Métodos computacionales avanzados para evaluar la correlación temporal entre apellidos y ancestría biológica en Argentina

Lic. Arturo Leonardo Morales

Director: Dr. Claudio Delrieux

CoDirectora: Dra. Virginia Ramallo

30 de julio de 2024

Prefacio

Esta tesis se presenta como requisito para la obtención del grado académico de Doctor en Ciencias de la Ingeniería, otorgado por la Universidad Nacional de la Patagonia San Juan Bosco. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el ámbito del Instituto Patagónico de Ciencias Sociales y Humanas "Dra. María Florencia del Castillo Bernal" CONICET-CENPAT y el Laboratorio de Ciencias de las Imágenes de la Universidad Nacional del Sur durante el período comprendido entre el 1 de Abril de 2018 y el 1 de Abril de 2024, bajo la dirección del Dr. Claudio Delrieux y la Co-dirección de la Dra. Virginia Ramallo.

A Gladis, Rubén y Leone.

Agradecimientos

En primer lugar, quiero expresar mi más profundo agradecimiento a Claudio y Virginia, quienes con su guía y apoyo me acompañaron a lo largo de todo el doctorado. Una mención especial merece Virginia, mi directora, compañera de oficina y amiga. Bajo tu dirección, he aprendido mucho más de lo que podría expresar aquí, sos un ejemplo de generosidad intelectual y humana Vir.

Extiendo mi gratitud al GIBEH, que me proporcionó un ambiente propicio para desarrollar mi investigación, brindándome el espacio y los recursos necesarios para avanzar en este proceso. También agradezco a aquellas argentinas y argentinos que, con su esfuerzo cotidiano y su determinación por un ideal de país, hacen posible la universidad pública, la ciencia, la tecnología y la investigación en nuestro territorio.

A mis compañeros y amigos Pablo, Bruno, Papablo, Paula, Emiliano y Diego, les agradezco su compañía y apoyo a lo largo de esta etapa. Sin duda hicieron que el recorrido fuera mucho más llevadero y me hacen crecer día a día.

A mi familia, Gladis y Rubén, su apoyo y amor incondicional son los pilares de todo lo bueno que me sucede. Y a Leone y Maxi, su compañia, energía y respaldo me motivan en cada paso del camino.

Finalmente, quiero dedicar unas palabras a Cindy, la persona más importante en mi vida. Gracias por ser mi apoyo incondicional, mi refugio y mi fortaleza. Soy muy afortunado de que nuestros caminos se hayan encontrado, este logro también es tuyo.

A todos ustedes, muchas gracias.

Abreviaturas

API Application Programming Interface

BILISA Indicadores Locales Bivariados de Asociación Espacial

CABA Ciudad Autónoma de Buenos Aires

CIE Clasificación Internacional de Enfermedades

CIEMP Clasificación Internacional de Enfermedades de Mortalidad Perinatal

CRSED Coverage Ratio of Stretched Exponential Distribution

CSMR Cause Specific Mortality Rates

DEIS Dirección de Estadísticas e Información de la Salud

DTN Defectos de Cierre del Tubo Neural

EA Enfermedad de Alzheimer

ENT Enfermedades No Transmisibles

EPOF Enfermedades Poco Frecuentes

ESDA Exploratory Spatial Data Analysis

ETL Extract, Transform, and Load

HTML HyperText Markup Language

INDEC Instituto Nacional de Estadísticas y Censos

IRAB Infecciones Respiratorias Agudas Bajas

LISA Indicadores Locales de Asociación Espacial

MF Muerte Fetal

MIM Mendelian Inheritance in Man

NBI Necesidades Básicas Insatisfechas

NOA Noroeste argentino

OMIM Online Mendelian Inheritance in Man

OMS Organización Mundial de la Salud

OPS Organización Panamericana de la Salud

PyPI Python Package Index

PySAL Python Spatial Analysis Library

RSV Respiratory Syncytial Virus

SQL Structured Query Language

SRMI Single-Regional Migration Intensity

SUS System Usability Scale

TMEA Tasas de mortalidad relacionadas con la enfermedad de Alzheimer

TMF Tasas de Muertes Fetales

VS apellidos Volga

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La presente tesis, denominada "Métodos computacionales avanzados para evaluar la correlación temporal entre apellidos y ancestría biológica en Argentina" ha sido posible gracias a una Beca Doctoral Temas Estratégicos CONICET 2017. El objetivo principal consistió en investigar y desarrollar algoritmos y métodos computacionales eficientes destinados a facilitar y diversificar el estudio de la estructura poblacional a través del análisis de apellidos, con el fin de generar diferentes formatos de informes y representaciones gráficas. Con este fin, se emplearon registros censales, padrones electorales y bases estadísticas del sistema de salud, siendo todos estos repositorios digitales de acceso público y abierto. Todas las fases de la investigación velaron por la privacidad de los individuos, atendiendo al espíritu de la Ley Nacional Nº 25.326 de Protección de Datos Personales.

1.1. Contexto

Los apellidos son rasgos culturales transmitidos de un antepasado a sus descendientes a través de un mecanismo vertical, constituyendo un sistema de herencia único de nuestra especie. El análisis de apellidos, o *método isonímico*, puede dar información cuantitativa sobre la estructura genética de las poblaciones. Los apellidos no se distribuyen de forma homogénea en diferentes lugares ni entre diferentes grupos sociales. El grado de isonimia proporciona una medida para evaluar las probabilidades de encontrar los mismos apellidos en diferentes poblaciones, grupos o parejas matrimoniales, lo que permite inferir una historia común. El método isonímico constituye una manera de estimar la consanguinidad y la

endogamia, fenómenos cuya evaluación resulta compleja mediante otros enfoques. Asimismo, posibilita también el análisis del parentesco y la migración en un país, en sus distintos niveles de administración territorial, por ejemplo departamental, provincial y regional.

Conjugados con información proveniente de otras fuentes disponibles, tales como variables sociales, médico-sanitarias y económicas, permitieron conocer aspectos demográficos clave. Entre ellos se destacan la identificación de potenciales grupos aislados dentro de la población, la inferencia de patrones históricos de poblamiento y migración, el análisis de las relaciones inter-fronterizas e intra-regionales, así como también la comprensión de la dinámica de ocupación del espacio y sus factores condicionantes. Las tasas altas de endogamia se han relacionado consistentemente con una mayor incidencia de enfermedades autosómicas recesivas y con la prevalencia de anomalías congénitas. Ahondar en el conocimiento de la estructura poblacional es estratégico para la salud, ya que permite la identificación temprana de factores de riesgo tanto genéticos como ambientales.

La investigación de nuevos algoritmos basados en Big Data y técnicas de Minería de Datos y Aprendizaje de Máquina permitió clasificar todos los apellidos por origen geográfico o lingüístico más probable, correlacionar datos de salud, así como la investigación de procesamiento, análisis y visualización de información. Se implementaron modelos de datos, ontologías específicas y Sistemas de Información Geográfica, permitiendo estudios de aplicación en gran escala y cobertura territorial, así como la identificación de patrones de agrupamiento en el espacio y su evolución temporal.

1.2. Estructura de la Tesis

Este primer capítulo cumple la función de introducción, luego, en el segundo capítulo se aborda la teoría isonímica, que implica la derivación de la estructura demográfica a partir de apellidos. En él, se introducen los conceptos biológicos y genéticos asociados, se destacan sus beneficios y aplicaciones, y se culmina con la presentación de una herramienta interactiva diseñada por el autor para explorar la isonimia en la población argentina. El tercer capítulo se centra en el marco teórico relacionado con la migración, analizado a través de la perspectiva de los apellidos. Conforme avanza su desarrollo, se presentan dos casos de estudio: el primero examina el desplazamiento de una población bien conocida y su

relación con la epidemiología del Alzheimer de tipo 4, mientras que el segundo caso describe la sistematización del análisis isonímico a partir de padrones electorales en Argentina. El cuarto capítulo retoma la metodología de estudio epidemiológico presentada en el capítulo anterior y la expande hacia la investigación de nuevos fenómenos. En esta línea, se presentan tres casos de estudio en distintas escalas poblacionales de Argentina, que evidencian la eficacia de implementar enfoques flexibles para investigaciones sanitarias. El quinto capítulo de la presente tesis expone las conclusiones derivadas de los estudios llevados a cabo, y se destacan las potenciales líneas de investigación que podrían ser exploradas en futuros trabajos

1.3. Contribuciones

1.3.1. Publicaciones referidas a esta tesis en revistas internacionales

- **Morales, L.**, Navarro, P., Cintas, C., Gonzalez-Jose, R., Ramallo, V., & Delrieux, C. (2021). Bulsarapp: Interactive visual analysis for surname trend exploration. IEEE Computer Graphics and Applications, 42(4), 28-39.

- Pazos, B. A., **Morales, A. L**., Ramallo, V., González-José, R., de Azevedo, S., & Taire, D. L. (2023). Mapping spatial morbidity patterns for bronchiolitis related to socioeconomic estimators: A spatial epidemiology approach to identify health disparities in Puerto Madryn, Argentina. American Journal of Human Biology, 35(10), e23938.

- Bronberg, R., Martinez, J., **Morales, L.**, Ruderman, A., Taire, D., Ramallo, V., & Dipierri, J. (2023). Prevalence and secular trend of neural tube defects in fetal deaths in Argentina, 1994–2019. Birth Defects Research, 115(18), 1737-1745.

- **Morales, A. L.**, Figueroa, M. I., Navarro, P., Chaves, E. R., Ruderman, A., Dipierri, J. E., & Ramallo, V. (2024). Volga German surnames and Alzheimer's disease in Argentina: an epidemiological perspective. Journal of Biosocial Science, 1-14.

1.3.2. Trabajos completos incluidos en actas de conferencias nacionales e internacionales

- **Morales, A. L.**, Figueroa, M., Delrieux, C., Ramallo, V., & Dipierri, J. E. (2023). Herramienta para la exploración de tendencias y detección de patrones epidemiológicos en Argentina. Memorias de las JAIIO, 9(5), 137-142.

1.3.3. Otras publicaciones en revistas.

- Trujillo-Jiménez, M. A., Navarro, P., Pazos, B., **Morales, L.**, Ramallo, V., Paschetta, C., ... & Gonzalez-José, R. (2020). Body2vec: 3D point cloud reconstruction for precise anthropometry with handheld devices. Journal of Imaging, 6(9), 94.

- Marticorena, L. G., **Morales, L. A.**, Antonelli, L., Rossi, G., & Firmenich, D. (2023). Development iterations based on web augmentation and context tasks. Multimedia Tools and Applications, 82(8), 11793-11817.

- Ruderman, A., Luisi, P., Paschetta, C., Teodoroff, T., Pérez, L. O., de Azevedo, S., **Morales, L.** ... & Ramallo, V. (2023). Genetic and self-perceived ancestries in Argentina: Beyond the three-hybrid model. American Journal of Biological Anthropology.

1.3.4. Trabajos completos publicados en actas de congresos.

- Pazos, B. A., & **Morales, A. L.** (2018). Computación corporal: Expansión de la sensibilidad computacional hacia mejores experiencias de usuario. In XXI Concurso de Trabajos Estudiantiles (EST)-JAIIO 47 (CABA, 2018).

- Navarro, J. P., Cintas, C., Ramallo, V., de Azevedo, S., **Morales, A. L.**, Trujillo, A., ... & Delrieux, C. (2018). Escaneo corporal 3D de bajo costo para monitoreo y seguimiento remotos de sobrepeso. In IX Congreso Argentino de Informática y Salud (CAIS)-JAIIO 47 (CABA, 2018).

- Pazos, B. A., Navarro, J. P., **Morales, A. L.**, Cintas, C., Trujillo, A., de Azevedo, S., ... & Delrieux, C. (2018). Detección automática de tejido blando nasal en CT-Scan

y MRI utilizando Random Forests. In IX Congreso Argentino de Informática y Salud (CAIS)-JAIIO 47 (CABA, 2018).

- **Morales L.A**, J.E. Dipierri, A.C. Cardoso Dos Santos, V. Ramallo. (2023). Epidemiología de las muertes por EPOF en Argentina. BAG I Journal of Basic and Applied Genetics I Vol XXXIV Issue 1 (suppl.): 123-136.

Capítulo 2

Apellidos

2.1. Introducción

El presente capítulo resume los fundamentos teóricos de la isonimia, método que permite abordar el estudio de una población a partir de los apellidos de los individuos que la componen. Se presentarán diversos antecedentes de investigación relevantes, así como también las fuentes de datos habitualmente utilizadas en este ámbito y aquellas que fueron específicamente utilizadas en el presente estudio, sirviendo como insumo de entrada para el diseño de una herramienta informática propia. Dicha herramienta fue creada con el objetivo de agilizar la exploración y el análisis de patrones isonímicos en Argentina. Finalmente, se proponen las maneras y los esfuerzos necesarios para expandir estos análisis en el futuro, a medida que se disponga de nuevas bases de datos en años venideros.

2.2. Apellidos

Con el propósito de contextualizar este trabajo, se seleccionó, de manera deliberadamente arbitraria, uno de los acontecimientos históricos que contribuyeron a consolidar el uso de apellidos como sistema de herencia: la conquista de Inglaterra por los franco-normandos (Redmonds et al., 2011). Si bien se reconoce que el desarrollo del sistema de apellidos ha estado influenciado por una diversidad de fuerzas y dinámicas históricas que han moldeado su evolución en múltiples culturas a nivel global, este evento se emplea aquí como punto de partida para explorar las dinámicas poblacionales que emergen y evolucio-

nan a medida que los apellidos comienzan a transmitirse como un rasgo hereditario. Aquel evento, ocurrido el 14 de octubre de 1066, tuvo lugar cuando el ejército del rey Haroldo II fue derrotado en Hastings, una localidad en el sur de Inglaterra, por las fuerzas lideradas por Guillermo II, que posteriormente sería conocido como Guillermo "El Conquistador".

La nueva configuración política del territorio derivada de tal suceso, requeriría una notable creatividad para su administración y organización. En este contexto, regido por cambios cruciales para organizar las tierras anexadas (y también dirigir aquellas que quedaron en Normandía), se produjo el paso de una cultura oral hacia una escrita (Redmonds et al., 2011). El gobierno de Guillermo II elaboró registros detallados, en los que se asignaba un jefe de familia a cargo de los distintos territorios. Este uso de los primeros nombres progresaría hasta la configuración de apellidos tal como se conocen en el presente. Los más frecuentes estuvieron relacionados con ocupaciones u oficios en los que se desempeñaba el jefe de familia, con apodos, o con relaciones o vínculos entre las personas (patronímicos, es decir, nombres que derivan del apelativo del padre o de otro antecesor masculino), así como con nombres de lugares (denominados como toponímicos). Para preservar el patrimonio dentro del mismo ámbito de parentesco, se volvió fundamental la transmisión de este distintivo nominal a los sucesores.

La adopción de los apellidos fue un proceso extenso, desde las etapas de utilización del nombre propio como identificador familiar, seguido del asentamiento y popularización de la práctica, hasta el afianzamiento como carácter hereditario. Su cronología en Europa abarca desde los siglos XI al XV, pudiéndose señalar eventos que contextualizan su paulatina gradualidad. Por ejemplo, en el siglo XIV, la pandemia de peste bubónica, conocida como Peste Negra, mató a un tercio de los habitantes de Eurasia. Este evento crítico obligó a reconfigurar la manera en la que se organizaban las sociedades y cómo se consignaba a las personas que las integraban. Otro hecho, que podemos traer a colación para ilustrar este marco desafiante, tiene que ver con que la adopción de un apellido no significó lo mismo para todas las clases sociales. Portar un determinado nombre de familia manifestaba el estatus de la persona. La inmensa mayoría de la población, masas excluidas de bienes o propiedades, no eran registradas en los escritos administrativos ni se observaba sobre sus nombres ningún tipo de regularidad.

Contemporáneamente, otras sociedades del mundo establecieron otros sistemas para

nombrar a los individuos y señalar las relaciones de parentesco o de afinidad que se establecieran entre ellos. Muchos de estos sistemas co-existieron y co-existen aún con el uso de los nombres de familia (Foerster, 2010).

Desde su formalización en la Edad Media y su posterior uso como procedimiento de control sobre la propiedad y la población durante la expansión a los reinos de ultramar, los apellidos permanecen como rasgos culturales que se transmiten entre ancestros y descendientes. Este mecanismo de herencia sigue un sistema vertical comparable a la transmisión de algunas variantes genéticas, como un locus con múltiples alelos (Guglielmino et al., 1991). Debido a esta característica, su distribución en una población se ajustaría a la de los alelos selectivamente neutros y son una variable sociocultural que funciona como un proxy de la dimensión biológica (Manrubia and Zanette, 2002).

Los apellidos no aportan información sobre periodos anteriores a finales de la Edad Media. Aún así, esta limitación temporal representa una ventaja si se desea estudiar procesos como migración, deriva o aislamiento recientes, que hayan modificado significativamente las estructuras genéticas de las poblaciones (Manni et al., 2005). La comparación entre la variabilidad de apellidos con la variabilidad de datos genéticos y lingüísticos torna evidente algunos cambios dramáticos, en especial en nuestro continente. A través de la imposición colonial del uso de apellidos en América, muchos de los nombres propios autóctonos se transformaron en nombres de familia. Existen documentos desde el siglo XVI (libros de bautismo, censos, padrones de encomiendas) en los que se consigna la ubicación y número de la población originaria sojuzgada en estructuras políticas de control como reducciones, encomiendas o yanaconazgos. En estos libros de asientos, los funcionarios eclesiásticos o civiles detallan los nombres utilizados, describen si son femeninos o masculinos y a lo largo de los sucesivos censos, en algunas regiones de Argentina (especialmente en el Noroeste y la Patagonia), los descendientes de las primeras personas censadas son registrados con esos mismos nombres como apellidos (Alfaro et al., 2005). En la Fig. 2.1 se presenta un ejemplo de esta clase de registro, el folio 10 del libro de bautismos de los años 1884 a 1892 de la parroquia María Auxiliadora de la localidad de Rawson, Chubut. El asiento número 11 en dicho folio deja constancia de la conversión al catolicismo de un niño, quien es bautizado como Antonio Carranza No-Taú. El nombre de pila de su padre era No-Taú.

