Software "*isonymic*" para la aceleración de estudios poblacionales a partir de apellidos

Morales Arturo Leonardo[1,2][0000−0002−3980−8862], Toledo Margalef Pablo[1,2][0009−0008−8099−2578], Navarro Jose Pablo[1,2][0000−0003−2180−449X], Ramallo Virginia[1][0000−0002−7856−4856], and Delrieux Claudio[3][0000−0002−2727−8374]

[1] Instituto Patagónico de Ciencias Sociales y Humana (IPCSH), CONICET-CENPAT, Puerto Madryn, Argentina
https://cenpat.conicet.gov.ar/ipcsh/
[2] Departamento de Informática Trelew (DIT), Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB), Trelew, Argentina
https://www.ing.unp.edu.ar/dpto-informatica
[3] Departamento de Ingeniería Eléctrica y de Computadoras, UNS, Bahía Blanca, Argentina. http://www.diec.uns.edu.ar

Resumen. Este trabajo presenta una biblioteca de software liviana denominada "*isonymic*", diseñada para el lenguaje de programación Python, con la finalidad de agilizar la obtención de información relativa a la estructura demográfica de una población a partir de sus apellidos. El método subyacente para adquirir dicha información se conoce como método isonímico y tiene el potencial de ejercer un impacto significativo en la investigación de la salud de las poblaciones. Los apellidos proporcionan información relevante acerca de las relaciones familiares y de grupos más extensos, que pueden vincularse con información sanitaria y epidemiológica sobre prevalencia e incidencia de ciertas enfermedades y los factores socioeconómicos concomitantes. El paquete propuesto permite obtener rápidamente los cinco indicadores isonímicos principales y facilita el cálculo de distancias entre poblaciones según las frecuencias de sus apellidos. También posibilita calcular en forma sencilla los principales estadísticos isonímicos relacionados con fenómenos migratorios (μ de Karlin McGregor, m de Wright, generación de gráficas log-log, asignación de etiquetas según origen geo-lingüístico y análisis espacial). En este trabajo se exhiben los resultados que se pueden obtener utilizando la interfaz de programación proporcionada por el software *isonymic*, empleando conjuntos de apellidos recolectados de padrones electorales argentinos correspondientes a dos períodos de tiempo diferentes.

Palabras clave: Nombres de familia · Isonímia · Paquete Python.

1 Introducción

Los apellidos han sido extensamente utilizados para comprender fenómenos poblacionales en todo el mundo. El trabajo principal en este campo fue propuesto por

los profesores James F. Crow y Arthur P. Mange, en 1965. En dicho aporte, se presentó una primera versión de un estimador de endogamia obtenido a partir del análisis de información de matrimonios isonímicos y de las frecuencias de portadores de apellidos idénticos en una población. Los matrimonios isonímicos son aquellos en donde los dos integrantes de la pareja tienen el mismo apellido. A partir de allí, el procedimiento que se centra en el análisis de las frecuencias de los apellidos de una población se conoce como método isonímico. La potencia de este análisis reside en lo económico de su aplicación, dado que se centra en el examen de datos ampliamente disponibles en la actualidad, como padrones electorales, registros bautismales, de registros civiles, etc. Asimismo, presenta una gran versatilidad, ya que es factible de aplicar incluso en situaciones donde la información genealógica es escasa o inexistente. También supone un alcance notable, ya que posibilita trabajar con conjuntos de datos de poblaciones tanto del pasado reciente como remoto. A partir del conjunto de indicadores que se generan utilizando los apellidos, se ha podido estimar la estructura genética de la población [14, 16, 15], la movilidad social [17], la segregación étnica [18, 19] y las desigualdades en salud [20, 21], entre otras tantas aplicaciones.

En este trabajo, presentamos un paquete software para el lenguaje de programación Python que busca acelerar la obtención de esta información isonímica en situaciones donde se conozcan los apellidos portados por cada individuo de la población que se encuentre en análisis. En la siguiente sección se describirán muy brevemente los indicadores demográficos calculables a partir de apellidos. Luego se presentará el paquete isonymic y la manera para obtener los indicadores y reportes introducidos previamente. Acto seguido se dispondrán algunos resultados alcanzados utilizando el paquete. Y por último, se expondrán las conclusiones y las líneas futuras en busca de la evolución del software presentado.

2 Estructura demográfica a partir de apellidos

Los indicadores se utilizan para caracterizar unidadades territoriales dentro de los países, como puede ser una provincia, un departamento, una región, una localidad o incluso una sección de una localidad.

