayor importancia y, por lo tanto, los que más han podido contribuir a la actual estructura de la población española. Estos movimientos se producen preferentemente en el interior de la gran fragmentación territorial que hemos identificado y siguen principalmente los arcos atlántico (en el Norte-Oeste) y mediterráneo (en el Sur-Este).
3. A nivel peninsular la estructura que regionaliza a la población española, es muy semejante a la estructura de la diversidad de fauna vertebrada. Este hecho junto a las diferencias climáticas que existen entre ambas zonas geográficas invita a pensar que en la estructura genética de la población española la influencia de los factores biogeográficos ha podido ser determinante. La organización interna de la estructura genética dentro de cada uno de los dos grandes grupos difiere ligeramente de la estructura de la diversidad de fauna vertebrada. Estas divergencias parecen relacionadas con factores etno-lingüísticos en el caso del Galicia, País Vasco, Cataluña y Valencia e históricos en las provincias que conforman la Ruta de la Plata.
4. La consanguinidad de la población española está determinada por la presión ambiental y la actitud social. La consanguinidad derivada de la presión ambiental obedece a la influencia del carácter rural y la orografía. La consanguinidad derivada de la actitud social se encuentra relacionada con el nivel educacional. El nivel económico se revela como la variable de mayor relevancia dado que se muestra vinculada a ambas presiones.