Es posible así establecer un origen lingüístico, étnico o geográfico más probable, ras-

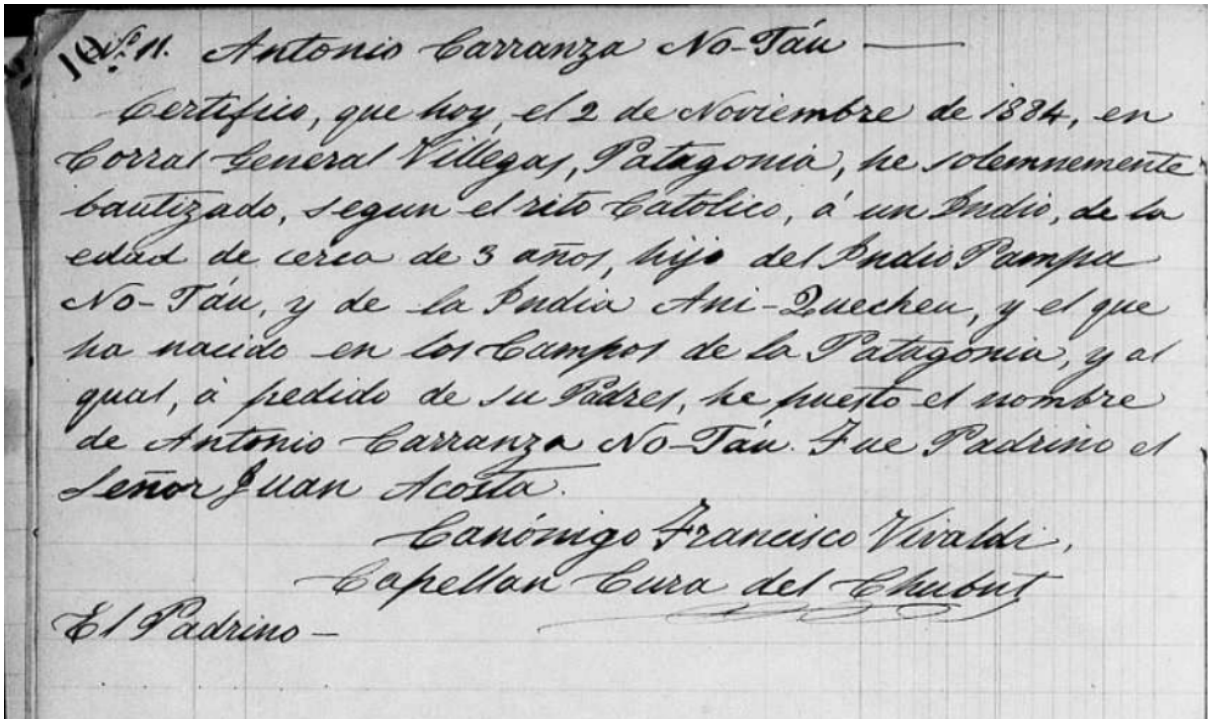


Figura 2.1: Asiento 11, folio 10, libro de bautismos 1884-1892, parroquia María Auxiliadora (Rawson, Chubut). Transcripción del texto manuscrito: *"N°11 Antonio Carranza No-Taú. Certifico que hoy, el 2 de noviembre de 1884, en Corral General Villegas, Patagonia, he solemnemente bautizado, según el rito Católico, a un Indio, de la edad de cerca de 3 años, hijo del Indio Pampa No-Taú, y de la India Ani-Quecheu, y el que ha nacido en los Campos de la Patagonia, y al qual, a pedido de sus Padres, he puesto el nombre de Antonio Carranza No-Taú. Fue Padrino el Señor Juan Acosta. Canónigo Francisco Vivaldi. Capellán Cura del Chubut. El Padrino-"*

go que robustece a los apellidos convirtiéndolos en un instrumento de análisis muy útil en las sociedades multiculturales contemporáneas. Este análisis discriminante por origen constituye un método adecuado para distinguir grupos y estructuración al interior de un conjunto social aparentemente homogéneo y describir escenarios migratorios complejos. Su uso puede ayudar a minimizar los errores de muestreo al indicar dónde y cuándo es probable que permanezca un patrón genético preexistente (Chakraborty et al., 1989, Dipierri et al., 1999, Mateos, 2007).

Los actuales estudios con apellidos combinan saberes y praxis de diversas disciplinas como la lingüística, la genealogía, la historia, la genética, la geografía, entre otras. Los avances en la tecnología referidos a la digitalización y el análisis de la información han contribuido positivamente a estos estudios interdisciplinarios. Como se puede inferir, las fuentes documentales con las que se trabaja pueden ser de muy diversa naturaleza y plantear diferentes desafíos para su sistematización, y abarcan desde libros parroquiales, actas de registros civiles, censos de población, hasta padrones electorales o cualquier otro listado relevado por el Estado. Como producto del largo proceso de adopción de este sistema nominal, los apellidos no están exentos de cambios (ya sea por errores en su escritura, por pérdida de documentación probatoria, por persecuciones político-religiosas que obligan a las personas a cambiarlos) y la relación entre herencia genética y herencia de nombre de familia no se mantiene inmutable a lo largo de las generaciones (ya sea por adopciones o filiaciones ilegítimas). Sin embargo, cuando el volumen de datos sobrepasa la magnitud de millones, el efecto de estas divergencias es desestimable. En esta tesis focalizamos en el creciente rol de la ciencia de datos en estos análisis, destacando los medios aplicados a la diversificación y mayor potencia de los estudios isonímicos.

2.3. El método isonímico

El uso de apellidos para estudios poblacionales tiene su origen en 1875, cuando George Darwin estimó la consanguinidad en matrimonios isonímicos, es decir aquellas uniones entre personas que compartían el mismo nombre de familia (Darwin, 1875). El término "consanguinidad" es derivado de la palabra en latín *consanguineus*, que significa "de la misma sangre". No debe confundirse con el término "endogamia", que se refiere al tipo

de unión preferencial en una sociedad determinada. Diferentes motivos, ya sea culturales (como economía, política, religión, idioma) o espaciales (proximidad entre localidades y personas) determinan que exista una tendencia a seleccionar posibles cónyuges dentro del mismo grupo (endogamia) o entre grupos diferentes (exogamia). Si la tendencia a la endogamia persiste durante muchas generaciones y las comunidades mantienen un tamaño pequeño, invariablemente se llegará a un punto en que los individuos comparten antepasados comunes y la diversidad genética de la población habrá disminuido. El conocimiento del grado de endogamia resulta importante para la salud pública, ya que guarda correlación con un alto riesgo de enfermedades genéticas de herencia recesiva (Torres-Hernández et al., 2023). George Darwin, uno de los diez hijos del matrimonio entre el famoso Charles Darwin y su prima hermana, Emma Wedgwood, se interesó en investigar si la unión de sus padres, a pesar de ser próspera, no resultaría inconveniente desde el punto de vista biológico, pudiendo derivar en padecimientos de salud para su descendencia. Demostró en (Darwin, 1875) que no había evidencia de aumento de mortalidad para los hijos de primos hermanos. Posteriormente se pudo comprobar que la mortalidad de la progenie entre primos de primer orden es 3.5 % más alta que en descendencia no consanguínea y que es primordial observar el fenómeno como multifactorial, en perspectiva con la demografía y los factores económicos y sociales (Bittles and Black, 2010).

Posteriormente, con los avances en el conocimiento sobre estos riesgos, se dieron lugar a nuevas legislaciones en las sociedades del siglo XIX, que prohibieron las uniones entre primos de primer orden. La similitud entre la forma en que un apellido pasa de una persona a otra, y la manera en que se heredan los genes, ha sido el disparador de los estudios poblacionales a partir de apellidos. El trabajo fundamental que puso en valor su uso para obtener información demográfica fue el publicado por los profesores James F. Crow y Arthur P. Mange, en 1965 (Crow and Mange, 1965). Trabajando con información de matrimonios isonímicos y con las frecuencias de portadores de apellidos idénticos en una población, presentaron una primera versión de su estimador de endogamia, aplicable incluso en casos en los que la información genealógica está incompleta o es nula. Este estimador, en su forma más simple, puede considerarse como un cuarto de la fracción de matrimonios isonímicos (Yasuda and Furusho, 1971). El punto destacado de esta metodología radica en que los apellidos satisfacen las expectativas de la teoría neutra

de la evolución (Cavalli-Sforza and Bodmer, 1999). Por ejemplo, en aquellas sociedades en donde los apellidos se transmiten por línea paterna, pueden homologarse a la herencia del cromosoma Y, parte del genoma que determina el sexo biológico y los padres pasan en forma exclusiva a sus hijos varones (Yasuda and Furusho, 1971, Yasuda and Morton, 1967, Yasuda et al., 1974). La evolución de este dispositivo cultural heredable puede ser descrita íntegramente mediante la deriva genética aleatoria, la mutación y la migración (Kimura, 1983). Con el paso del tiempo, el método isonímico se ha perfeccionado y ha servido para estudios demográficos en diversas poblaciones del mundo, funcionando también con distintos sistemas de transmisión. En el caso de Japón (Yasuda and Furusho, 1971), por ejemplo, se demostró que el traspaso del apellido, ya sea patrilineal (de parte del padre) como matrilineal (de la madre), no afecta la relación entre coeficiente de endogamia propuesto y la frecuencia de los apellidos. Los desarrollos teóricos en torno a la relación entre apellidos e individuos se fueron complejizando, permitiendo incluso alcanzar métodos de cálculos fiables para estimación de tasas migratorias (Piazza et al., 1987).

Relethford (Relethford, 1982), a partir de los listados de apellidos de 1890, analizó la diversidad poblacional en siete localidades irlandesas. Detectó que la variación entre ellas estuvo fuertemente influida por la separación geográfica, siendo este factor el principal determinante en el modelo de aislamiento por distancia.

En el siglo XXI se continúa utilizando el método isonímico para describir la estructura demográfica de las poblaciones. Por ejemplo, casi 50 años después del aporte fundamental de Crow y Mange, en el trabajo de (Dugène and Bauduer, 2013), se analiza una comunidad francesa de escasa movilidad, en el período entre 1800 y 1899. Utilizaron datos de archivos parroquiales, civiles y notariales para evaluar variables o marcadores biodemográficos como la estacionalidad (momentos del año en que se celebraban mayor cantidad de matrimonios), la tasa de endogamia (si esas uniones tendían a ser más frecuentes dentro de la comunidad o no) y el coeficiente de consanguinidad (cuántos de esos matrimonios involucraban personas emparentadas). Estos trabajos ponen de manifiesto otras de las ventajas que ofrece trabajar con apellidos: su bajo o nulo costo y la posibilidad de analizar poblaciones pasadas, con las cuales ya no es posible la interacción directa.

Estos estudios han sido aplicados a distintos niveles de organización administrativa, ya sea continental, nacional, regional, provincial, departamental y municipal. En este en-

tramado, algunos niveles organizacionales incluyen poblaciones muy grandes e involucran cantidades masivas de datos y, por otro lado, estudios sobre unidades administrativas más pequeñas requieren replicaciones ordenadas del camino exploratorio y de los cálculos involucrados para el análisis integral final. En los siguientes párrafos se presentará un panorama general de los indicadores que pueden obtenerse a partir de los apellidos y se resumirán trabajos fundamentales realizados en poblaciones con distintas características del mundo, de Latinoamérica y puntualmente Argentina.

2.4. Indicadores a partir de los apellidos

El indicador de isonimia I (Crow and Mange, 1965), que se calcula a partir de la frecuencia de matrimonios entre personas con el mismo apellido, representa el estimador principal en los estudios del campo. Podemos medir esta proporción en una población tomando pares aleatorios de apellidos. Entonces, el cálculo se realizaría con las siguientes sumatorias,

$$\sum q_m q_f \qquad \sum q_m^2 \qquad \sum q_f^2 \qquad \sum q_{m+f}^2, \tag{2.1}$$

donde q_m, q_f y q_{m+f} son las frecuencias de un apellido en varones, mujeres y en el total de la población, respectivamente (Yasuda and Furusho, 1971). Por regla general, se puede obviar la distinción de los sexos y directamente realizar la sumatoria de las proporciones de cada apellido considerando todas las personas de una población. Por lo tanto el indicador I puede estimarse como

$$I = \sum x_i^2, \tag{2.2}$$

siendo x_i la frecuencia relativa del i-ésimo apellido de la población.

Si consideramos el ejemplo de una sociedad donde la herencia del apellido es patrilineal y tomamos un caso de matrimonio entre primos, puede involucrar a un varón casado con la hija de su tía materna (hermana de su madre), la hija de su tío materno (hermano de su madre), la hija de su tía paterna (hermana de su padre), o la hija de su tío paterno (hermano de su padre) (ver Fig. 2.2). Solo en el último de estos cuatro casos el matrimonio sería isonímico, es decir, ambos integrantes tendrían el mismo apellido (Lasker, 1968).

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 obtener entonces

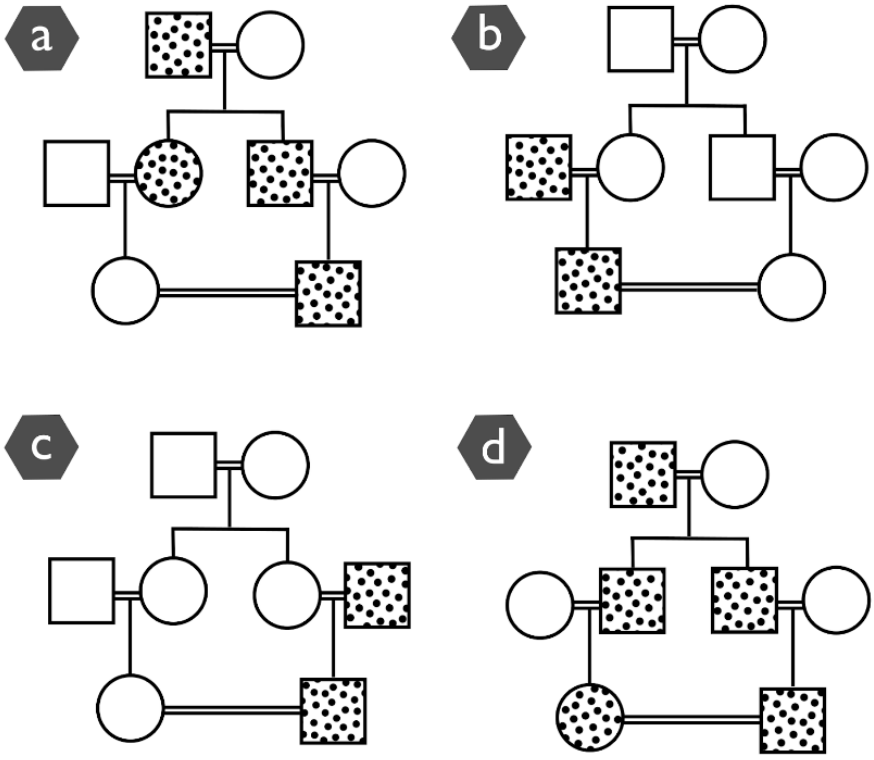


Figura 2.2: Cuatro posibles genealogías de un matrimonio consanguíneo entre primos en primer grado. La herencia del apellido es patrilineal (patrón de puntos). Los círculos representan a las mujeres y los cuadrados a los hombres. Las uniones están señaladas mediante una línea doble. Se representan, en cada subfigura: a) un hombre casado con la hija de su tía paterna (hermana de su padre); b) un hombre casado con la hija de su tío materno (hermano de su madre); c) un hombre casado con la hija de su tía materna (hermana de su madre); d) un hombre casado con la hija de su tío paterno (hermano de su padre). Únicamente este último casamiento es isonímico.

un indicador de isonimia aleatoria entre dos poblaciones, aplicando la fórmula

$$I_{ij} = \sum p_{ik} p_{jk}, \tag{2.3}$$

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 (Rodríguez Larralde and Barrai, 1998) se define con la fórmula:

$$L = -\log(I_{ij}). \tag{2.4}$$

2. Distancia Euclídea entre dos grupos (Cavalli-Sforza and Edwards, 1967). Se define según la fórmula:

$$E = \sqrt{1 - \sum_k \sqrt{p_{ki} p_{kj}}}. \tag{2.5}$$

3. Distancia de Nei. Según (Nei, 1973) se define con la fórmula:

$$N_d = -\log\left(\frac{I_{ij}}{\sqrt{I_{ii} I_{jj}}}\right). \tag{2.6}$$

Los estudios de aislamiento por distancia evalúan la correlación lineal de estas distancias de apellidos con las distancias geográficas, en los distintos niveles de separación administrativa (Dipierri et al., 2005a). Las distancias geográficas se calculan generalmente entre las ciudades capitales o entre los puntos centrales de los polígonos que representan a las unidades administrativas en un mapa.

Retomando a los indicadores propios de una unidad territorial (como puede ser una provincia, un departamento o una región) veremos cómo se construyen tanto el estimador de endogamia como el α de Fisher a partir del análisis de sumatorias de proporciones de apellidos, es decir, del valor I.

El estimador de endogamia F_{st} se calcula con la fórmula:

$$F_{st} = \frac{I}{4}. \tag{2.7}$$

El otro indicador, denominado α de Fisher, surge originalmente como indicador de diversidad de la población, enmarcado en el campo de la biología (Fisher et al., 1943). Este ha sido adaptado luego para poder medir la riqueza en apellidos y se estima mediante la

fórmula $\alpha = \frac{1}{I}$ (Rodríguez-Larralde et al., 1993). También puede pensarse este indicador como un estimador del número de apellidos que, teniendo la misma frecuencia, nos dan como resultado la misma isonimia observada en una población. Un valor bajo de α puede representar alta endogamia y deriva, mientras que un valor elevado podría significar migración y baja endogamia (Dipierri et al., 2005a).

Además, al analizar 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. Para su obtención, se computan todas las personas que son las únicas representantes de un nombre de familia y se establece la relación sobre el número total de individuos. Este indicador fue propuesto en 1986 por Rodríguez Larralde, con el objetivo de observar el aislamiento y sedentarismo de las poblaciones (Rodriguez-Larralde, 1986). 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. Por ejemplo, los migrantes llegan a un lugar por motivos laborales, posteriormente en el corto plazo se establecen y aportan su nuevo apellido a la población receptora pero, hasta no tener descendencia, no existirá la posibilidad de transmitirlo. De esta forma, estos nuevos apellidos habrán aportado a la suba del indicador A. Por otro lado, en un nuevo escenario hipotético, muchos portadores del mismo apellido se mudan hacia otras localidades (por motivos académicos, laborales, religiosos, conflictos bélicos, etc), convirtiendo a la última persona residente en la única portadora del apellido que queda en la población. Estos nuevos casos, ahora de emigración, también habrán aportado a la suba del indicador A. La información de contexto puede ser evidente, por ejemplo, en aquellas regiones que son un polo atrayente de estudiantes universitarios o en las que la expansión de actividades económicas requiere grandes cantidades de trabajadores y trabajadoras. Estos patrones y otros aún más variados, modifican nuestra interpretación del indicador A. Generalmente, el aumento en su valor se relaciona con una intensificación en el movimiento de la población.

Por último, el indicador B representa el porcentaje de individuos cuyo apellido se encuentra dentro de los siete más frecuentes en una población determinada (Rodriguez-

Larralde, 1986). Es un indicador de aislamiento relativo, en donde valores más altos representan sedentarismo, al no recibir nuevos apellidos por migración.

2.5. Estadísticos isonímicos

Los resultados reportados con mayor recurrencia en la literatura incluyen la elaboración de tablas de frecuencia de apellidos, tablas de resumen de indicadores y la creación de representaciones cartográficas de diversa complejidad. Estas representaciones en mapas pueden presentar formas incrustadas o utilizar gradientes de colores para ilustrar la información. Junto a estas salidas principales, son comunes otros métodos según las preguntas a responder en cada investigación, como los gráficos log-log de frecuencias *vs.* ocurrencias, las matrices de distancia geográficas o isonímicas, la estadística espacial, y la clasificación por orígenes. Describiremos cada uno de estos dispositivos a continuación.

2.5.1. Gráficos log-log

El sistema de distribución de apellidos exhibe una dinámica regida por una ley de potencia, en la que la frecuencia de un apellido presenta una relación inversamente proporcional con su posición en el ránking de frecuencias. Esta correlación da lugar a la representación gráfica conocida como gráfico log-log (de ocurrencias *vs.* frecuencias). 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. Por ejemplo, el apellido González es el de mayor popularidad en Argentina, aun así, se sitúa en la posición más baja del listado de frecuencias ordenadas. Esto significa que al confeccionar un ranking según la cantidad de portadores por cada frecuencia identificada (el número de apellidos con un solo portador, dos portadores, tres, y así sucesivamente), el apellido González, con sus 535,785 portadores registrados para el año 2021, ocupará el último lugar debido a la ausencia de otros apellidos que compartan esa misma frecuencia. Este fenómeno confirma que un reducido número de apellidos son extraordinariamente frecuentes, mientras que la mayoría de los elementos se presentan con una frecuencia considerablemente menor. Es interesante ana-

lizar la forma del gráfico resultante. En caso de que presente una concavidad, sugiere una limitada presencia de conjuntos con el mismo apellido, mientras que si exhibe convexidad, denota una sobreabundancia de dichos conjuntos (Tarskaia et al., 2009). Por su parte las distribuciones en cola larga implican que los pocos apellidos más comunes acaparan una gran proporción de la población total. Según las características de la población estudiada, este último fenómeno puede sugerir un fuerte efecto de deriva (Liu et al., 2012).