El indicador principal en este campo, es el de isonimia I, que mide la proporción de las parejas de individuos que comparten el mismo apellido en una población, tomando pares aleatorios de apellidos. Por regla general, se ignora la distinción de los sexos y directamente se realiza la sumatoria de las proporciones de cada apellido considerando todas las personas de una población. Por lo tanto el indicador I puede estimarse como $I = \sum x_i^2$, siendo x_i la frecuencia relativa del i-ésimo apellido de la población [1]. A partir del valor I se construyen dos indicadores más. Por un lado el estimador de endogamia F_{st}, que se calcula con la fórmula $F_{st} = \frac{I}{4}$. Y otro indicador, denominado α de Fisher, que mide la riqueza en apellidos y se estima mediante la fórmula $\alpha = \frac{1}{I}$ [3]. Un valor bajo de α puede representar alta endogamia y deriva, mientras que un valor elevado podría significar migración y baja endogamia [4].

Adicionalmente, cuando se analizan las frecuencias de los apellidos registrados en una población, podemos obtener dos indicadores adicionales: el indicador A y el indicador B. El indicador A representa el porcentaje de la población que es única portadora de su apellido y permite observar el aislamiento y sedentarismo [5]. Para interpretar correctamente el significado de la magnitud de este indicador es necesario contar con información de contexto, ya que valores altos podrían producirse tanto por inmigración como por emigración. Como regla general, un aumento en su valor se relaciona con una intensificación en el movimiento de la población. El indicador B representa el porcentaje de individuos cuyo apellido se encuentra dentro de los siete más frecuentes en la población en observación [5]. Este, es un indicador de aislamiento relativo, en donde valores más altos representan sedentarismo, debido a que no reciben nuevos apellidos por migración.

El estimador I también se puede adaptar para medir la relación existente entre poblaciones, a partir de la frecuencia de los apellidos compartidos. Es posible entonces obtener un indicador de isonimia aleatoria entre dos poblaciones, aplicando la fórmula

$$I_{ij} = \sum p_{ik} p_{jk}, \tag{1}$$

en donde p_{ik} y p_{jk} son la frecuencia del apellido k en la población i y la población j respectivamente. A partir de esta distancia isonímica se derivan otras tres medidas:

1. Distancia de Lasker. Según [6] se define con la fórmula:

$$L = -\log(I_{ij}). \tag{2}$$

2. Distancia Euclídea entre dos grupos [7]. Se define según la fórmula:

$$E = \sqrt{1 - \sum_k \sqrt{p_{ki} p_{kj}}}. \tag{3}$$

3. Distancia de Nei. Según [8] se define con la fórmula:

$$N_d = -\log\left(\frac{I_{ij}}{\sqrt{I_{ii} I_{jj}}}\right). \tag{4}$$

En cuanto al analisis de migraciones humanas a partir de apellido, se destacan dos indicadores mas. Por un lado el indicador μ, que resume la tasa de migración reciente [12, 13]. Para su obtención, se necesita el datos demograficos referido a la cantidad de habitantes de una unidad territorial. Entonces, se calcula con la fórmula propuesta en [11] de la siguiente forma:

$$\mu = \frac{\alpha}{(N + \alpha)}, \tag{5}$$

donde α corresponde al α de Fisher y N a la cantidad de personas en población en estudio.

Por otro lado, el indicador m de Wright [9] estima la proporción de intercambio de apellidos o tasa de migración por generación usando la fórmula:

$$m = 1 - \sqrt{\frac{2N_e F_{st}}{(2N_e - 1)\, F_{st} + 1}}, \tag{6}$$

donde N_e es el tamaño efectivo de la población, o sea la población que contribuye efectivamente a la próxima generación y F_{st} corresponde al coeficiente de endogamia por isonimia al azar antes descrito. De acuerdo a Nei e Imaizumi [10] el N_e puede ser obtenido a partir de la relación $N_e = N * 0,65$, siendo N el tamaño censal del departamento, provincia o región considerada.

Un ultimo dispositivo visual de amplia utilización en el campo de apellidos es el gráfico log-log. Esta gráfica se basa en el hecho de que la distribución de apellidos sigue una dinámica caracterizada por una ley de potencia, donde la frecuencia de un apellido guarda una relación inversamente proporcional con su posición en el ranking de frecuencias. A partir de esto, este instrumento representa la relación entre la frecuencia de ocurrencia de los apellidos (en el eje de abscisas) y el número de apellidos que se repiten exactamente esa misma cantidad de veces (eje de ordenadas) (Fox and Lasker, 1983). Sobre ambos ejes se aplica la operación logarítmica para facilitar su interpretación visual, de allí su denominación. Analizando la concavidad de la línea ajustada a los puntos de la gráfica es posible obtener información que facilite la detección de migración, deriva y aislamiento. Es interesante obtener este gráfico de forma muy flexible para observar su comportamiento según se modifique el conjunto de apellidos de trabajo. Esta diversidad en la composición puede estar influenciada por el género de la población (representada gráficamente con distinción entre hombres y mujeres), la popularidad de los apellidos (mediante el análisis de los 50 o 100 más comunes), e incluso comparaciones entre unidades administrativas dentro de un país.

3 Paquete *isonymic*

Como se pudo apreciar en la sección anterior, los indicadores isonímicos son fáciles de calcular pero abundantes en cantidad. El paquete propuesto para el lenguaje de programación Python, se trata de una biblioteca liviana que expone operaciones para obtener estos indicadores, ya sea por separado o todos de una vez, a partir de una lista de apellidos cualquiera. La Tabla 1 lista las funciones disponibles y una breve descripción de las tareas que llevan a cabo.

El paquete se encuentra publicado en el repositorio de software para Python Package Index "PyPI" (Fig. 1).

4 Resultados

Hemos utilizado *isonymic* para analizar la estructura demográfica a partir de apellidos en distintos niveles de la organización administrativa de la Argentina

Fig. 1. Portada del paquete "isonymic" en el repositorio indexado de paquetes para el lenguaje de programación Python, conocido como PyPI (Python Package Index).

(escala nacional, regional, provincial y departamental) y para tres puntos en el tiempo basados en los registros electorales a los que el grupo de investigación tuvo acceso (2001, 2015 y 2021). En este entramado, algunos niveles organizacionales incluyen poblaciones muy grandes e involucran cantidades masivas de datos, mientras que, por otro lado, estudios sobre unidades administrativas más pequeñas requieren replicaciones ordenadas del camino exploratorio y de los cálculos involucrados para confluir en el análisis integral final.

El paquete *isonymic* nos permitió rápidamente evaluar la tendencia de cada uno de los indicadores en el período 2001-2021 en los tres momentos temporales indicados mas arriba. A partir de los datos en esta escala departamental, es posible la aceleración de la confección de mapas para visualizar los patrones geográficos de cada indicador, bajo una perspectiva migratoria (Fig 2).

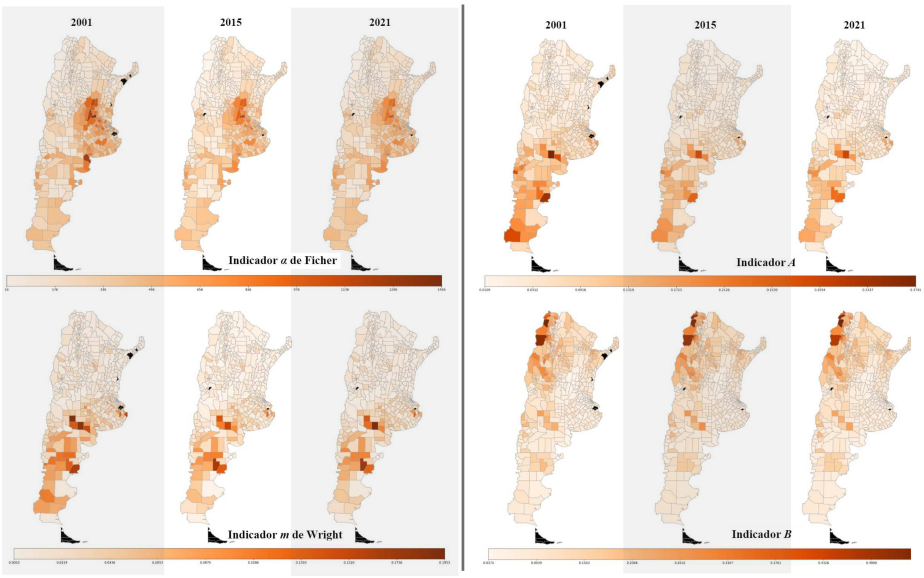

Fig. 2. De izquierda a derecha, mapas coropléticos para los valores departamentales de los indicadores α de Fisher, A, m de Wright y B a partir de los 529 valores departamentales obtenidos con *isonymic*.