Algunos autores analizan este gráfico haciendo la distinción por sexo (Barrai et al., 1992). Esto permite observar la similitud de las tendencias e inclusive identificar deficiencias de representatividad de algunos apellidos para un sexo u otro. Otros investigadores optan por seleccionar únicamente el conjunto de apellidos con el mayor número de portadores, como los 100 más frecuentes, y contrastar este grupo con los resultados obtenidos para un subconjunto más reducido, como los 10 o 50 apellidos más populares (Cardoso-dos Santos et al., 2021). En este contexto, cobra una importancia adicional la versatilidad del gráfico, ya que nos permite obtener nuevas perspectivas en función de las características propias de la población en análisis.

2.5.2. Estadística espacial

El índice de autocorrelación de Moran, conocido como el I_{Moran}, proporciona un análisis para la distribución espacial de la isonimia y sus múltiples indicadores, permitiendo la detección de posibles agrupamientos. Este índice cuantifica la correlación entre la variable isonímica de un área específica y los valores correspondientes a las áreas circundantes. En esencia, al considerar un conjunto de unidades espaciales y sus atributos asociados, el I_{Moran} ofrece la capacidad de determinar si el patrón espacial observado presenta agrupación, dispersión o más bien es de naturaleza aleatoria (Moran, 1950). Las unidades espaciales generalmente se corresponden a divisiones departamentales y la obtención de este estadístico implica llevar a cabo una serie de pasos detallados a continuación. En primer lugar, es necesario establecer un vecindario para cada unidad específica dentro de su contexto espacial con el fin de construir una matriz de pesos. Las unidades espaciales son típicamente representadas como polígonos en un mapa. Las estrategias de selección de vecinos pueden basarse en relaciones de contigüidad/adyacencia o en pesos derivados de relaciones basadas en la distancia. Ejemplos de estas estrategias pueden ser el criterio

de contigüidad de *Rook* (torre del ajedrez, dos regiones se consideran vecinas si comparten un límite común) o el criterio de *Queen*, (reina del ajedrez, dos regiones son vecinas si tienen un punto en común). En segundo lugar, se calcula el número medio de casos vecinos (denominado *lag* espacial) y se representa gráficamente la relación atributo frente al lag espacial del atributo para cada unidad espacial. Se conforma así el diagrama de dispersión de Moran. La línea de regresión por mínimos cuadrados que mejor se ajusta a la relación entre atributo-lag espacial del dato tras normalizar las variables, es el coeficiente I_{Moran} global. Como toda correlación, el índice puede variar entre -1 y 1. Los valores positivos indican asociación espacial positiva, en la cual las unidades vecinas tienen valores cercanos, lo que indica una tendencia a agruparse (los valores altos se agrupan cerca de otros valores altos o a la inversa con los valores bajos). Los valores negativos indican asociación espacial negativa, donde las unidades vecinas tienen tendencia a dispersarse (los valores altos repelen a otros valores altos y tienden a estar cerca de los valores bajos). Los valores cercanos o iguales a 0 indican ausencia de autocorrelación espacial o distribución aleatoria. Si bien el Índice de Moran global ofrece una visión general del comportamiento espacial de los datos, también es capaz de desagregarse en sus componentes individuales. Esto proporciona una medida localizada de autocorrelación que permite la generación de mapas con identificación de las áreas con patrones de asociación espacial notablemente altos o bajos ("hotspots" y "coldspots" respectivamente) (Rey et al., 2023). La idea central de los Indicadores Locales de Asociación Espacial (LISA) es determinar cuándo un valor dado y la media de sus vecinos son más similares (alto-alto, bajo-bajo) o diferentes (alto-bajo y bajo-alto) de lo que cabría esperar por azar (Anselin, 1995). La determinación de la significatividad se logra mediante la ejecución de permutaciones que desvinculan de manera aleatoria las unidades y los valores correspondientes al atributo analizado, comparando los estadísticos de Moran Local resultantes en cada iteración. Este análisis puede extenderse para examinar la autocorrelación espacial entre dos variables simultáneamente, mediante el cálculo del I_{Moran} bivariado (Anselin, 1995). Mientras que los LISA miden la agrupación espacial (vecinos similares) y la dispersión espacial (vecinos dispersos o diferentes) de las características de una variable y el lag de la misma variable entre vecinos, los Indicadores Locales Bivariados de Asociación Espacial (BILISA) se centran en la agrupación espacial

y la dispersión espacial entre las características de una variable y otra variable diferente entre vecinos (Anselin et al., 2002).

Matrices de distancias

A partir del cálculo de distancias isonímicas (de Nei, Lasker o Euclídea) y del cálculo de distancias espaciales entre dos conjuntos de población, se confeccionan matrices para evaluar la correlación entre ambas variables. Esto permite detectar la tendencia de agrupamientos no aleatorios o clusters de aquellas unidades administrativas a las que los conjuntos de apellidos representan. Es un insumo fundamental para el contraste de hipótesis sobre el movimiento poblacional, su sentido, si este movimiento encuentra barreras geográficas o de otra naturaleza, y para verificar información previa acerca de las poblaciones y el espacio que ocupan (Tarskaia et al., 2009).

2.5.3. Indicador μ de Karlin-McGregor

Para analizar la migración desde la isonimia, se emplea el marco teórico propuesto por Karlin y McGregor en 1967 (Karlin and McGregor, 1967), que considera a los apellidos como múltiples alelos neutros de un mismo locus. Así, prevé que los apellidos de los individuos que fallecen serán reemplazados interna o externamente. Internamente, por portadores herederos del mismo apellido, es decir, las generaciones de descendientes de los mismos grupos fundadores. Externamente, y con una tasa notada con μ, por apellidos nuevos. Estos nombres adicionales pueden surgir de dos maneras: a través de mutaciones (alteraciones en la escritura de los apellidos) o mediante la incorporación de nuevos apellidos debido a la llegada de nuevos individuos a la población. Como la mutación es un fenómeno poco frecuente en las sociedades contemporáneas, que mantienen registros administrativos continuos y donde la transmisión de apellidos es mayoritariamente regular, el indicador μ resumiría la tasa de migración reciente (Piazza et al., 1987, Zei et al., 1983). El μ se calcula con la fórmula propuesta en (Karlin and McGregor, 1967) de la siguiente forma:

$$\mu = \frac{\alpha}{(N + \alpha)}, \tag{2.8}$$

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

2.5.4. Tasa de migración m de Wright

En 1943, S. Wright estableció un modelo teórico para calcular los cambios en las frecuencias genéticas en una población, discriminando los efectos que tendrían la tendencia a la endogamia, el tamaño efectivo o cantidad de individuos en condiciones de tener descendencia y la tasa de migración. En un extremo de este modelo, las poblaciones demasiado pequeñas tendrían pérdida de diversidad genética (deriva) si las condiciones de aislamiento persistían a lo largo de generaciones (Wright, 1943). Dado que los apellidos funcionan como un proxy de la variabilidad genética, a partir de su distribución y la medición de las distancias isonímicas entre poblaciones (Scapoli et al., 2005), es posible estimar la proporción de intercambio de apellidos o tasa de migración por generación (m) usando la fórmula:

$$m = 1 - \sqrt{\frac{2N_e F_{st}}{(2N_e - 1)\, F_{st} + 1}}, \tag{2.9}$$

donde N_e es el tamaño efectivo de la población, o sea la población que contribuye efectivamente a la próxima generación. De acuerdo a Nei e Imaizumi (Nei and Imaizumi, 1966) el N_e fue obtenido a partir de la relación $N_e = N * 0,65$, siendo N el tamaño censal del departamento, provincia o región considerada. El F_{st} corresponde al coeficiente de consanguinidad por isonimia al azar antes descrito.

2.5.5. Clasificación por orígenes

Dada su historia, es posible establecer un origen lingüístico, étnico o geográfico más probable para cada apellido. Este análisis clasificatorio ha sido extensamente utilizado para reconstruir rutas migratorias, para entender la estructura subyacente de las poblaciones, para monitorear posibles sesgos en el diseño de instrumentos como los censos, incluso para la aplicación de políticas de salud pública, al permitir conocer la fracción de una población potencialmente expuesta a un riesgo (ampliaremos sobre esta línea de investigación a lo largo del Capítulo 2) (Mateos, 2007)

2.6. Estudios isonímicos a nivel mundial

La teoría isonímica ha sido aplicada en numerosos contextos y escalas demográficas. En este apartado se mencionan algunos ejemplos. En el año 2007, Scapoli y colaboradores estudiaron un conglomerado conformado por ocho naciones en el occidente europeo, incluyendo 24,2 millones de apellidos. Se detectó un mayor valor de endogamia por isonimia en España y un valor mínimo en Francia. Además, se registró una similitud general entre las naciones en cuanto a la abundancia de apellidos medida a través del α de Fisher. Esta diversidad se analizó tanto a nivel de bloque subcontinental, como también entre los países y las localidades comprendidas en su interior (Scapoli et al., 2007). El trabajo concluye que la estructura de la población de Europa Occidental se ve influida por la acción conjunta de la deriva y la migración, siendo las diferentes lenguas un factor de aislamiento tan importante como las altas cadenas montañosas.

Barrai y colaboradores estudiaron la estructura isonímica de Estados Unidos, en base a 18 millones de apellidos registrados en 247 municipios (Barrai et al., 2001). Este conjunto de apellidos incluye a los usuarios del servicio de telefonía para el año 1996, listados provistos por las compañías. Se observó que la geografía del país parece no ser un obstáculo para los movimientos humanos. A diferencia de estudios realizados en Venezuela y Europa, el aislamiento calculado utilizando apellidos no crece con la distancia. Según el α de Fisher, en promedio, las regiones hacia el centro-sur son más endogámicas que las del norte. Otro ejemplo de análisis de gran magnitud son los desarrollados en la República Popular China (Liu et al., 2012), en donde se estudió la estructura de la población a través de los 7,327 apellidos portados por 1,280,000,000 personas. Se georeferenciaron los valores de la isonímia a nivel de provincia, prefectura y condado y se detectaron claros patrones espaciales. Los valores bajos de isonímia seguían el camino medio y bajo del Río Yangtze, fenómeno explicado por los largamente documentados movimientos migratorios internos. En un trabajo posterior, para el mismo punto temporal, se detectaron concentraciones de ciertos apellidos en zonas específicas, como la región costera del sur, claramente diferenciada del norte y noreste de China (Meng et al., 2016). A partir de conjuntos extensos de datos, se analizó la correlación espacial de la frecuencia de apellidos, revelando que la variabilidad de nombres de familia disminuye a medida que aumenta la distancia. En un trabajo más reciente, se presentó un nuevo índice de migración denominado índice Single-

Regional Migration Intensity (SRMI), que midió la intensidad histórica de los movimientos poblacionales (Fan et al., 2023). Se utilizaron los mismos 7,327 apellidos portados por 1,280,000,000 personas, cubriendo 362 prefecturas. Las prefecturas con valores similares de SRMI tienden a encontrarse cercanas en el mapa administrativo, dando lugar a zonas bien diferenciadas. La intensidad migratoria es mayor en el noreste, seguido por la llanura central y la cuenca del río Yangtsé, mientras que es menor en el sur, zona en la que se establece una mayor población de minorías étnicas. El trabajo resalta la efectividad del enfoque basado en apellidos ante el desafío de la escasez de otros datos. Su aplicación, además de revelar patrones migratorios diversos, se valida de manera mutua con el índice Coverage Ratio of Stretched Exponential Distribution (CRSED), otro indicador de migración basado en apellidos propuesto en (Chen et al., 2019).

2.7. Estudios isonímicos en Latinoamérica y Argentina

Uno de los primeros trabajos para América Latina fue el de Azevedo y colaboradores, (Azevedo et al., 1969) que estudiaron 1068 familias del noreste de Brasil en el año 1962. En aquella investigación se relevó información de una muestra poblacional, con el objetivo de estudiar mortalidad y morbilidad bajo control de variables. Los autores detectaron niveles diferenciales de endogamia y consanguinidad, tanto a través de los apellidos como de análisis de polimorfismos sanguíneos y de estudios genealógicos. Se concluyó entonces que las comunidades del noreste tienen una preferencia regional a establecer matrimonios isonímicos, que prevalece más allá de la distancia real que separe a los cónyuges. Décadas después, Cardoso dos Santos y colaboradores analizaron en profundidad las características poblacionales en la misma región del Brasil (Cardoso-dos Santos et al., 2021), ya que su tendencia a presentar altas tasas de endogamia, así como una alta incidencia de enfermedades autosómicas recesivas, ha persistido a lo largo del tiempo. La elaboración de políticas de salud pública enfocadas a la vigilancia epidemiológica se enfrenta al desafío de sortear la vasta extensión territorial, la heterogeneidad social y las complejidades económicas. El estudio de apellidos permitió una aproximación a la estructura genética de la población, analizando datos de más de 37 millones de personas junto con indicadores demográficos y sanitarios, en los 1,794 municipios de los nueve estados de

la región. Se encontró una correlación positiva entre el índice de isonimia y la frecuencia de nacidos vivos con anomalías congénitas y una frecuencia significativamente mayor de apellidos repetidos en la región del Quilombo dos Palmares, el mayor conglomerado de esclavos fugitivos en América Latina.

Estudios similares se han publicado con datos de Bolivia (Rodriguez-Larralde et al., 2011), Paraguay (Dipierri et al., 2011), Chile (Barrai et al., 2012), Honduras (Herrera Paz et al., 2014), Uruguay (Carrieri et al., 2020) y Venezuela (Rodríguez-Larralde et al., 2000)

En Argentina se ha analizado la estructura demográfica a partir de los apellidos en diferentes niveles. Por ejemplo, el estudio de Dipierri y colaboradores (Dipierri et al., 2005a) investigó la estructura isonímica del país y sus subdivisiones utilizando un corpus de 414.441 apellidos distintos portados por los 22,6 millones de electores registrados en el padrón electoral del año 2001. La distribución geográfica del α de Fisher en los 541 departamentos reveló una diversidad significativamente mayor hacia el este y menor en el oeste del país. Los autores concluyeron que la estructura poblacional es reflejo de fenómenos migratorios sobre los procesos de deriva genética. A nivel regional, también con la información del mismo padrón electoral, se calculó la isonimia de los 117 departamentos pertenecientes a la región Noroeste, que agrupa las provincias de Jujuy, Salta, Catamarca, La Rioja y Tucumán (Dipierri et al., 2007, 2005b). Entre los 2.576.548 votantes se observó mayor diversidad de apellidos y bajo grado de parentesco en los departamentos ubicados en el centro de la región, mientras que en los periféricos se dio un patrón opuesto, con mayor aislamiento. La estructura isonímica está en concordancia con el historial de asentamiento y las características geográficas del Noroeste argentino (NOA). En el año 2009, Bromberg y colaboradores examinaron exclusivamente los apellidos de la Ciudad Autónoma de Buenos Aires (Bronberg et al., 2009). Se utilizó información de 2,552,359 electores para estimar el coeficiente de parentesco por isonimia de Lasker, la diversidad de apellidos según el α de Fisher, el coeficiente de consanguinidad resultante de la isonimia aleatoria, y las distancias de Nei, de Lasker y Euclidiana. Estas distancias se correlacionaron con las distancias geográficas, que se calcularon asignando un punto arbitrario a cada distrito y midiendo los kilómetros entre ellos en un mapa de la ciudad. Los promedios de la distancia isonimica de Lasker y la isonímia aleatoria de los distritos ubicados al sur de la Avenida Rivadavia fueron mayores que los situados al norte de dicha aveni-

da. Se observó una correlación significativa entre la distancia geográfica y las distancias de Nei y Euclidiana, indicando una subdivisión de la población metropolitana, con mayor consanguinidad y menor variedad de apellidos en los distritos ubicados en el sector sur de la ciudad. Esta estructura concuerda con la fragmentación y las diferencias sociales, culturales y económicas entre los diferentes barrios de esta metrópolis latinoamericana.

2.8. Fuentes de datos para la isonimia

Como se mencionara en el capítulo 2, en los estudios isonímicos se suelen emplear listados de directorios telefónicos, registros de bautismos o actas de uniones matrimoniales (Colantonio et al., 2003). Igualmente, otros conjuntos que sobresalen dentro de la diversidad de fuentes, son los padrones electorales. Estos registros ofrecen datos considerablemente más representativos, abarcando una mayor fracción de la población y manteniendo un grado más alto de estandarización y periodicidad en su publicación. En nuestro país, estos documentos están encuadrados bajo la Ley Nacional N° 22.864 (modificatoria del Código Electoral Nacional o Ley N° 19.945), que desde el año 2012 considera como electores a todos "(...) los argentinos nativos y por opción, desde los dieciséis (16) años de edad, y los argentinos naturalizados, desde los dieciocho (18) años de edad", siendo el sufragio un deber individual y universal (P.E.N. Codigo Electoral Nacional, 1972). La Cámara Nacional Electoral tiene la potestad y obligación de confeccionar el Registro de Electores en formato digital y documental, constando, para cada persona, de "(...) apellidos y nombres, sexo, lugar y fecha de nacimiento, domicilio, profesión, tipo y número de documento cívico, especificando de qué ejemplar se trata, fecha de identificación y datos de filiación. Se consignará la condición de ausente por desaparición forzada en los casos que correspondiere". El soporte documental del padrón incluye además huellas dactilares, fotografía y firma de los electores. Todos estos datos están ordenados por distrito y sección, resultando fácilmente georeferenciables. Aunque se trata de información sensible, su acceso o análisis está exento de la obligación de evaluación de un comité de ética, ya que fueron datos tomados para el ejercicio de funciones propias de los poderes del Estado y/o por una obligación legal. No obstante, en el desarrollo de esta tesis, así como en los estudios elaborados durante los años formativos del doctorado, las comunicaciones

en diferentes foros académicos y las publicaciones de los artículos científicos, se observó plena atención a la Ley 25.326 de Protección de Datos Personales (Ministerio de Justicia, 2000). A través de peticiones cursadas a la Cámara Nacional Electoral, se obtuvieron copias de los padrones de los años 2015 y 2021, los que conforman el corpus de datos que servirá de insumo principal. La información fue exclusivamente relativa a los apellidos de los electores, su sexo y su distribución en los establecimientos de votación, dejando ausente cualquier otro dato potencialmente identificatorio y/o de naturaleza sensible.

Antes de emprender el cálculo de cualquier indicador, es necesario aplicar operaciones de limpieza y preparación sobre los datos. Estas operaciones involucran transformaciones sobre el dataset inicial que contiene los datos crudos y son de complejidad variada. Algunas están orientadas a eliminar registros duplicados, renombrar valores o extraer apellidos a partir de nombres de familia. Otras tareas son más complejas, como la asignación de códigos de unidad administrativa según el lugar de votación. Argentina se encuentra organizada administrativamente en cinco regiones: Noroeste, Noreste, Cuyo, Centro y Patagonia. Contenidas en estas regiones se reparten 24 jurisdicciones: 23 provincias y un distrito federal, Ciudad Autónoma de Buenos Aires (CABA). Cada una de estas jurisdicciones está organizada a su vez en 530 divisiones de segundo orden: 380 departamentos, 135 partidos (provincia de Buenos Aires) y 15 comunas (CABA). Es importante entonces poder enmarcar dentro de esta jerarquía a cada uno de los registros de los conjuntos de datos con los que trabajamos, para tener análisis completos en los tres niveles de organización territorial: regional, provincial y departamental.

Para representar la información se utilizaron las capas cartográficas o de información geoespacial, en formato vectorial, puestas a disposición en el catálogo de datos espaciales del Instituto Geográfico Nacional. Estos datos se encuentran expresados en coordenadas geodésicas, utilizando el Sistema de Referencia WGS 84 y el Marco de Referencia POSGAR 07 (Código EPSG:4326) (Instituto Geográfico Nacional, 2019).