Por último, utilizando la función get_surname_frequencies de *isonymic* se experimento en distintas escalas administrativas respetando la jerarquía establecida país-región-provincia-departamento. Se generaron funciones de utilidad para la rápida creación de visualizaciones efectivas entre los que se encontraban gráficos log-log. Estas gráficas se adaparon para presentar no sólo la pendiente de la diferencia entre cantidad de portadores de los apellidos más frecuentes y los menos frecuentes, sino también señalar explícitamente cuáles son dichos apellidos. La Fig 3 muestra los gráficos log-log para las cinco regiones de la Argentina y para el país en general. La flexibilidad de esta gráfica que ha sido mencionada mas arriba, permite contrastar el resultado de una provincia con el de su región contenedora, o de un departamento con su provincia contenedora.

La Fig 4 exhibe el primero de estos dos escenarios correspondiente al año 2015. Esto, a modo ilustrativo del caudal de gráficos que se pueden obtener en esta escala, como paso previo a revisar sus estadísticos derivados. El mismo procedimiento se realiza a escala departamental, para diferentes años, desagregando por sexos, o filtrando según frecuencia de cada apellido.

5 Conclusiones

Los apellidos proporcionan una perspectiva valiosa para enriquecer la comprensión de la estructura demográfica de una población. Esta contribución resulta

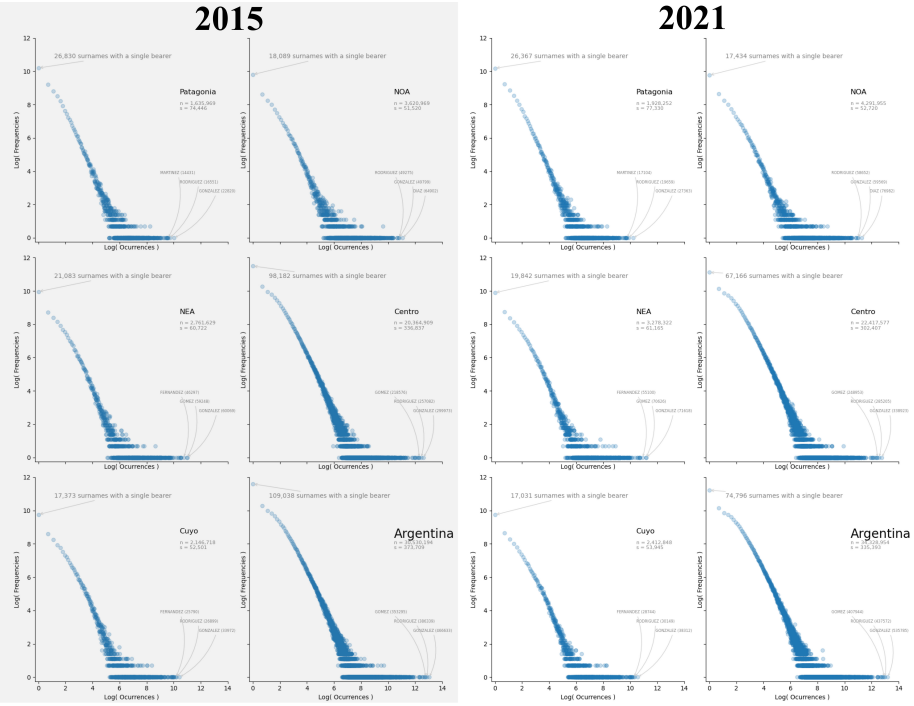

Fig. 3. Gráficos log-log de frecuencias versus ocurrencias a partir de apellidos de las cinco regiones y de la Argentina completa. El cálculo es acelerado por el paquete *isonymic* para los padrones electorales de los años 2015 y 2021.

de suma importancia en el ámbito de la salud, ya que contribuye a delinear barreras sociales, identificar grupos poblacionales aislados, detectar momentos de intensificación en los desplazamientos de las personas que inciden en el desarrollo urbano, entre otros aspectos relevantes.