Es esencial considerar aquí también la estructura jerárquica ya mencionada de las divisiones administrativas dentro del territorio. Esto es, cuáles departamentos (unívocamente identificados) están incluidos dentro de cuáles provincias (unívocamente identificadas), del mismo modo que entre las provincias y las regiones. Toda la información georreferenciada debe disponerse estandarizada, de forma tal que se permita entrelazar registros de diferen-

tes fuentes (como censos, registros de salud, entre otros). Para este cometido, se utilizó la Application Programming Interface (API) del Servicio de Normalización de Datos Geográficos (Ministerio De Modernización, 2016). Esta interfaz permite normalizar y codificar los nombres de unidades territoriales contenidos en los registros de nuestros conjuntos de datos. Este es un paso muy importante, ya que muchas veces las provincias, departamentos, municipios o localidades de la Argentina pueden ser indicados de diversa manera. Por ejemplo, para referirse a la provincia de Santiago del Estero, podríamos encontrar los términos "Stgo. del Estero", "S. del Estero" o "Sgo. del Estero"; el departamento Doctor Manuel Belgrano de la provincia de Jujuy podría encontrarse como "General Manuel Belgrano", "Manuel Belgrano", o simplemente "Belgrano"; o, en casos más enredados, un mismo nombre podría ser compartido entre más de una unidad, como los departamentos Rawson presentes en las provincias de San Juan y también de Chubut. La API permite resolver estos problemas de forma sencilla, en virtud del cumplimiento de la Política de Datos Abiertos impulsada desde el Gobierno de la Nación Argentina en 2016, a través del Decreto N° 117/2016 (Dirección Nacional de Datos e Información Pública, 2016). En secciones posteriores, cuando se presenten las distintas propuestas de software de análisis de datos que han sido implementadas, se describirán con mayor detalle las tareas de limpieza y preparación que debieron cumplirse.

2.9. La informática y la isonimia

Entre los lenguajes interpretados, Python ha desarrollado una amplia y activa comunidad de computación científica y análisis de datos. En los últimos años, Python ha pasado de ser un lenguaje de computación científica de vanguardia a uno de los lenguajes más importantes para la ciencia de datos, el aprendizaje automático y el desarrollo general de software en el mundo académico y la industria (McKinney, 2012). El fortalecimiento del soporte de Python para bibliotecas como `Pandas` y `SciKit-Learn`, sumado a su impulso en el ámbito de la ingeniería de software de propósito general, ha consolidado su posición como una elección prevalente para realizar tareas de análisis de datos.

Para nuestras labores se diseñó e implementó, como núcleo software, un paquete Python denominado *"isonymy"* que permite, dados los apellidos de un conjunto de perso-

nas, realizar el cálculo de frecuencias y obtener rápidamente los indicadores isonímicos. Dicho paquete ha sido además publicado en el repositorio de software para el lenguaje de programación Python llamado Python Package Index (PyPI), como muestra la Fig. 2.3.

El paquete propuesto se trata de una biblioteca liviana que expone operaciones para obtener los indicadores por separado, o todos de una vez:

- *get_unbiased_random_isonymy*: Retorna la isonímia no sesgada (I) a partir de un listado de apellidos dado.
- *get_fishers_alpha*: Retorna el valor del α de Fisher a partir de un listado de apellidos recibido como parámetro.
- *get_a_index*: Retorna el valor del indicador A según un listado de apellidos recibido como parámetro.
- *get_b_index*: Retorna el valor del indicador B a partir de un listado de apellidos recibido como parámetro.
- *get_isonymic_indicators*: Retorna una estructura (dict) que contiene todos los valores de los indicadores listados previamente.
- *get_wright_m_vect(total_pop_serie, fst_serie)*: 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: 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.

  Esta función es primordial para la elaboración de los gráficos log-log de apellidos.

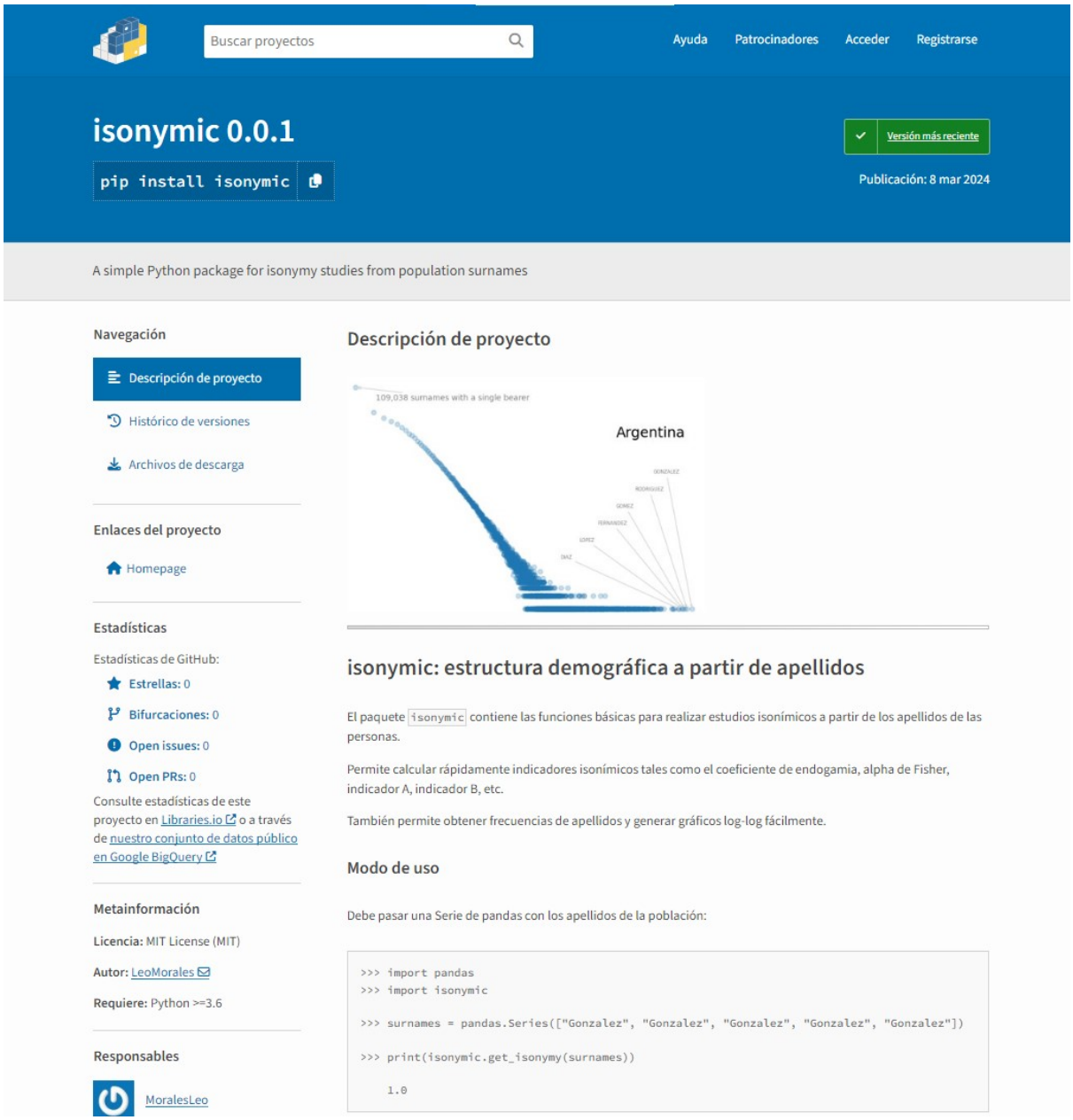


Figura 2.3: Paquete isonímico publicado en el portal PyPI, repositorio oficial de software para el lenguaje de programación Python.

- get_distances(surnames_i, surnames_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.

Este paquete conforma el medio principal a partir del cual se acelera el análisis de la isonímia de la población, variando las listas con los apellidos de las personas según la población que estamos analizando. Como bien se describió anteriormente en la sección *"Fuentes de datos para la isonimia"*, al conformar un conjunto de datos ordenado, con una etiqueta apropiada de unidad administrativa, la operatoria de obtención de información isonímica en distintos tamaños poblacionales se ve muy facilitada, alternando entre los distintos niveles administrativos de Argentina: nacional, regional, provincial y departamental.

Para el desarrollo de los artefactos de software del presente trabajo, se utilizó el conjunto de herramientas de software libre ofrecidas por la comunidad del lenguaje de programación Python. Aquellas más relevantes se describen a continuación. Para el análisis de datos se utilizó el popular paquete `Pandas`, que consiste en una biblioteca de operaciones sobre datos tabulares en forma de dataframes. Para la visualización de los datos se utilizaron `MatPlotLib`, `Seaborn` y `Bokeh`. Las tres son librerías muy populares que permiten generar gráficos de un amplio abanico de variedades. Para el trabajo con datos geoespaciales, se utilizó `GeoPandas`, la versión de `Pandas` acondicionada para trabajar con datos espaciales. Para la contextualización de las visualizaciones en mapas se utilizó el paquete `Contextily`, que permite obtener mapas de mosaicos (o mapas base) de Internet. Para la escritura del código, pruebas y su mantenimiento se utilizó la herramienta de ciencia de datos `Jupyter Notebook`. Esta es una herramienta web que permite la creación y elaboración de documentos interactivos denominados cuadernos (o notebooks). Estos documentos se construyen a partir de la inserción de celdas que pueden contener código

Python, texto plano, texto enriquecido, imágenes, texto `HyperText Markup Language (HTML)`, entre otras. Esta posibilidad los dota de una gran expresividad al momento de su presentación. Cada celda de código es ejecutable y dicha ejecución es independiente de la ejecución de las demás, por lo cual el control es responsabilidad del usuario. Su entorno de ejecución base puede ser tanto local como en la nube. Para la construcción de procedimientos ordenados sobre datos se utilizó el paquete `Ploomber`, que consiste en un marco de trabajo para disponer los cuadernos Jupyter o las funciones Python y generar procedimientos de transformaciones encadenadas sobre los datos. Este es un paquete muy importante en el desarrollo de esta tesis por lo cuál su forma de uso se describe más adelante, en el Capítulo 3, cuando se aborden los casos de uso que necesitan tratamientos ordenados de los datos.

2.10. Bulsarapp

Considerando la gran cantidad de información producto del procesamiento automatizado de datos demográficos a partir de padrones electorales, se desarrolló un prototipo de software orientado a la exploración interactiva de la isonimia en Argentina. En esta sección nos proponemos exponer y analizar en detalle dicho artefacto de software, que ha sido desarrollado en el formato de una aplicación web y al que hemos denominado `Bulsarapp`. Esta aplicación constituye la capa de presentación final de un pipeline de visualización de información isonímica Morales et al. (2021). Con ello se busca responder a la demanda de especialistas en esta área de estudio, quienes señalaron la necesidad de contar con una plataforma exploratoria ágil y modular. En la actualidad, no existen otros desarrollos similares.

En la literatura especializada sobre el tema encontramos diferentes propuestas de visualización de la información que van desde presentaciones sencillas para mostrar distribución de apellidos, referencias geográficas y estadísticas básicas hasta visualizaciones más complejas de los diferentes índices, resumiendo los resultados en tablas o gráficos estáticos que no son interactivos ni accionables. Bulsarapp, por su parte, incorpora muchas de estas visualizaciones de los índices de población a partir del método isonímico, que resultan intuitivas, en una única interfaz. En este producto, se destaca la presentación de la distri-

bución espacial de apellidos y de la estimación de los orígenes geolingüísticos para todos ellos.

En la etapa del diseño de Bulsarapp se efectuaron diversas reuniones con académicos y expertos en estudios de apellidos, quienes identificaron las tareas y problemas de investigación más comunes. Se definieron entonces los tres objetivos principales siguientes.

Exploración de tendencias y relaciones isonímicas: Era necesario disponer de una herramienta que facilitara la exploración de valores isonímicos en distintos niveles y unidades territoriales, su distribución geográfica, el estudio de las relaciones entre ellos y la identificación de valores límites, extremos o atípicos.

Traza de apellidos: Los apellidos pueden analizarse conformando grupos según varios criterios, como se ha mencionado anteriormente. Por ejemplo encontramos aquellos que comparten una derivación o fuente etimológica común (basada en toponimia, ocupaciones, apodos o características físicas), otros tienen una trazabilidad hacia un origen geográfico específico, un punto crucial en la historia, o asociados a diferentes grupos de ancestría. La representación visual de la distribución de estos grupos de apellidos resulta fundamental para complementar y caracterizar los estudios isonímicos. De esta manera, los investigadores pueden identificar conjuntos y enfocar sus análisis en aspectos específicos de las poblaciones. Estos análisis pueden ser especialmente útiles en investigaciones sobre enfermedades hereditarias, como se verá en los capítulos siguientes. Asimismo, el estudio de la distribución geográfica de los apellidos contribuye a ampliar la comprensión de los procesos migratorios.

Tendencia espacial del origen de los apellidos: En Argentina, al igual que en otros países latinoamericanos y antiguas colonias, el mestizaje de su población repercute en una diversidad de apellidos notable. La herramienta a desarrollar debía permitir a los investigadores explorar la frecuencia de un origen geolingüístico particular en una región, para compararla con la de otros orígenes o contextualizarla con otra información demográfica, ya sea propia o no de los apellidos. Para lograr esto, se definieron y diseñaron las vistas necesarias, así como su interacción y coordinación. Y como sucede en el punto anterior, este estudio de la distribución espacial de los oríge-

nes de los apellidos contribuye significativamente a la comprensión de los procesos migratorios.

Con el propósito de brindar una solución informática que atendiera a las necesidades planteadas, se identificaron las transformaciones necesarias sobre un conjunto de datos inicial que incluya los apellidos de una población. Tal conjunto de datos primario correspondía al registro electoral argentino del año 2015. Las transformaciones a las que se hace referencia conformaron un pipeline de visualización que dio lugar a una aplicación web. En su primera versión, la aplicación web Bulsarapp permite al usuario la exploración interactiva de los cinco indicadores isonímicos, presentados en el apartado "Indicadores a partir de los apellidos": Endogamia por isonímia, estimador de consaguinidad, α de Fisher y los indicadores A y B. Se sumó también el cálculo de la cantidad de personas (N), la cantidad de apellidos diferentes (S) y el resultado del cálculo de distancia de Lasker entre provincias y departamentos. Permite examinar la dispersión espacial tanto de un conjunto de apellidos de interés como la distribución geográfica de orígenes geo-lingüísticos de apellidos. Otorga las interacciones necesarias para indagar esta información a nivel nacional, provincial y departamental.

Aunque los registros electorales no fueron inicialmente creados con la intención de ser utilizados para análisis geográficos, su información espacial puede ser reconstruida de manera sencilla mediante un proceso de georreferenciación usando los distintos campos de sus registros. Suman un total de 13 campos, que pueden ser el apellido, el género o la información necesaria para ubicar geográficamente a cada individuo en términos de departamento, provincia y región, a partir del lugar que tenga asignado para votar.

El modelo de visualización de datos elegido especifica cuatro pasos generales que se aplican sobre los datos crudos hasta obtener las visualizaciones objetivo. Una visión general de este proceso se presenta en la Fig. 2.4. Así, un primer paso consiste en las transformaciones necesarias para convertir los datos crudos en datos abstractos. A partir de estos últimos otro paso especifica cuáles serán los datos de presentación. La definición del qué se va a mostrar depende del caso de uso, por lo cual en el pipeline se pueden ver tres ramificaciones acorde a los tres objetivos generales que se describieron antes. Un tercer paso especifica el mapeo visual de los datos para adecuarse a la semántica de los gráficos elegidos. Un último paso aplica el mapeo visual para así obtener las representaciones fina-

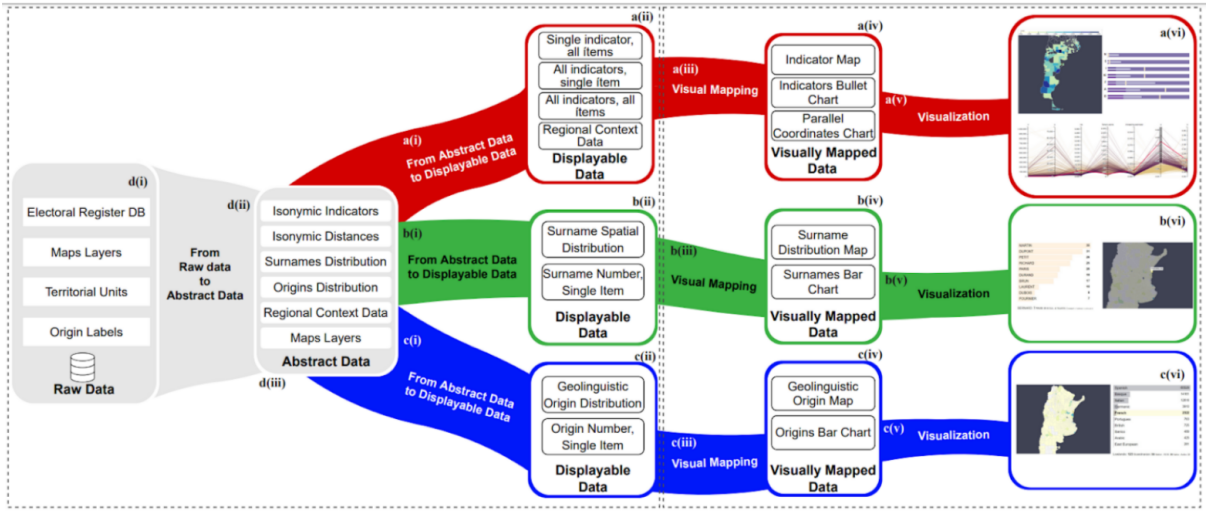


Figura 2.4: Representación gráfica del pipeline de visualización de información isonímica propuesto. Dos recuadros punteados separan la especificación de los conjuntos de datos reales a mostrarse (a la izquierda), de la definición de cómo se mostrarán (a la derecha). Las ramas roja, verde y azul representan transformaciones sobre los datos y las cajas representan las etapas resultantes (intermedias y finales) de la aplicación de tales manipulaciones.

les. Como se ha sugerido en esta descripción, entre cada paso existen diferentes instancias que atraviesan los datos. A continuación se describen estas etapas de los datos.

Cuatro conjuntos principales componen el cuerpo de datos crudos que sirven de insumo de entrada. Al ya mencionado padrón electoral, se le suman:

- Las capas geográficas, a partir de las cuales se podrán representar regiones, provincias y departamentos.
- La codificación de la jerarquía de unidades territoriales, especificada según la división administrativa de la Argentina (necesaria para contextualizar los valores isonímicos de una unidad con respecto a la unidad que la contiene).
- Las etiquetas de origen de apellido, que se tratan de listas con pares apellido-etiqueta de origen a partir de dos fuentes principales, (Monasterio, 2017) y (Albeck et al., 2017). Así, el origen de cada apellido del padrón, siempre que fuera posible de identificar, fue uno de 27 orígenes distintos.

Cinco de las categorías fueron consideradas del trabajo de Monasterio (ibérico, italiano,

japonés, alemán y este europeo). Es de destacar que en su labor se constituye un ejemplo paradigmático que ejemplifica de manera precisa la diversidad de fuentes de datos empleadas en este tipo de estudio con orígenes asociados a apellidos. El autor reunió apellidos a partir de documentación proveniente del Museo de Immigración de San Pablo en Brasil, de censos de América del Norte, de estudios sobre el tráfico de esclavos al sur de Brasil, la Heráldica de Apellidos Españoles, y de los servicios de ciudadanía italiana en Brasil, entre otras (Monasterio, 2017). A partir de estas fuentes, conformó un corpus de 71,791 pares de apellido y ancestría, que utilizó para aplicar métodos de Fuzzy Matching y un método de categorización de textos basado en N-gramas propuesto por Cavnar y Trenkle (Cavnar et al., 1994). Del listado total, 30,405 pares se ofrecen de forma pública en el portal web del trabajo. El resto de las categorías de orígen fueron obtenidas a partir del laborioso esfuerzo de Albeck y colaboradores (Albeck et al., 2017). En su investigación el origen de los apellidos se determinó mediante criterios geográficos y lingüísticos. Primeramente la clasificación se realizó a partir del conocimiento empírico y de búsqueda bibliográfica, complementandose posteriormente con documentación colonial (de Jujuy y Tucumán) para el caso de los apellidos autóctonos, y bases digitales disponibles en internet para los de origen foráneo. Los apellidos autóctonos se subclasificaron según su origen geográfico en: Andinos (Bolivia, Chile, Perú y Ecuador), Jujeños (registrados en documentos coloniales de la Puna de Jujuy y Quebrada de Humahuaca entre 1557 y 1786), Tierras Altas del Noroeste Argentino (incluidos en documentos históricos del sector andino de Salta, Catamarca, La Rioja y Tucumán), Tierras Bajas de Noroeste Argentino (Santiago del Estero y tierras bajas de Salta, Jujuy y Tucumán), Sureños (antropónimos registrados desde la provincia de Córdoba y San Juan hacia el Sur) y Sin Clasificar. Los apellidos foráneos se agruparon en categorías menores según su origen geográfico y/o lingüístico en: africanos, alemanes, árabes, armenios, belgas, británicos, coreanos, chinos, escandinavos, españoles, este europeo, franceses, griegos, hebreos, hindúes, holandeses, italianos, japoneses, portugueses, suizos, tailandeses y vascos. Con las cinco clasificaciones autóctonas y las 22 foráneas (entre Albeck y Monasterio), se obtuvieron las categorías de orígenes utilizadas en este trabajo.

Sobre el padrón electoral se aplicaron las tareas de limpieza mencionadas en el apartado "Fuentes de datos para la isonimia" de este capítulo, obteniendo un apellido a partir de cada

registro individual del padrón. Ese nombre de familia está vinculado a un departamento unívoco, que a su vez permite trabajar en el resto de los niveles administrativos, provinciales y regionales.

El conjunto de datos abstractos se obtiene a partir de transformaciones sobre los datos crudos, siendo:

- Indicadores isonímicos, para todas las unidades de todos los niveles administrativos del país, calculados a partir del padrón electoral correctamente limpio y ordenado.
- Datos de contexto regional. Con los indicadores isonímicos calculados y la jerarquía de divisiones administrativas, se generó una tabla de valores de contexto. Sabiendo que en todos los niveles administrativos (menos en el último), cada unidad es a su vez contenedora de otras, se generó una tabla con los datos para cada administración, los valores máximo, mínimo y medio de cada indicador según las unidades contenidas. Por ejemplo, para la región Patagonia, esta tabla contiene el valor máximo, mínimo y medio del estimador de consanguinidad entre las provincias de Neuquén, Rio Negro, Chubut, Santa Cruz y Tierra del Fuego. Este mismo enfoque se aplica a todos los indicadores y para las otras cuatro regiones. Luego se aplica la misma metodología para el nivel nacional.
- Distancias isonímicas, en forma de matrices que reflejan las distancias de Lasker entre departamentos y entre provincias.
- Distribución de apellidos: por cada unidad, en todos los niveles administrativos, se calcula cuántas personas son portadoras de cada apellido distinto encontrado en el padrón.
- Distribución de orígen: por cada unidad, en todos los niveles administrativos, se calcula cuántas personas son portadoras de un apellido distinto con un origen geo-lingüístico específico, según se encuentre dicho nombre de familia en el listado unificado de orígenes.
- Representaciones geográficas de las unidades en capas cartográficas, con códigos estandarizados para poder combinarse con los datos provenientes del padrón.

A partir de este punto, el proceso se separa en tres subramas de acuerdo a las tres tareas principales a llevar a cabo: Visualización isonímica, Distribución geográfica de apellidos y Distribución geográfica de orígenes de apellidos. Esta división se ilustra con ramas de colores distintos en la Fig. 2.4.

En el primer caso, de visualización isonímica, los indicadores y sus valores de contexto confluyen en tres formas diferentes de gráficos según se trate de la exhibición individual o de la presentación colectiva de la información. La Fig. 2.5 muestra la interfaz de esta sección. Los valores de los indicadores isonímicos de cada unidad territorial se representan en un mapa de coropletas, con la granularidad correspondiente al nivel en análisis (provincial o departamental). Los mapas son los dispositivos visuales ideales para representar de manera gráfica y espacial la distribución, relaciones y patrones de datos (MacEachren and Ganter, 1990). A través de un selector (Fig. 2.5a), se especifica el indicador para el cual se desea explorar y detectar algún patrón geográfico. Cuando el usuario elige otro indicador en el selector, éste se colorea según la nueva opción. Los indicadores isonímicos de todas las unidades territoriales se presentan en una vista de coordenadas paralelas (Fig. 2.5b). Esto permite a los usuarios tener una aproximación a todos los índices y luego filtrar y ver detalles a voluntad (una de las reglas principales de la visualización) para explorar las relaciones entre ellos. Los ejes pueden reordenarse para facilitar la comparación entre dos o más indicadores. También es posible establecer rangos en cada eje para limitar los valores isonímicos y explorar cómo repercute esta operación sobre los rangos y distribuciones de los demás índices. Las líneas contenidas dentro del rango son más oscuras. Cuando se indican rangos, se genera un criterio de filtro. Los elementos que quedan fuera del criterio ya no aparecen coloreados en el mapa (sólo se conservan sus contornos). De este modo, los usuarios pueden visualizar geográficamente el conjunto especificado de relaciones de datos. El mapa se colorea automáticamente según los lugares cuyos valores entran dentro de los criterios. Esto resulta extremadamente útil, por ejemplo, para detectar e identificar poblaciones aisladas.

Cuando se selecciona una determinada región en el mapa (Fig. 2.5d), la línea correspondiente a la unidad territorial se resalta con un color diferencial para el elemento seleccionado en la vista de coordenadas paralelas. Debajo se completa un detalle que incluye el nombre del departamento, la provincia o región a la que pertenece la unidad te-

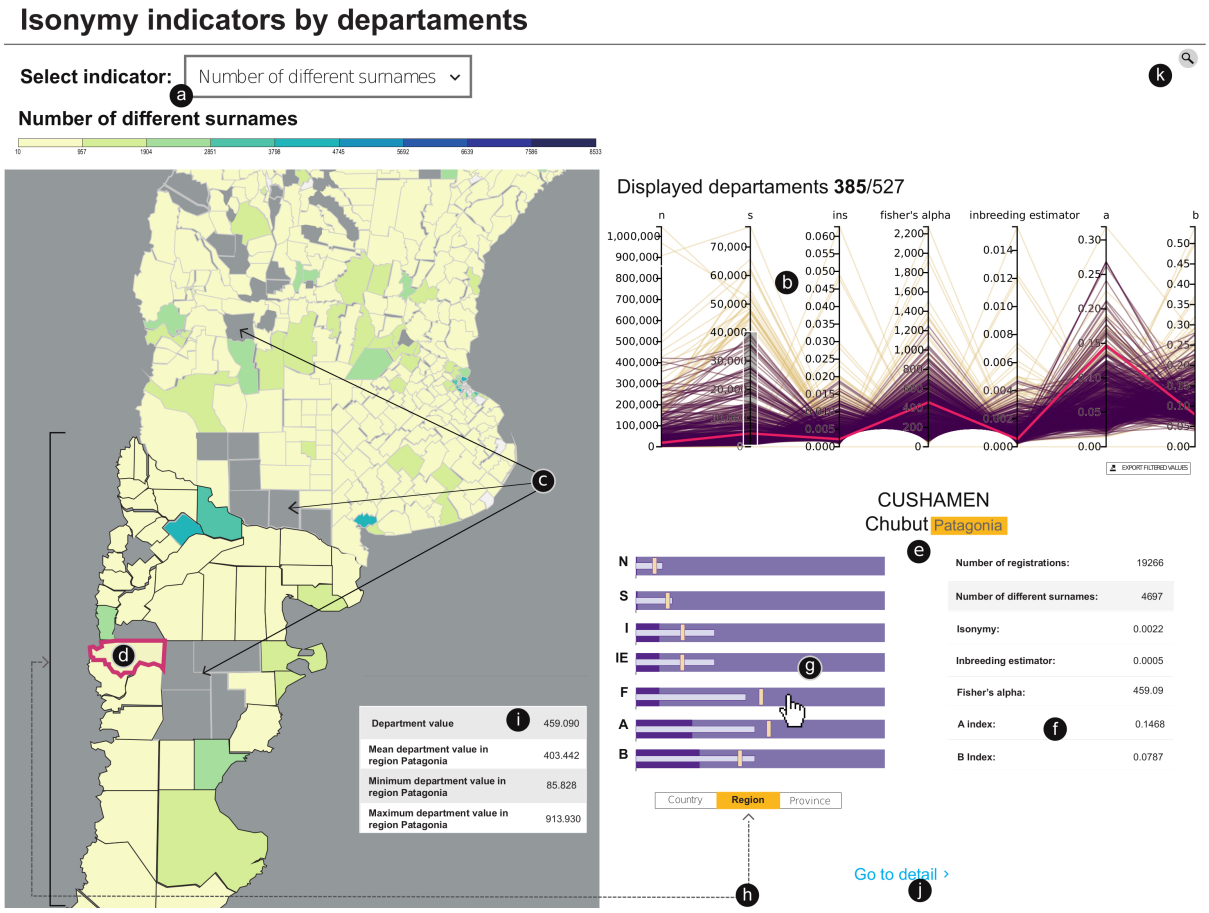


Figura 2.5: Interfaz de exploración de la información isonímica a nivel departamental en la aplicación web Bulsarapp. El usuario puede interactuar con el mapa eligiendo cualquiera de los siete índices isonímicos desde el selector (a) y/o estableciendo rangos en los ejes de un gráfico de coordenadas paralelas (b). La imagen muestra un rango establecido en los valores del segundo eje. La línea correspondiente a la unidad territorial seleccionada en el mapa aparece resaltada en un color diferente en la vista de coordenadas paralelas. El usuario puede elegir una vecindad para contextualizar a la unidad seleccionada entre tres opciones: país, región o provincia (h). El área de contexto aparece resaltada con bordes más oscuros en el mapa. Cuando el usuario pasa el ratón por encima de cualquiera de las barras de este gráfico de viñetas, aparece un tooltip con los valores de referencia de la barra (i). Se presenta un botón para pasar a la sección de detalles (j). Se muestra un botón de búsqueda en la esquina superior derecha para permitir la búsqueda por nombre de departamento/provincia (k).

rritorial seleccionada (Fig. 2.5e), y los respectivos valores de índices isonímicos (Fig. 2.5f) en una tabla y en un gráfico contextualizado. Este gráfico se trata de una visualización de viñetas (en inglés, *bullet chart*) que permite al usuario comparar valores isonímicos de la unidad seleccionada con los valores máximo, mínimo y medio su contexto o vecindario (Fig. 2.5g). Debajo del gráfico aparece un selector con tres opciones contextuales, que el usuario puede cambiar según su interés, e indicar el nivel administrativo deseado (Fig. 2.5h). Entonces, los valores contextuales dependen de la jerarquía geográfica seleccionada (provincial, regional o nacional). Estos valores contextuales se dibujan en barras de distintas longitudes. La barra más larga representa el valor máximo del indicador en el contexto de la unidad seleccionada. La barra más oscura con el mismo grosor representa el valor más bajo del indicador en el vecindario. La barra más fina representa el valor medio. El marcador, o viñeta (punto sobre la barra), representa el valor del indicador para la unidad territorial seleccionada. Los bordes de los ítems en cada vecindario aparecen resaltados en el mapa para proporcionar contexto geográfico. Cuando el usuario hace clic en un nuevo elemento del mapa, el gráfico de viñetas se actualiza con nuevos marcadores. Estas tres vistas, mapa, coordenadas paralelas y gráfico de viñetas están vinculadas de forma coordinada.

En otra vista en la aplicación (y otra rama del pipeline) se presenta la salida correspondiente a la distribución de apellidos. El usuario debe especificar el conjunto de apellidos a abordar, la consulta se envía al servidor y con la respuesta se elaboran dos gráficos: un mapa de gradiente de colores y un gráfico de barras (Fig. 2.6a). Cada valor isonímico de las entidades individuales se mapea visualmente a una forma (polígono) del mapa con un color asociado. El color dependerá de la cantidad de portadores que sumen los apellidos consultados. Junto al mapa aparece una tabla resumen con las provincias y departamentos con mayor número de portadores. Cuando se selecciona una unidad del mapa (Fig. 2.6a(i)), se muestra el nombre de la unidad y número de individuos portadores de alguno de los apellidos de la consulta (Fig. 2.6a(ii)). Debajo, se despliega una lista, ordenada por prevalencia, presentada en un gráfico de barras horizontales según la cantidad de portadores de cada apellido del conjunto ingresado (Fig. 2.6a(iii)).

La tercera vista presenta la distribución por orígenes, presentada en la Fig. 2.6b, para la exploración a nivel departamental. La herramienta genera un mapa de gradiente coloreado

por frecuencia de un origen geolingüístico especificado a través de un selector desplegable. A un lado del mapa se ofrece la información de resumen para el origen seleccionado: Cantidad de personas que portan un apellido de tal origen, qué porcentaje representa esa cantidad sobre el total de la población, un ranking (top 3) con la cantidad de portadores para el origen en provincias y otro (top 10) en los departamentos. Cuando se selecciona una unidad del mapa (Fig. 2.6b(i)), se despliega al costado un resumen con las cantidades de apellidos por cada origen de la base de datos, ordenadas en un gráfico de barras horizontales. La Fig. 2.6b(ii) muestra este momento. En el gráfico desplegado se resalta la barra que corresponde al origen que el usuario ha seleccionado (Fig. 2.6b(iii)).

El usuario puede acceder a un informe detallado de cada unidad cuando la selecciona en el mapa. Esta sección de detalle muestra un resumen con información de cada una de las tres tareas principales. Se incorporan también las distancias calculadas según Lasker (a partir de los datos abstractos) y se presenta un mapa con el gradiente de colores según los valores mínimos y máximos de las diferencias de las demás unidades con respecto a la unidad detallada. Para medir el tiempo de consulta general utilizando las facilidades incluidas en la herramienta Bulsarapp, se desarrollaron una serie de experimentos que tenían como objetivo hacer navegar a los usuarios (investigadores del área de la isonimia) por las tres secciones del desarrollo web y luego completar un cuestionario. Incluía 20 tareas, cada una con una pregunta de opción múltiple asociada. En las preguntas se caracterizaban poblaciones (provincias y departamentos) a través de sus valores isonímicos, rastreaban un grupo de apellidos en una unidad territorial concreta y se indagaba sobre los orígenes de los apellidos en diferentes zonas. Al finalizar, se solicitó completar una encuesta con una escala de usabilidad del sistema clásica, en inglés System Usability Scale (SUS), con preguntas en escala Likert (del 1 al 5 según nivel de satisfacción).

En la Tabla 2.1 se listan las tareas de la experimentación, agrupadas según las vistas de las funcionalidades ofrecidas por Bulsarapp: índices isonímicos (T1-T6), sección de detalle de isonimia (T7-T13), frecuencias de apellidos (T14-T18) y distribución de orígenes geo-lingüístico de apellidos (T19, T20). El experimento fue realizado por 9 participantes (Mujeres=4, Hombres=5, edad media=35), concluyendo que la presentación a través del formato aplicación web acorta los tiempos de consulta. Los nueve participantes, con conocimientos básicos de isonimia pero que nunca habían tenido contacto con la aplicación,

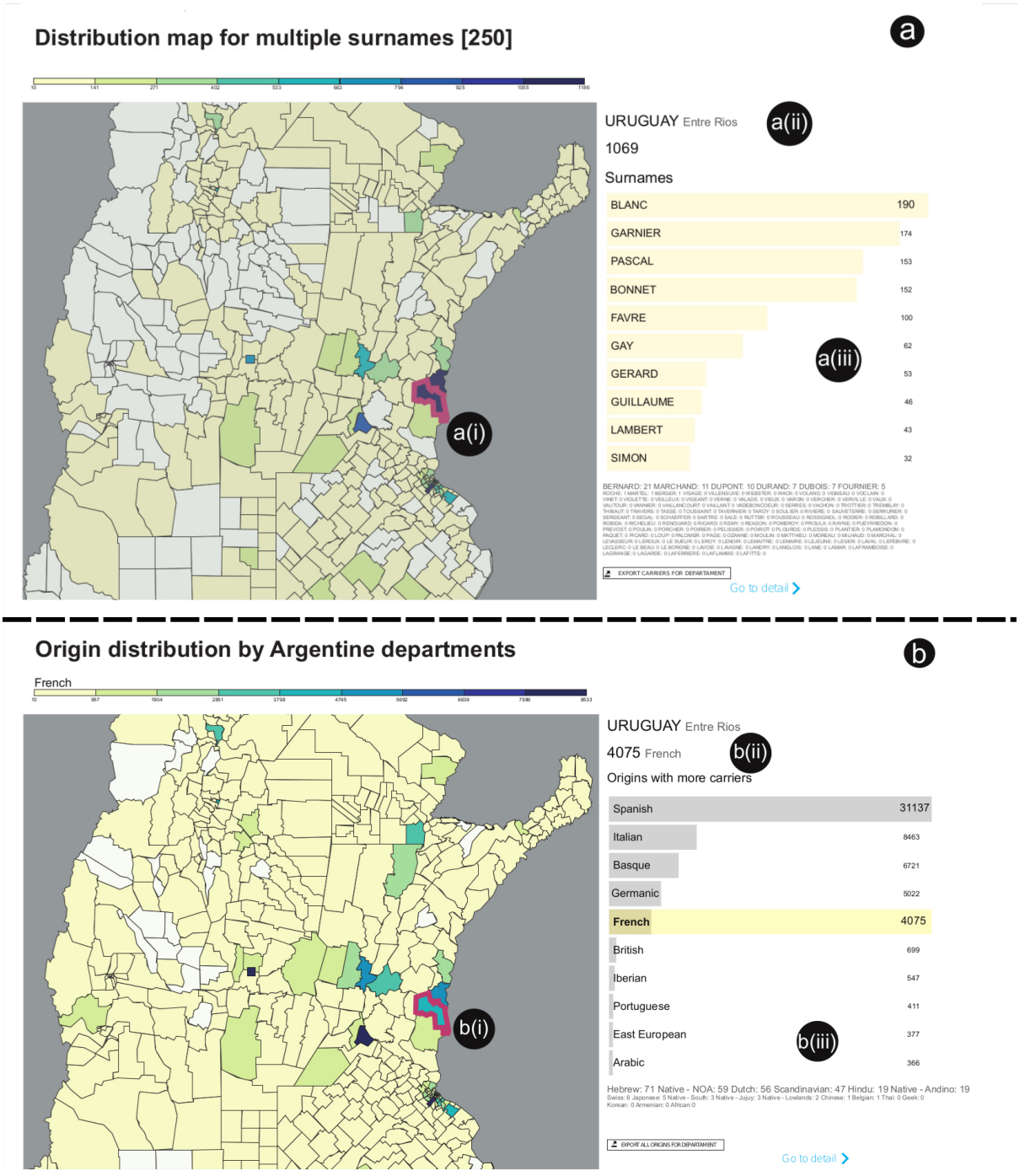


Figura 2.6: Dos interfaces de la aplicación para las distribuciones de un conjunto de apellidos y para un origen específico, ambas a nivel departamental, y con patrón de distribución espacial similar. A la izquierda (a), se muestra el resultado de una consulta de portadores de un conjunto de 250 apellidos de origen francés recogidos de la web. A la derecha de la figura (b), se muestra la visualización de la distribución de apellidos etiquetados con origen bien conocido, en este caso francés. En ambas, el detalle se muestra para el departamento de Uruguay, provincia de Entre Ríos.