El paquete *isonymic* ha evidenciado su eficacia al acelerar significativamente el cálculo de los indicadores relacionados con los apellidos, además de posibilitar el establecimiento de procedimientos de datos precisos, mantenibles y reproducibles. Los conjuntos de datos isonímicos resultantes proporcionan una instantánea de la estructura demográfica por apellidos de una población. Instantáneas isonímicas a nivel departamental, provincial, regional y nacional de la Argentina para el año 2015 se exploraron a través de la aplicación web Bulsarapp [22]. La implementación de *isonymic* promueve el desarrollo y la evolución de tales herramientas, permitiendo la expansión del período de tiempo considerado y la creación de nuevas formas de interacción y exploración. Además, los grandes cuerpos de datos isonímicos, transversales en el tiempo, cuya conformación se ve facilitada por el paquete propuesto en este trabajo, se incorporan como nuevos insumos en estudios epidemiológicos en nuestro país. Por ejemplo, el trabajo [23], estudia el fenómeno migratorio bajo la perspectiva de apellidos combinada con

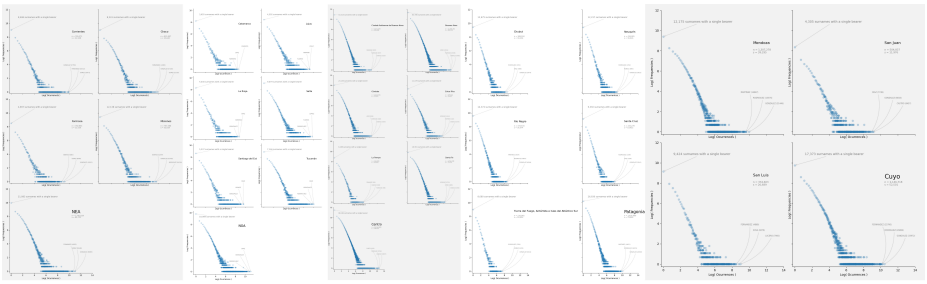

Fig. 4. Gráficos log-log de frecuencias versus ocurrencias a partir de apellidos de las provincias de las cinco regiones. De izquierda a derecha, se presentan los gráficos para las provincias de las regiones Noroeste, Noreste, Centro, Patagonia y Cuyo. La biblioteca Python *isonymic* acelera los cálculos en todas las escalas administrativas del país.

el estudio epidemiológico de enfermedades específicas. Puntualmente, se observa la distribución espacial conjunta de portadores de apellidos alemanes del Volga y de fallecimientos causados por Alzheimer tipo 4 en la Argentina, para la segunda década de este siglo.

La difusión del paquete *isonymic* en repositorios públicos tiene como objetivo ampliar la comunidad de investigadores interesados en utilizar los apellidos como herramienta de análisis en estudios poblacionales.

References

1. Yasuda, N. & Furusho, T. Random and nonrandom inbreeding revealed from isonymy study. I. Small cities of Japan.. *American Journal Of Human Genetics.* **23**, 303 (1971)
2. Fisher, R., Corbet, A. & Williams, C. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal Of Animal Ecology.* pp. 42-58 (1943)
3. Rodríguez-Larralde, A., Formica, G., Scapoli, C., Beretta, M., Mamolini, E. & Barrai, I. Microevolution in Perugia: isonymy 1890–1990. *Annals Of Human Biology.* **20**, 261-274 (1993)
4. Dipierri, J., Alfaro, E., Scapoli, C., Mamolini, E., Rodriguez-Larralde, A. & Barrai, I. Surnames in Argentina: A population study through isonymy. *American Journal Of Physical Anthropology.* **128**, 199-209 (2005)
5. Rodriguez-Larralde, A. Estimadores de aislamiento en base a distribución de apellidos. *XXXVI Convención Anual De AsoVAC, Valencia, Venezuela.* (1986)
6. Rodríguez Larralde, A. & Barrai, I. Estudio genético demográfico del Estado Zulia, Venezuela, a través de isonimia. *Acta Cient. Venez.* pp. 134-43 (1998)
7. Cavalli-Sforza, L. & Edwards, A. Phylogenetic analysis. Models and estimation procedures. *American Journal Of Human Genetics.* **19**, 233 (1967)
8. Nei, M. Analysis of gene diversity in subdivided populations. *Proceedings Of The National Academy Of Sciences.* **70**, 3321-3323 (1973)
9. Wright, S. Isolation by distance. *Genetics.* **28**, 114 (1943)