	ID	Descripción	Opciones
1	T1 y T2	Indicar cuál es la región del país con los valores mas altos para un índice específico	Dos opciones categóricas: "Hacia el sur" o "Hacia el norte"
2	T3 y T6	Indicar el valor de un índice específico para una provincia o departamento	Tres opciones numéricas
3	T4 y T5	Indicar si los valores de un indicador exceden un límite específico	Verdadero o falso
4	T7	Indicar cuáles son los apellidos más populares en un departamento	Tres listas de apellidos
5	T8	Indicar cuántos portadores tiene un apellido específico en Argentina	Tres listas de apellidos
6	T9	Indicar la cantidad de personas que portan un apellido categorizado con un origen específico	Tres opciones numéricas
7	T10 y T11	Indicar si el valor de un índice en un departamento es mayor que la media provincial del mismo índice	Verdadero o falso
8	T12	Indicar si, para un departamento en particular, aquella menor distancia Lasker hasta otro departamento pertenece a un departamento vecino	Verdadero o falso
9	T13	Determinar si la distancia Lasker entre dos departamentos dados corresponde al valor mas alto en comparación con la registrada hacia el resto de los departamentos	Verdadero o falso
10	T14 y T15	Dado un conjunto de apellidos, determinar la cantidad de personas que portan alguno de ellos en una provincia/departamento dado	Tres opciones numéricas
11	T16	Dado un apellido específico, determinar la cantidad de portadores el total de la población de un departamento	Tres opciones numéricas (porcentajes)
12	T17	Dado un conjunto de apellidos, determinar cuáles son los departamentos/provincias en donde existe mayor cantidad de portadores	Tres listas de provincias/departamentos
13	T18 y T19	Determinar la cantidad de portadores de un apellido dado en un departamento en particular	Tres opciones numéricas
14	T20	Determinar si un apellido posee el origen que mas portadores tiene en una región	Verdadero o falso

Tabla 2.1: Consignas de la experimentación sobre la aplicación web Bulsarapp.

respondieron las 20 preguntas necesitando en promedio de 28 minutos y 30 segundos, con una tasa de acierto de 158/160. Posteriormente a la experimentación, los usuarios otorgaron coeficiente de usabilidad de promedio 88,4.

El uso de herramientas de visualización interactiva como la presentada en esta sección es una vía prometedora para apoyar estudios multidisciplinares sobre poblaciones humanas. Bulsarapp posibilita la investigación dinámica a la vez que proporciona representaciones de los apellidos mediante vistas enlazadas. Los estudios isonímicos constaban tradicionalmente de reportes estáticos y/o generalizados sobre la información de los apellidos. Por lo tanto, si el objetivo es caracterizar una nueva población por su isonimia, los investigadores deben tener acceso a los datos y realizar los cálculos necesarios nuevamente desde cero. Bulsarapp hace posible una consulta rápida de la información desde tres perspectivas vinculadas: la estructura isonímica, la frecuencia de apellidos a partir de un conjunto de variables especificado y el número de portadores de apellidos de un origen geolingüístico concreto. Además, es posible la exploración espacial mediante el análisis de unidades territoriales. También es extremadamente útil en la detección de valores atípicos y descripciones de tendencias, representando las relaciones entre los valores isonímicos de una unidad territorial dentro de un radio determinado.

Bulsarapp significó nuestra primera aplicación de tareas de minería de datos sobre los padrones electorales argentinos. Este software de analítica de datos, encargado de la extracción, transformación, carga de los apellidos, con su posterior tratamiento para la visualización de información y la generación de reportes, constituye un pipeline general que será ampliado incorporando nuevos padrones como datos de entrada. Estos resultados se presentan en el siguiente capítulo, para abordar el estudio de las migraciones a través de los apellidos y la relación que los desplazamientos o aislamientos de población tienen en la salud general.

Capítulo 3

Apellidos y migraciones

3.1. Introducción

Estudiar la dinámica poblacional constituye uno de los aspectos más relevantes del análisis demográfico y el más complejo en el plano metodológico. El crecimiento de una población está determinado por cuatro factores:

- A: Nacimientos.
- B: Fallecimientos
- C: Número de emigrantes.
- D: Número de inmigrantes.

La diferencia entre nacimientos y muertes constituye el crecimiento vegetativo o natural de un área determinada en un período definido. Puede ser positivo (si los nacimientos superan a las defunciones) o negativo (si la mortalidad es mayor que la natalidad). La razón entre emigrantes e inmigrantes nos da el saldo migratorio, puede también ser positivo o negativo, aumentando o disminuyendo el tamaño efectivo de la población en un período de tiempo determinado (Foschiatti, 2011). En un conjunto de tamaño considerable, como la población entera de un Estado o país, el crecimiento vegetativo suele ser el factor más importante, aunque pueden registrarse excepciones temporales. Está fuertemente interrelacionado con los procesos de desarrollo o modernización económica, dando lugar a una gran variabilidad en el ritmo de los cambios de la fecundidad y de otras variables relacionadas (estado nutricional, salud de la población, conductas matrimoniales y de planificación familiar). Volveremos a hablar sobre esta relación entre factores económicos y salud poblacional al final de este capítulo.

El número de emigrantes y de inmigrantes son el conjunto de la migración, un factor muy importante en la evolución biológica de las comunidades humanas. Modifica la estructura demográfica y genética y junto con la mortalidad y la fecundidad, moldea la estructura de la pirámide poblacional. A diferencia de estos últimos dos factores, la migración es un proceso que puede ser reversible. Puede presentarse en diferentes grados, desde un traslado (pasando un límite geográfico o no) hasta un cambio de residencia dentro de la misma área delimitada. Dicha mudanza puede ser temporal o permanente.

Estudiar la distribución de la población en el territorio conforma un problema político para todas las sociedades, incluyendo la preocupación por las relaciones entre la producción y distribución de los alimentos, el acceso equitativo a infraestructura que garantice la calidad de vida o resolver los problemas derivados del aumento de la densidad poblacional, entre otros (Busso, 2007).

Al migrar, las personas llevan consigo sus identidades, entendidas tanto en sus múltiples aspectos biológicos e históricos como en el puramente apelativo, es decir, sus nombres y apellidos. En el actual contexto organizacional de nuestro país la movilidad de las personas deja un rastro burocrático, con una división espacial precisa y bajo el control territorial por parte de los estados provinciales y del estado nacional. Esta huella administrativa se materializa en múltiples registros estatales, desde los censos hasta los padrones electorales y son materia de análisis complementarios a los estudios demográficos.

Si la migración continúa en el tiempo, se convierte en uno de los principales factores del mestizaje. Argentina, al igual que otros países de Latinoamérica, tiene una historia de extenso mestizaje que se dio en forma gradual e intermitente, con características particulares en cada región. En el período de migraciones masivas pueden identificarse dos momentos, el primero entre fines del siglo XIX y mediados del XX, y el segundo desde mediados hasta finales del XX. El primero se caracterizó por contar con mayoría de los inmigrantes europeos, mientras que durante el segundo la migración de ultramar paulatinamente disminuye hasta casi desaparecer y predominan los inmigrantes latinoamericanos, en especial, de países limítrofes.

Por su parte, las migraciones internas fueron y son de considerable magnitud. Estos movimientos tuvieron una importancia creciente hasta 1960-1970, punto desde el cual comenzaron a descender pero nunca han cesado. Se estima que desde 1895 cambiaron

de provincia unos 7 millones de personas. La dinámica migratoria no es un mero efecto o consecuencia de un fenómeno económico, sino que genera nuevos resultados que se retroalimentan, tanto en las áreas receptoras como en las emisoras (Velázquez and Lende, 2004).

Analizar la dinámica de la migración es una estrategia para estudiar el patrimonio cultural y biológico propio de cada región desde una perspectiva interdisciplinaria, utilizando un dato único de las poblaciones humanas: la herencia de apellidos. Como se ha presentado en el capítulo anterior, desde su formalización en la Edad Media y su posterior expansión mundial como procedimiento de control de la propiedad y la población durante el periodo colonial, los apellidos son rasgos culturales que se transmiten entre ancestros y descendientes. Este mecanismo de herencia sigue un sistema vertical comparable a la transmisión de algunas variantes genéticas.

Además del apellido, los integrantes de una familia comparten un acervo génico común. Sin embargo, dado que la transmisión de un nombre no está exenta de alteraciones, la utilidad de los apellidos como indicadores de filiación puede verse comprometida. Un ejemplo son los posibles errores en la grafía al transcribir documentos, que pueden resultar en dos apellidos distintos para un único vínculo familiar. A medida que se suceden las generaciones, los nombres continúan y la memoria oral puede llegar a perder el registro del nexo biológico original. El uso del apellido del cónyuge una vez establecido un matrimonio o las adopciones son otros ejemplos. En estas circunstancias, aunque las personas estén vinculadas jurídicamente a través del nombre como una familia, la conexión no se corresponde a nivel biológico. Cuando se trabaja con grandes volúmenes de datos, como los padrones electorales en los países en los que el voto es obligatorio, el efecto de estos casos de disrupción en el nexo apellido-vínculo biológico es desestimable.

Además de parentesco, los apellidos permiten evaluar el aislamiento, el sedentarismo, la semejanza y relación de poblaciones en un área geográfica determinada (Castro de Guerra et al., 1990). Al ser posible establecer un origen lingüístico, étnico o geográfico más probable, los apellidos también son marcadores de identidad de grupo (Alford, 1987) y pueden ser rastreados a través del tiempo, destacándose su utilidad en los estudios sobre minorías étnicas. Marcadas discontinuidades en su distribución espacial pueden ser producidas por migración reciente o pasada y también por relocalización forzada de grupos

en contextos bélicos, en tensión sociopolítica o por desastres ambientales, con el consecuente flujo génico posterior (Sokal et al., 1992, Mascie-Taylor and Lasker, 1985). La comparación de las tasas de migración inferidas por apellidos con las proporcionadas por las fuentes demográficas tradicionales, indica que es posible obtener estimaciones confiables de los patrones migratorios recientes y de los cambios en la distribución geográfica de las poblaciones subdivididas (Piazza et al., 1987, Mourrieras et al., 1995).

Aquí se presentan dos casos de análisis sobre Argentina, que combinan diversas fuentes de datos, aplicando los estadísticos y metodologías presentadas en el capítulo anterior. En el primero se estudian las consecuencias biológicas de la tendencia al aislamiento, luego de una migración de larga data en un contexto multicultural. En el segundo, se identifican los patrones espaciales y modificaciones de la estructura poblacional producto de las migraciones internas, es decir, aquellas que se producen entre provincias o regiones del mismo país.

3.2. Fuentes de datos utilizadas

3.2.1. Estadísticas Vitales

En Argentina, la Dirección de Estadísticas e Información de la Salud (DEIS) es el nivel nacional del Sistema de Estadísticas de Salud, dependiente del Ministerio de Salud de la Nación. Fue planificada para responder a objetivos primordiales, entre los que se destacan:

- Producir, difundir y analizar estadísticas relacionadas con condiciones de vida y problemas de salud, suministrando datos sobre Hechos Vitales (Nupcialidad, Natalidad y Mortalidad), Morbilidad y Rendimientos Hospitalarios.
- Aplicar en todo el territorio normas y procedimientos uniformes para la captación de la información, la elaboración y el procesamiento de los datos.
- Difundir y publicar la información en todos los niveles nacionales y también a los organismos internacionales encargados de la difusión de estadísticas.

El Anuario, denominado *"Estadísticas Vitales-Información Básica"* es la publicación que reúne la información estadística sobre los hechos vitales –nacimientos, defunciones, defunciones fetales y matrimonios- ocurridos en Argentina.

Para que dicha publicación sea posible, deben observarse una serie de pasos protocolizados a diferentes niveles, empezando por el local. El personal de salud de los establecimientos certifica los hechos y capta los datos básicos a partir de los instrumentos de recolección de datos normalizados. Los registros civiles y sus delegaciones inscriben y registran legalmente los hechos vitales, es decir, todos los acontecimientos relacionados con el comienzo y fin de la vida de los individuos y con los cambios en su estado civil que pueden ocurrir durante su existencia. El siguiente nivel es el jurisdiccional, las unidades de Estadísticas Vitales y de Salud de las provincias y de la Ciudad Autónoma de Buenos Aires deben recibir, controlar, codificar y elaborar los datos, suministrando anualmente los archivos a nivel nacional. La DEIS, como responsable última, es la encargada de elaborar, publicar y difundir las estadísticas sobre hechos vitales para el total del país.

3.2.2. Clasificación Internacional de Enfermedades

Poder contar con instrumentos de recolección de datos normalizados es resultado de un largo proceso. Las estructuras administrativas de los Estados, en especial después de la Segunda Guerra Mundial, se vieron enfrentadas al problema urgente de llevar estadísticas precisas y comparables sobre la morbi-mortalidad poblacional. La herramienta consenso es la Clasificación Internacional de Enfermedades, para el registro y notificación nacional e internacional de las causas de enfermedad y muerte. Tiene su origen en la Clasificación Bertillon o Lista Internacional de Causas de Muerte, instrumento estadístico del 1893. La Clasificación Internacional de Enfermedades (CIE) es un nomenclador que estandariza códigos como una combinación de letras y números. En 1967, la XX Asamblea Mundial de la Salud definió que el certificado médico de defunción debía incluir "todas aquellas enfermedades, estados morbosos o lesiones que produjeron la muerte o contribuyeron a ella, y las circunstancias del accidente o de la violencia que produjo dichas lesiones". La responsabilidad de quien firma el certificado es indicar la afección que condujo directamente a la defunción -denominada causa básica- y establecer las condiciones antecedentes o sucesos que le dieron origen. Las estadísticas de mortalidad son una de las principales fuentes de información demográfica y en muchos países constituyen el tipo de dato de salud más confiable. La versión en vigor del CIE es la décima, desarrollada en 1992. Se divide en capítulos, desde I a XXII, según el conjunto de enfermedades que agrupe. Por

ejemplo, el capítulo II comprende todas las neoplasias y el XVII todas las malformaciones congénitas y anomalías cromosómicas. En forma anual se realizan actualizaciones menores y cada tres años se introducen cambios mayores. Algunos países, incluyendo Argentina, han creado sus propias extensiones del código CIE-10.

En este capítulo se utilizaron los registros de la Dirección de Estadísticas e Información de la Salud, desagregados a nivel nacional, provincial y departamental, de todos los fallecimientos y las causas entre los años 2005 y 2014. Estas bases de datos -y muchas otras- se publican de forma abierta y están disponibles en una plataforma web [1].

3.2.3. Indicadores demográficos

Otra plataforma en línea que integra distintos indicadores demográficos y los pone a disposición para acceso público es la del Instituto Nacional de Estadísticas y Censos (INDEC). Permite consultar datos sobre la evolución de la población para cada una de las jurisdicciones del país desde el primer censo nacional en 1869 hasta el último, en 2022. Reúne indicadores de enorme sensibilidad, como las tasas de crecimiento intercensal, el crecimiento vegetativo, la natalidad, mortalidad, fecundidad, la esperanza de vida, la mortalidad infantil, neonatal y posneonatal. Las series elaboradas a partir de estadísticas vitales comprenden un período de más de cien años. Para poder estimar el número efectivo poblacional y calcular las tasas de migración interna, se emplearon las estimaciones anuales desagregadas por nivel administrativo, así como los valores finales de los censos [2].

3.2.4. Bases OMIM (Online Mendelian Inheritance in Man)

Existe una relación clara y evidente entre la movilidad de la población, la migración y la epidemiología de algunas enfermedades (Gushulak and MacPherson, 2006). Aunque históricamente se ha prestado mayor atención a las enfermedades transmisibles, en tiempos recientes, el incremento del impacto global de los movimientos migratorios ha generado un renovado enfoque hacia otros elementos relativos a la salud poblacional y las enfermedades no transmisibles vinculadas con factores genéticos.

Las enfermedades de base genética constituyen un grupo de patologías muy importante, tanto por su incidencia como por el tipo de problemas que producen. Comprometen

[1] https://www.argentina.gob.ar/salud/deis
[2] https://www.indec.gob.ar/indec/web/Nivel4-Tema-2-41-158

la calidad de vida de los afectados, causando graves discapacidades. Es frecuente que estas enfermedades tengan además un carácter progresivo y condicionen una mortalidad precoz (Strachan and Read, 2018). Aunque en la mayor parte de los casos aún no se dispone de soluciones terapéuticas definitivas, las mejoras en los diagnósticos y tratamientos se deben a los avances de la capacidad tecnológica. La posibilidad de estudiar cada componente de la molécula de ADN llevó al hito de secuenciar el genoma humano por completo. De forma paralela a estos avances en biomedicina, hubo un enorme desarrollo en las técnicas de comunicación y de acceso a la información, que llega prácticamente en tiempo real a profesionales, investigadores y público general (Pennisi, 2001). La base de datos Mendelian Inheritance in Man (MIM) y su versión electrónica Online Mendelian Inheritance in Man (OMIM) es un registro de conocimiento sobre genes humanos y sus enfermedades asociadas. Fue iniciada en 1960 y hoy contiene información sobre todos los trastornos conocidos, incluyendo más de 16.000 genes. Es un compendio completo de variantes génicas y fenotipos asociados (es decir, las características observables de un individuo que son consecuencia de la expresión génica). Cada código de identificación está seguido del nombre o de los distintos nombres que recibe la enfermedad. En una tabla se presenta el fenotipo y su código, el código del gen, la incidencia de la enfermedad, sus bases moleculares, síntomas, fecha en que la enfermedad fue descrita por primera vez, etc. Está disponible en forma gratuita y se actualiza diariamente (OMIM, 1966-2024).

Además, se utilizan aquí nuevamente los cuatro conjuntos de datos principales que componen el cuerpo de datos crudos o insumo de entrada, tal como se presentaron en el Capítulo 2:

- Padrón electoral
- Capas geográficas para representar regiones, provincias y departamentos.
- Codificación de la jerarquía de unidades territoriales, especificada según la división administrativa de Argentina, para contextualizar los valores isonímicos.
- Etiquetas de origen geográfico/étnico/lingüístico más probable para cada uno de los apellidos, a partir de listas validadas.

3.3. Metodologías de extracción de información

La analítica de datos combina la aplicación de principios matemáticos y estadísticos junto con programas informáticos con el propósito de obtener conocimiento a partir de conjuntos de datos. Una de sus formas es la analítica descriptiva, que se centra en caracterizar eventos del pasado y sugerir sus potenciales causas subyacentes mediante la consulta básica, la elaboración de informes y la visualización de datos históricos (Fleckenstein et al., 2018). Los análisis realizados tanto en el Capítulo 2 como en el Capítulo 3 de la presente tesis se enmarcan dentro de este enfoque. Todos los escenarios descritos comparten la característica fundamental de haber integrado datos demográficos y de salud de la población argentina con el propósito de identificar patrones que expliquen -o ayuden a explicar- distintos fenómenos de interés.

La analítica de datos permite obtener información que no es fácilmente deducible a partir de los datos crudos e incluye las tareas preliminares de búsqueda y extracción de datos, así como su posterior preparación. Aprovechando el desarrollo de diversos métodos de visualización, producto del contínuo progreso en las técnicas de representación gráfica, es posible obtener una nueva vista a partir de estos datos ordenados, en busca de una comprensión mas pulida del objeto de estudio. Estas técnicas son capaces de transformar la información -a veces implícitamente oculta- en figuras que ayudan a comprender más fácilmente su significado (Nelli, 2015). Algunos autores insertan este proceso en un marco más general denominado DIKW, por las siglas en inglés de las palabras Datos, Información, Conocimiento y Sabiduría. En su forma más simple, y sin importar su volumen, los datos son la materia prima de la que se deriva la información ("Information" en el acrónimo DIKW). Luego, la información debe sistematizarse y analizar sus patrones para generar conocimiento ("K" de "Knowledge" en DIKW). Un paso final en el proceso integra y sintetiza el conocimiento que se ha obtenido sobre el fenómeno de interés para producir sabiduría ("Wisdom", en DIKW). Dada la gran cantidad de dominios de aplicación, existen muchas definiciones de marcos de trabajo, pero destacamos el enfoque DIKW porque nos permite explicitar el valor de trabajar en equipos interdisciplinarios, combinando una gran variedad de perfiles informáticos (analistas, desarrolladores, expertos en bases de datos, etc) con investigadores de otras áreas de la ciencia.

También suele denominarse minería de datos al esfuerzo de reunir una gran base de

datos y estudiar posibles patrones de respuesta en un sistema puntual. Dentro de esta minería se destaca un proceso crucial conocido como Extract, Transform, and Load (ETL), responsable de la extracción, limpieza y transformación de datos para su posterior análisis. Este proceso garantiza la calidad y confiabilidad de los datos a utilizar (Han et al., 2012) y es fundamental para la generación de información significativa a partir de datos crudos. "La transformación y estructuración adecuadas permiten a los analistas centrarse en la exploración de patrones y relaciones, en lugar de perder tiempo en la manipulación manual de datos desorganizados" (Linoff and Berry, 2011).

Dentro de este contexto, se debía facilitar la construcción del conocimiento demográfico basado en los apellidos de las personas. Desarrollamos entonces lo que se conocen como "tuberías de datos" o `pipelines`. Estas tuberías reúnen, en forma específica y detallada, los pasos para el procesamiento de datos. Son producto de identificar y diseñar las transformaciones necesarias, con un orden de ejecución y salida que mejor se ajusten a las preguntas que se busca responder. Para su implementación debe utilizarse el lenguaje de programación que ofrezca el marco de trabajo más adecuado.

Para todos los pipelines de datos desarrollados en esta tesis se utilizó la biblioteca Ploomber, en su versión 0.20, del lenguaje de programación Python. Este paquete permite realizar tareas de minería a partir de la definición de las transformaciones que queremos aplicar sobre los datos. En especial, permite diagramar e implementar los pasos del ETL de forma controlada, bien documentada, de fácil mantenimiento, prueba, crecimiento y despliegue.

Como se mencionó en el capítulo anterior (apartado "La informática y la isonimia"), los notebooks de Jupyter son una herramienta fundamental para el trabajo de ciencia de datos. Ploomber surge sobre el ecosistema de herramientas propuesto por los Jupyter Notebooks, con el objetivo de disminuir la dificultad en el paso desde estos documentos dinámicos hacia procesos activos y robustos para la analítica de datos. El trabajo intensivo en notebooks Jupyter puede generar documentos monolíticos difíciles de mantener, cuyo formato además no se complementa muy bien con los sistemas de control de versiones de código fuente.

A continuación se describe de forma muy breve el framework propuesto por Ploomber para nuestros objetivos. El código encargado de completar un procesamiento de los datos

se denomina tarea. Cada tarea se implementa entonces en forma de funciones Python. La definición del pipeline se realiza incorporando los nombres de las tareas (el nominal y el nombre específico de la función Python) a un archivo en formato YAML. Este es el archivo principal del framework Ploomber, y es regido por un formato que nos permite listar las tareas del pipeline y especificar otros campos además del nombre y de la ubicación de la función Python que implementa dicha tarea, como la ubicación en disco en donde se guardará el resultado de su ejecución o parámetros que modifiquen su comportamiento. Así, el pipeline se construye de forma clara, trazable y mantenible. Los resultados que genera cada tarea -a los que se denomina productos- pueden tener formato `csv`, `parquet`, `binary`, `pdf`, entre otros. Estas tareas implementadas como funciones reciben además un argumento especial, denominado upstream, que les permite trabajar sobre productos de otras tareas. A través del upstream y sabiendo el nombre de las tareas previas necesarias, es posible acceder a los resultados ya computados para combinarlos, filtrarlos, agregarlos, etc. y generar nuevas representaciones.

Ploomber ofrece un comando que ejecuta todo el pipeline definido y obtiene los productos de cada tarea. Éstas se ejecutan sólo cuando es necesario, es decir, la primera vez, al definirlas e incorporarlas, y luego solo ocuparan recursos de procesamiento cuando se re-ejecuten por la aplicación de algún cambio en su implementación. En posteriores ejecuciones del pipeline entero, se reutilizan los productos ya procesados. Cómo última mención acerca de esta herramienta, se destaca la posibilidad para indicar parámetros generales, es decir, parámetros disponibles para todas las tareas del pipeline completo. Éstos son accesibles de forma muy sencilla y cualquier cambio en sus valores forzará la re-ejecución del pipeline para las tareas que hagan uso de ellos, y así en cascada según la dependencia de tareas que se haya definido.

Ploomber sirvió de soporte para todos los proyectos de datos desarrollados en esta tesis. Resultó fundamental para el mantenimiento de los procesos de ETL y las tareas de minería de datos en un contexto de proyectos en ejecución simultánea y de grado de complejidad diverso. Las tareas de extracción resultaron muy variadas ya que dependen del estado de madurez de la fuente principal de cada conjunto de datos. Por ejemplo, los padrones electorales se presentan en formato de base de datos transaccional y su extracción se realiza a través de Structured Query Language (SQL) mientras que otros conjuntos de

datos, como los sanitarios, están conformados por archivos tabulados separados por algún carácter específico y son accesibles de forma más directa a través de operaciones con paquete Pandas. Otros recursos se encuentran accesibles a través de internet y suponen esfuerzos de web scrapping para su obtención, como en el caso de algunos listados de apellidos y sus orígenes. Tareas de este estilo se convierten en tareas Ploomber, contenidas de una forma adecuada en funciones python individuales específicas.

Las tareas de transformación para la preparación y limpieza de datos, como puede ser la identificación y el manejo de valores faltantes, duplicados o ruidosos conformarán una tarea de limpieza que dará como resultado el dataset de trabajo principal para cada proyecto. Las transformaciones más comunes son el filtrado, normalización, estandarización, agregación, y conversión de tipos y se aplican según la naturaleza de cada problema. Una de las bondades del trabajo con pipelines simultáneos es la posibilidad de obtener retroalimentación en un proyecto que sirva para incorporarse en otros o corregir errores que fueron pasados por alto al inicio del proceso.

Por último, las tareas de carga en nuestro caso han sido de carácter heterogéneo cuyos resultados abarcan desde archivos con datos ordenados (Wickham, 2014) aptos para crear visualizaciones o la aplicación de modelos estadísticos, hasta elementos en bases de datos, como la variante de base de datos no estructurada utilizada en la aplicación Bulsarapp expuesta en el capítulo anterior. A continuación se describen de forma detallada dos casos de uso que integran apellidos y demografía en respectivos proyectos de minería de datos.

3.4. Caso 1 - Alemanes del Volga en Argentina: implicancias sanitarias de la migración y el aislamiento poblacional

3.4.1. Introducción, fuentes de datos y objetivo

La enfermedad o mal de Alzheimer recibe su nombre por el psiquiatra alemán Alois Alzheimer, quien en 1906 realizó la primera publicación donde describe los síntomas que había observado en sus pacientes (pérdida de memoria, desorientación, alucinaciones) combinado con el análisis post mortem de los cerebros, encontrando placas de una sustancia anómala (amiloide) y acumulaciones de materia fibrosa en las neuronas. La enfermedad de Alzheimer es la principal causa de demencia y se está convirtiendo rápidamente

en una de las enfermedades más costosas y letales de este siglo, en particular para los países desarrollados, cuya pirámide demográfica es de tipo regresivo. Esta enfermedad es un continuo que se extiende durante un período de 15 a 25 años. Al principio la patología puede estar presente sin ningún síntoma, pasando luego por una etapa de deterioro cognitivo leve hasta una demencia manifiesta y el fallecimiento posterior. Inicialmente, el diagnóstico llegaba en la etapa de demencia, un síndrome clínico caracterizado por un deterioro cognitivo progresivo sustancial que afecta a varios dominios, o síntomas neuroconductuales de suficiente gravedad como para causar un impacto funcional evidente en la vida diaria.

Una persona con demencia ya no es completamente independiente, y esta pérdida de independencia es la principal característica que diferencia la demencia del deterioro cognitivo leve (Jack Jr et al., 2018). El diagnóstico de la enfermedad ha pasado de tener un enfoque clínico y excluyente, a un encuadre clínico y biológico combinado, que incorpora determinación por biomarcadores y diagnóstico por imágenes. Los estudios de gemelos mostraron que el riesgo de enfermedad de Alzheimer depende en un 60-80 % de factores hereditarios. Los estudios de genealogías familiares indican que variantes raras específicas tienen un efecto causal en la enfermedad de Alzheimer, en algunos casos con una edad de inicio tan temprana como los 40 años. Esta forma familiar de la enfermedad presenta una proporción que varía entre el 1 % al 5 % de los casos, es autosómica dominante y está causada por mutaciones en los genes PSEN1 (presenilina 1) (MIM *104311) y PSEN2 (presenilina 2) (MIM *600759). El PSEN1 está asociado a la enfermedad de Alzheimer de tipo 3 (MIM 607822) y el PSEN2 a la enfermedad de Alzheimer de tipo 4 (MIM 606889) Las formas familiares se caracterizan por un inicio precoz y una sobrevida menor (Scheltens et al., 2021). La forma esporádica de la enfermedad representa el 95 % de los casos y no tiene una causa genética conocida. Es de aparición tardía, se presenta en pacientes mayores de 65 años y se considera dentro del proceso degenerativo del envejecimiento natural (Andrade-Guerrero et al., 2023).

En 1988 se describieron casos de Alzheimer en 5 familias residentes en Estados Unidos. Todas eran descendientes de un grupo de inmigrantes conocidos como los alemanes del Volga, que llegaron al país entre 1870 y 1920. La enfermedad, confirmada por autopsia y heredada como un rasgo autosómico dominante, se observó tanto en hombres como

en mujeres durante varias generaciones (Bird et al., 1988). Sus antepasados se habían mudado de Alemania a la región del sur del Volga, en Rusia, en la década de 1760. A su vez, estos antepasados eran descendientes de personas que, antes de la migración a Rusia, vivían en dos pequeñas aldeas alemanas adyacentes y compartían varios apellidos.

Para comprender el contexto en que se da esta migración, debemos retroceder hasta la Guerra de los Siete Años, un conflicto bélico internacional por el control de colonias en América del Norte e India, que comenzó en 1756 e involucró diversos imperios y reinos, estructuras políticas asentadas en territorios que en la actualidad conforman otros estados-nación, entre ellos, Alemania. Catalina II, emperatriz del Imperio Ruso, era de ascendencia alemana y en 1763 promovió una serie de beneficios para los emigrantes con el fin de que la colonización del Volga (regiones fronterizas cercanas a Asia) resultara atractiva para la debilitada población alemana de la posguerra. Estos beneficios incluían libertad religiosa, exención temporal de impuestos, préstamos sin intereses, autogobierno interno y exención permanente del servicio militar obligatorio. Entre 1764 y 1769, oleadas de colonos, principalmente de Hesse, un estado de Alemania cuya capital actual es Wiesbaden, llegaron al bajo Volga, cerca de la ciudad de Saratov.

En el curso de una década, se fundaron 106 colonias de base económica agrícola. La zona se volvió rápidamente próspera, las colonias se expandieron pero permanecieron sin mayores contactos con el resto de Rusia, siendo una parte más del complejo mosaico étnico que componía el imperio zarista. La población total de las colonias llegó a 22.246 habitantes (Pohl, 2009). En 1864, el zar Alejandro II modificó el acuerdo original y exigió a los inmigrantes alemanes que se inscribieran en el servicio militar. Este acontecimiento marcó el inicio de la diáspora hacia Estados Unidos, Canadá, Australia y Brasil. En estos nuevos contextos se mantuvo su relativo aislamiento, fundamentalmente por barreras idiomáticas y religiosas que persisten hasta la actualidad.

En 1878, 1.100 alemanes del Volga llegaron a Argentina. Se concentraron, por orden de importancia, en las provincias de Entre Ríos, Buenos Aires, Santa Fe, Chaco y Córdoba, donde se fundaron numerosas colonias rurales que aún existen. La inmigración continuó hasta la Primera Guerra Mundial. Según estimaciones de asociaciones de descendientes, hay 2.000.000 de personas herederas de la migración Volga en Argentina. Las historias individuales están bien documentadas por las propias prácticas de auto-afirmación de los

descendientes, que llevan registro de las fechas de ingreso de las familias al país, sobre las colonias en las que se establecieron, sobre los casamientos y fallecimientos. Se puede acceder a estas bases de datos a través de los sitios web de las asociaciones y/o sus redes sociales:

- Alemanes del Wolga en Argentina (Wagner Raúl A., 2007).
- Centro Cultural Argentino Wolgadeutsche.
- Federación Argentina de Descendientes de Alemanes del Volga [3].
- Asociación Argentina de Descendientes de Alemanes del Volga Unser Licht.

En base a estos archivos, se seleccionó un conjunto de apellidos a los que se puede asignar la etiqueta de alemanes del Volga. En Argentina, la investigación de la enfermedad de Alzheimer se ha centrado en áreas geográficas y cohortes específicas (Larraya et al., 2004, Melcon et al., 2010, Méndez et al., 2018, Itzcovich et al., 2020, García and Comesaña, 2021). Actualmente, no existe un registro institucional nacional de individuos diagnosticados. En este contexto, los certificados de defunción pueden servir como una valiosa fuente de datos. La DEIS es responsable de la preparación de informes periódicos que incluyen un registro oficial de todas las defunciones, sus causas inmediatas y cualquier afección asociada o preexistente. Aunque puede haber variaciones entre años o regiones, la calidad de esta información está validada y sirve de insumo fundamental.

Como ya se ha mencionado, los apellidos funcionan como proxy del origen geográfico y los datos biológicos, resultando una herramienta valiosa para los estudios poblacionales. El objetivo fue contribuir a la epidemiología de la enfermedad de Alzheimer analizando la distribución espacial de todas las muertes relacionadas con esta dolencia, la distribución espacial de los apellidos de origen alemán del Volga y la asociación entre estos dos fenómenos en el país. En este estudio se utilizó el Padrón Electoral del año 2015, junto con los registros nacionales de defunciones del período 2005 a 2017, ya que comparten la misma organización territorial de datos. Las 24 divisiones administrativas principales (23 provincias, un distrito federal) se agrupan a su vez en cinco regiones. La región Centro o Pampeana es la más densamente poblada y la de mayor renta per cápita. Durante los

[3] http://fadadav.org.ar

siglos XIX y XX fue el destino más común de las migraciones transcontinentales. Esta región reúne a las provincias de Buenos Aires, Córdoba, Entre Ríos, La Pampa, Santa Fe y Ciudad Autónoma de Buenos Aires. El resto de las regiones NOA (Catamarca, Jujuy, La Rioja, Salta, Santiago del Estero y Tucumán); Noreste o NEA (Corrientes, Chaco, Formosa y Misiones); Cuyo (Mendoza, San Juan y San Luis); y Patagonia (Río Negro, Neuquén, Chubut, Santa Cruz y Tierra del Fuego).

3.4.2. Metodología

Se desarrolló un pipeline Ploomber, procedimiento descrito en la sección 3.3, que implememenenta un ciclo de extracción-transformación-carga con el propósito de adecuar los datos para la creación de visualizaciones e informes.

Se cumplieron tres tareas de extracción: Extracción del padrón electoral del 2015, extracción de los casos de Alzheimer por departamento y extracción de los apellidos de origen Volga. El procesamiento principal para la extracción a partir de padrones electorales fue presentado en la sección "Fuentes de datos para la isonimia" del primer capítulo. Básicamente se alcanza un dataset de trabajo que presenta todos los apellidos registrados a nivel departamental, en este caso, según el padrón electoral del año 2015. Estas tareas, potencialmente replicables para cualquier padrón electoral, fueron ordenadas en un proceso de ETL para padrones utilizando Ploomber. Los padrones electorales limpios y ordenados sirven de insumo para todos los proyectos de esta tesis. Para la extracción de los casos de Alzheimer, fueron desagregados a nivel departamental todos los registros de fallecimientos publicados por la DEIS para los años 2005 a 2017. A partir de este nivel se pueden agrupar para completar los niveles administrativos más abarcativos (es decir el provincial, el regional y el nacional). Aunque el certificado de defunción es un documento oficial firmado por un profesional médico, no está exento de imprecisiones cuando no se conocen bien los datos sobre el estado general de salud previo a la muerte. Se consideró entonces un criterio amplio para seleccionar los siguientes códigos según la Clasificación Internacional de Enfermedades (CIE-10):

- G30.0 (enfermedad de Alzheimer de aparición temprana, normalmente antes de los 65 años),

Figura 3.1: Fuentes consultadas en línea con registros de apellidos de origen Volga. De izquierda a derecha: www.alemanesdelwolga.com.ar; hilandorecuerdos.blogspot.com/; fadadav.org.ar/el_alto/

- G30.1 (enfermedad de Alzheimer de aparición tardía, normalmente después de los 65 años),
- G30.8 (otras enfermedades de Alzheimer),
- G30.9 (enfermedad de Alzheimer, sin especificar) y
- G31 (otras enfermedades degenerativas del sistema nervioso, no clasificadas en otro apartado)

Se filtraron los registros de fallecimientos según estuvieran o no clasificados con uno de cinco códigos de la familia de las G, "enfermedades inflamatorias del sistema nervioso central". Se obtuvo así el número total de muertes para el periodo y el número de fallecimientos específicamente relacionados a la enfermedad de Alzheimer.

La lista para clasificar apellidos, según sean de origen Volga o no, fue confeccionada a partir de recursos disponibles en internet. En la Fig. 3.1, se presentan las distintas fuentes de apellidos de interés. Las formas de consulta son potencialmente más automatizables, como en el caso del sitio *www.alemanesdelwolga.com.ar* y otros suponen una extracción manual y detenida.

Entre las tareas de transformación, las principales son las encargadas de obtener la frecuencia de apellidos Volga y las que calculan las tasas de mortalidad del Alzheimer. Los apellidos del padrón fueron clasificados según pertenezcan o no al conjunto de apellidos de origen Volga. La frecuencia de estos apellidos se calculó como el número de personas con apellidos Volga (VS) por cada 1000 votantes (VS*1000) a nivel departamental, provincial y regional. Para los datos de salud, tomando el período 2005-2017 como un todo, se

calcularon las tasas de mortalidad relacionadas con la Enfermedad de Alzheimer (EA) por cada 1000 defunciones (EA*1000). A esta tasa que relaciona fallecimientos totales y específicos se la denomina TMEA o tasa de mortalidad relacionadas con la EA. Se calculó a nivel departamental, provincial y regional.

Para visualizar la información geográficamente, se implementaron tareas de elaboración de mapas coropléticos nacionales tanto para las tasas Volgas como para las TMEA. En lo que respecta a las labores de minería de datos destinadas al análisis de la información recabada en las etapas previas, se ejecutó el proceso de cálculo del índice de Moran global y el indicador local de autocorrelación espacial (LISA) para analizar la distribución espacial de la frecuencia de VS y las Tasas de mortalidad relacionadas con la enfermedad de Alzheimer (TMEA), así como la combinación de ambas variables.