10. Nei, M. & Imaizumi, Y. Genetic structure of human populations. *Heredity*. **21** pp. 183-190 (1966)
11. Karlin, S. & McGregor, J. The number of mutant forms maintained in a population. *Proceedings Of The Fifth Berkeley Symposium On Mathematics, Statistics And Probability*. **4** pp. 415-438 (1967)
12. Piazza, A., Rendine, S., Zei, G., Moroni, A. & Cavalli-Sforza, L. Migration rates of human populations from surname distributions. *Nature*. **329**, 714-716 (1987)
13. Zei, G., Guglielmino, C., Siri, E., Moroni, A. & Cavalli-Sforza, L. Surnames as neutral alleles: observations in Sardinia. *Human Biology*. pp. 357-365 (1983)
14. King, T. & Jobling, M. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. *Trends In Genetics*. **25**, 351-360 (2009)
15. Kandt, J., Cheshire, J. & Longley, P. Regional surnames and genetic structure in Great Britain. *Transactions Of The Institute Of British Geographers*. **41**, 554-569 (2016)
16. Darlu, P., Bloothooft, G., Boattini, A., Brouwer, L., Brouwer, M., Brunet, G., Chareille, P., Cheshire, J., Coates, R., Dräger, K. & Others The family name as socio-cultural feature and genetic metaphor: From concepts to methods. *Human Biology*. **84**, 169-214 (2012)
17. Clark, G. & Cummins, N. Intergenerational wealth mobility in England, 1858–2012: surnames and social mobility. *The Economic Journal*. **125**, 61-85 (2015)
18. Mateos, P. & Others Names, ethnicity and Populations. *Advances In Spatial Science*. (2014)
19. Lan, T., Kandt, J. & Longley, P. Geographic scales of residential segregation in English cities. *Urban Geography*. **41**, 103-123 (2020)
20. Petersen, J., Longley, P., Gibin, M., Mateos, P. & Atkinson, P. Names-based classification of accident and emergency department users. *Health & Place*. **17**, 1162-1169 (2011)
21. Lewis, D. & Longley, P. Patterns of patient registration with primary health care in the UK National Health Service. *Geographies Of Health, Disease And Well-being*. pp. 261-271 (2016)
22. Morales, L., Navarro, P., Cintas, C., Gonzalez-Jose, R., Ramallo, V. & Delrieux, C. Bulsarapp: interactive visual analysis for surname trend exploration. *IEEE Computer Graphics And Applications*. **42**, 28-39 (2021)
23. Morales, A., Figueroa, M., Navarro, P., Chaves, E., Ruderman, A., Dipierri, J. & Ramallo, V. Volga German surnames and Alzheimer's disease in Argentina: an epidemiological perspective. *Journal Of Biosocial Science*. pp. 1-14 (2024)

Tabla 1. Funciones disponibles en la biblioteca isonymic. La función *get_surname_frequencies* es primordial para la elaboración de los gráficos log-log de apellidos

Nombre	Argumentos	Descripción
get_isonymy	Apellidos de los individuos de la población	Retorna la isonímia no sesgada (I) a partir de un listado de apellidos dado.
get_fishers_alpha	Apellidos de los individuos de la población	Retorna el valor del α de Fisher partir de un listado de apellidos recibido como parámetro.
get_a_index	Apellidos de los individuos de la población	Retorna el valor del indicador A según un listado de apellidos recibidos.
get_b_index	Apellidos de los individuos de la población	Retorna el valor del indicador B a partir de un listado de apellidos recibidos.
get_isonymic_indicators	Apellidos de los individuos de la población	Retorna una estructura (dict) que contiene todos los valores de los indicadores listados previamente.
get_wright_m_vect	Número de población total por unidades, Valor isonímia por unidades	Retorna el valor del indicador m de Wright a partir de los vectores de población y de coeficientes de consanguinidad por isonímia, recibidos como parámetros.
get_surname_frequencies	Apellidos de los individuos de la población	A partir de un listado de apellidos de personas (con repetición), retorna la frecuencia mínima y máxima de ocurrencias de apellidos, y los apellidos en cuestión según el siguiente orden: – 1: Frecuencia mínima de ocurrencia. Entero, por ejemplo 1. – 2: Apellidos con la mínima ocurrencia. Listado, por ejemplo: RaroA, RaroB. – 3: Frecuencia máxima de ocurrencia: Entero, por ejemplo 1000. – 4: Apellidos con la máxima ocurrencia: Listado, por ejemplo: PopularA, PopularB.
get_distances	Apellidos de los individuos de la población I y J	Devuelve una estructura (diccionario) con todas las distancias a partir de apellidos entre las dos listas de apellidos correspondientes a la población i y la población j. La estructura resultante tiene las siguientes entradas: – 'I_{ij}': Isonímia entre i y j. – 'L_{ij}': Lasker entre i y j. – 'E_{ij}': Euclidea entre i y j. – 'N_{ij}': Nei entre i y j. – 'I_{ii}': Isonímia en i.