Este proceso consta de un conjunto de funciones (una por cada atributo analizado) que leen la información de los departamentos y la combinan con la capa geográfica para así proceder a obtener las vecindades y calcular la autocorrelación espacial. Estas funciones se implementan utilizando el paquete Exploratory Spatial Data Analysis (ESDA) que es un subpaquete del Python Spatial Analysis Library (PySAL) (Rey and Anselin, 2009). Para cada atributo, el producto de ejecutar la tarea será un valor espacial y una asignación a un cluster LISA, para cada departamento. Esta información permitirá obtener un gráfico de Moran (Moran Scatterplot) y un mapa de hotspots, coldspots y outliers espaciales, respectivamente. La autocorrelación espacial calculada permite determinar si los departamentos y sus atributos (valores de VS y TMEA) presentan patrones agrupados, dispersos o aleatorios en el territorio argentino. El análisis puede ayudar a identificar relaciones y patrones espaciales que pueden no ser evidentes a primera vista. Para los LISA locales se determinó un nivel de confianza de 0,05 utilizando la prueba de Monte Carlo (999 permutaciones) bajo la hipótesis nula de ausencia de asociación. El análisis se repitió aplicando cortes nacionales, regionales y provinciales, para examinar cambios de patrones según varíen los conjuntos con los cuales se establecen las vecindades. El diseño del pipeline completo se observa en la Fig. 3.2, a partir del cual se implementan las tareas Ploomber.

Previamente, se mencionó que el framework Ploomber nos otorga la posibilidad de definir parámetros o argumentos generales en un archivo, a los cuales cualquier tarea del

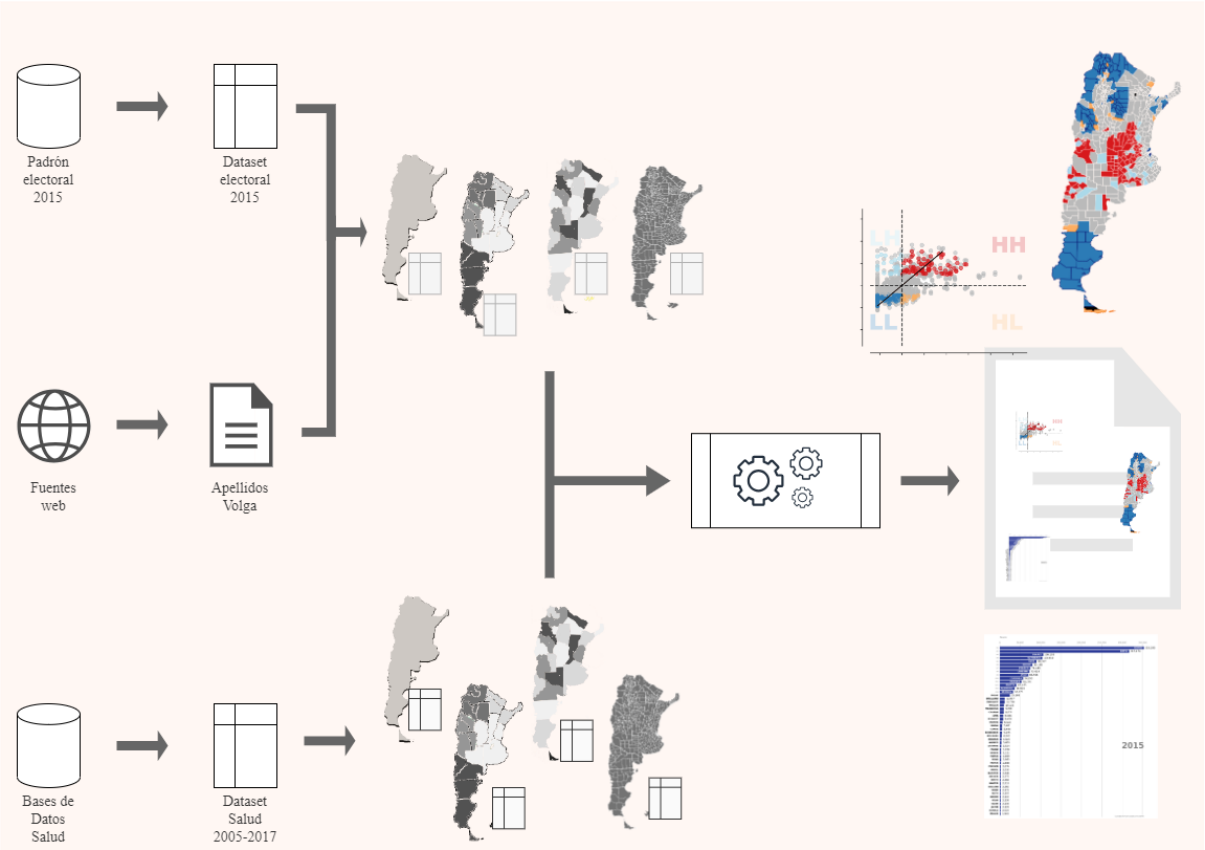


Figura 3.2: Pipeline Volga: Desde los datos crudos hasta la estadísitica espacial bi-variada. Tres fuentes de datos de distinta naturaleza sirven de entrada al procesamiento. Luego de una estructuración inicial de los datos crudos, se generan subconjuntos de datos para cada nivel administrativo. Luego, éstos se disponen para el entrecruzamiento y las trasformaciones necesarias, considerando además sus relaciones en el plano entre niveles. Como resultado se obtienen las visualizaciones, tablas y reportes que ayudan a la comprensión del fenómeno en estudio. Un conjunto de parámetros globales rigen el comportamiento de todo el procedimiento.

proceso puede acceder. Como desarrolladores, podemos aprovechar estos argumentos y caracterizarlos tanto como funcionales como no funcionales. Los primeros se utilizan para modificar el comportamiento de las tareas, como por ejemplo la estructura Python que se utiliza para especificar nuevos rangos de años sobre el período en análisis, y así agrupar los datos, calcular tasas, combinar, resumir, graficar, etc. Por otro lado, los no funcionales nos permiten especificar, por ejemplo, la ruta al archivo de la capa geográfica que se va a utilizar, o la estructura de datos que mapea correspondencias de nombres para los textos que figuran en las visualizaciones y los reportes. El código fuente de este proyecto de datos se encuentra en el repositorio Github, accesible a través del siguiente enlace: github.com/LeoMorales/volga-pipeline.

3.4.3. Resultados y conclusión

Según las estimaciones de población proporcionadas por el INDEC en el año 2023, Argentina tenía una población total de 43.131.966 habitantes en 2015. El padrón electoral de ese año contenía 30.530.194 electores, con 373.709 apellidos diferentes. De ellos, 326.922 individuos (1,22 %) presentaron un apellido Volga, lo que supone un total de 1.109 apellidos con este origen.

La Tabla 3.1 muestra la distribución geográfica de los portadores de apellido Volga (VS). La frecuencia más alta se encontró en la provincia de La Pampa (región Centro), mientras que San Juan (región Cuyo) mostró la más baja. La tabla también resume la información de los datos de certificados de defunción. Entre 2005 y 2017 se registraron 4.115.216 defunciones en Argentina, de las cuales 17.226 (4,19 %) estuvieron relacionadas con la enfermedad de Alzheimer. A nivel provincial, La Pampa tuvo la tasa de mortalidad relacionada con la dolencia (TMEA) más alta, mientras que la provincia de Santa Cruz tuvo la más baja. Del total de fallecidos por Alzheimer, el 68 % fueron mujeres y el 31 % hombres. En el resto de los casos no se pudo determinar el sexo biológico del fallecido. El código más frecuente en la base de datos fue G30.9 (enfermedad de Alzheimer, sin especificar), apareciendo en el 85 % de los certificados. La edad media en los certificados de defunción con el código G30.0 (enfermedad de Alzheimer de inicio precoz) fue de 73,82 años para las mujeres, con una desviación estándar de 14,89 años, y de 68,23 años para los hombres (desvío estándar=12,01).

Región	Provincia	Electores 2015	Electores con apellido Volga	VS*1000	Fallecimientos 2005-2017	Muertes relacionadas al Alzheimer	TMEA
Centro	CABA	2.541.076	20.75	8,17	422.702	1.129	2,67
	Córdoba	2.645.525	13.766	5,20	360.432	2.148	5,96
	Entre Ríos	979.546	46.768	47,74	128.128	736	5,74
	La Pampa	262.03	14.769	56,36	32.465	338	10,41
	Santa Fe	2.552.338	30.694	12,03	378.409	2.795	7,39
	Total	20.364.904	269.337	13,23	2.999.143	12.404	4,14
NOA	Catamarca	278.151	361	1,30	28.986	82	2,83
	Jujuy	478.463	477	1,00	51.76	101	1,95
	La Rioja	250.537	368	1,47	25.779	45	1,75
	Salta	885.984	1.351	1,52	92.08	271	2,94
	S. del Estero	648.777	1.102	1,70	70.362	132	1,88
	Tucumán	1.079.057	1.519	1,41	127.163	515	4,05
	Total	3.620.969	5.178	1,43	396.13	1.146	2,89
Cuyo	Mendoza	1.307.278	2.634	2,01	167.537	1.271	7,59
	San Juan	504.837	487	0,96	60.88	185	3,04
	San Luis	334.603	1.093	3,27	37.218	183	4,92
	Total	2.146.718	4.214	1,96	265.635	1.639	6,17
NEA	Chaco	815.907	8.359	10,25	92.655	268	2,89
	Corrientes	765.271	2.972	3,88	85.889	257	2,99
	Formosa	392.863	2.238	5,70	43.792	268	6,12
	Misiones	787.588	19.489	24,75	82.463	310	3,76
	Total	2.761.629	33.058	11,97	304.799	1.103	3,62
Patagonia	Chubut	388.934	3.208	8,25	38.443	194	5,05
	Neuquén	436.081	3.397	7,79	36.851	332	9,01
	Río Negro	474.634	5.979	12,60	50.768	373	7,38
	Santa Cruz	220.278	1.481	6,72	17.353	19	1,09
	T. del Fuego	116.042	1.07	9,22	6.094	16	2,63
	Total	1.635.969	15.135	9,25	149.509	934	6,25
Argentina	Total	30.530.189	326.922	10,71	4.115.216	17.226	4,19

Tabla 3.1: Producto del *pipeline Volga* que devuelve la frecuencia de apellidos Volga (VS) y tasas de mortalidad relacionadas con la enfermedad de Alzheimer (TMEA) por regiones, provincias y todo el país

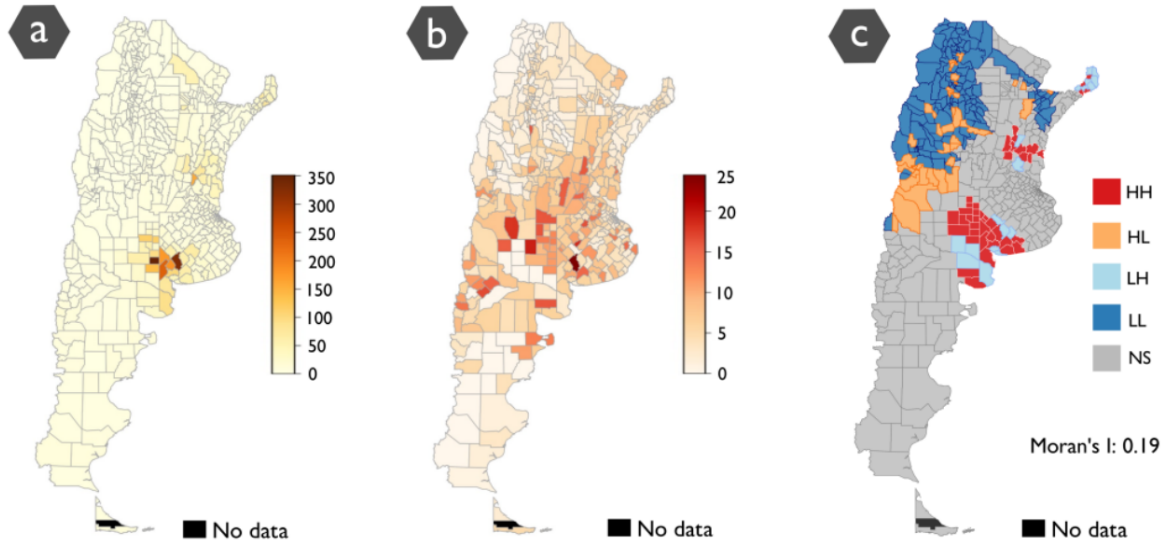


Figura 3.3: Mapa de coropletas de frecuencia por departamentos de apellidos de origen Volga (a), y tasas de mortalidad relacionadas con la enfermedad de Alzheimer (b). Los mapas LISA ilustran la agrupación geográfica de ambas variables (c). Los departamentos con valores altos que están rodeados por vecinos con valores altos están coloreados en rojo (HH). Los valores altos rodeados de valores bajos están coloreados en naranja (HL). Los departamentos con valores bajos rodeados de valores altos están coloreados en azul claro (LH). Los grupos de valores bajos o puntos fríos se colorean en azul (LL). En gris, no significativo (NS). Todos los mapas son producto de la *pipeline Volga*.

Todos los análisis espaciales indican una autocorrelación positiva con valores p inferiores a 0,001. Los mapas de coropletas de la Fig. 3.3 ofrecen una clara visualización de la variabilidad de la VS y la TMEA por departamentos. La mayor concentración de votantes con apellidos de origen Volga se encuentra en varios departamentos de la región Centro-Pampeana (Fig. 3.3a). Se observan valores de TMEA notablemente bajos en los departamentos de las regiones NOA, NEA y sur del país (Fig. 3.3b). Se identificaron tres conglomerados con valores altos y no aleatorios (I de Moran=0,19) tras calcular la I de Moran bivariada para las dos fuentes de datos (VS y TMEA). Estos hot spot se caracterizan por departamentos con altas tasas específicas de mortalidad relacionada con el Alzheimer y una alta frecuencia de portadores del apellido Volga. Se localizan en la región Centro/Pampeana y en el norte de la Patagonia. Por el contrario, los conglomerados de valores bajos o puntos fríos se encuentran en las regiones Noroeste y Noreste (Fig. 3.3c). No se detectaron asociaciones espaciales significativas en el resto del país.

Se determinó, mediante un modelo lineal generalizado, que la frecuencia de apellidos

de origen alemán del Volga explica el 43,53 % de la variación en las muertes relacionadas con la enfermedad de Alzheimer.

La enfermedad de Alzheimer, en todas sus formas, es un reto de política pública en términos de detección precoz y atención de la salud. Aunque la evaluación clínica debe combinarse con la neuroimagen, los biomarcadores y los estudios moleculares para el diagnóstico, las restricciones económicas limitan la capacidad de realizar estudios exhaustivos. Los estudios moleculares de los genes asociados a esta enfermedad rara vez se realizan en Argentina. Las familias descritas por Bird et al˙ (Bird et al., 1988, 1992) compartían una única mutación N141I en el gen PSEN2. Aunque se han identificado más de 10 mutaciones adicionales en PSEN2, la N141I sólo se ha encontrado en familias de origen Volga, lo que sugiere que esta mutación es específica de este grupo de población. En Latinoamérica, la presencia de la N141I fue investigada en un meta-análisis por Llibre-Guerra et al. (Llibre-Guerra et al., 2021). Veinticuatro variantes, típicamente atribuibles a efectos fundadores y en su mayoría de ascendencia europea, fueron detectadas en 3,583 individuos en riesgo, con mayor frecuencia en Colombia seguida de Puerto Rico y México. Un meta-análisis de 47 países mostró que la variante N141I sólo se encuentra en Argentina, Alemania y Estados Unidos (Dehghani et al., 2021).

En Alemania, un estudio realizado a principios del siglo XXI proporcionó valores estimados sobre la prevalencia y la incidencia de la demencia a través de análisis epidemiológicos y metaanálisis a gran escala (Bickel, 2000). La región de donde proceden los alemanes del Volga presentaba una mayor prevalencia de demencia, tendencia que cambió en las décadas siguientes por aumento de la enfermedad de Alzheimer de inicio tardío en pacientes femeninas (Ziegler and Doblhammer, 2009). Estas diferencias se atribuyen principalmente a la mayor esperanza de vida de las mujeres: la edad avanzada sigue siendo el mayor factor de riesgo (Chêne et al., 2015). En nuestro estudio de caso, del total de muertes por Alzheimer, el 68 % fueron mujeres y el 31 % hombres. Se observó una gran disparidad en la distribución de TMEA y VS entre las regiones Centro/Pampeana, Cuyo y Patagonia en comparación con la región NOA. Esta última abarca diversos entornos, incluida la precordillera andina, donde hay poblaciones que residen a altitudes superiores a los 2,500 metros. Estudios genómicos mostraron que aquí se registra la mayor proporción de componente nativo de ascendencia centroandina (Muzzio et al., 2018, Luisi et al., 2020). Por otro

lado, los individuos de la provincia de Misiones (NEA) representan la mayor proporción de ancestría del centro/norte europeo. Esto es consistente con el registro histórico de asentamiento de colonias polacas, alemanas, danesas y suecas en esta provincia.

Los factores geográficos y la historia biológica de una población también pueden influir en el riesgo de desarrollar demencia (alz, 2020). La migración de los alemanes del Volga a Argentina debe distinguirse de la migración de otros grupos alemanes. Según las fuentes germánicas, la migración masiva tuvo lugar entre 1883 y 1890. Sin embargo, hasta 1870, la mayoría de los contingentes que partieron por el puerto de Bremen se dirigieron a Estados Unidos, Canadá y Brasil. Argentina sólo se convirtió en un destino de interés con posterioridad a esa fecha. Según los Censos Nacionales de Argentina, los migrantes alemanes representaban una minoría, que no superaba el 2,35 %, el 1,70 % y el 1,14 % de la población total en los censos de 1869, 1895 y 1914, respectivamente. Casi 150 años después, la población descendiente de la migración del Volga permanece altamente concentrada. Combinando datos diacrónicos (apellidos) y sincrónicos (defunciones) se identificaron patrones de distribución territorial y co-espacialidad. Estos hallazgos sugieren que es posible combinar documentos históricos y oficiales, como padrones electorales, datos censales e información sanitaria, para analizar la dinámica poblacional y proporcionan una base sólida para orientar el muestreo en la investigación médica. La enfermedad de Alzheimer es un problema sanitario importante y creciente. El envejecimiento de la población ha añadido urgencia al desarrollo de políticas públicas eficaces y a la búsqueda de mejores terapias. Aunque las formas familiares representan un pequeño porcentaje de los casos de Alzheimer, es muy importante estudiarlas. Las mutaciones responsables de estas formas tienen consecuencias bioquímicas conocidas que probablemente estén en el origen de la enfermedad de Alzheimer esporádica. Para las personas y familias en riesgo, las intervenciones tempranas tienen el potencial de retrasar o incluso prevenir la demencia en personas asintomáticas, además de ralentizar la progresión en aquellas con síntomas (Bateman et al., 2011). Enfoques analíticos innovadores ofrecen la posibilidad de identificar los determinantes de las disparidades sanitarias en la enfermedad de Alzheimer y su impacto a nivel individual, comunitario y social (Akushevich et al., 2023). Algunos de estos determinantes están relacionados con el origen étnico, el género y la geografía. El análisis de los apellidos es un método adecuado y económicamente viable para distinguir entre

grupos y estructuras dentro de lo que puede parecer un grupo social homogéneo. También es útil para describir escenarios migratorios complejos en otros países latinoamericanos que experimentan procesos demográficos similares y puede ayudar a reducir los errores de muestreo al identificar dónde y cuándo es probable que persista un patrón genético preexistente.

3.5. Caso 2 - Migración Interna en Argentina: Patrones Espaciales y Modificaciones de la Estructura Poblacional

3.5.1. Introducción, fuentes de datos y objetivo

Para entender histórica y prospectivamente la dinámica y distribución territorial de la población, es central analizar las migraciones internas, ya que alteran la estructura por sexo y edad, contribuyendo a rejuvenecer o envejecer la edad promedio y la edad mediana de las provincias de origen o las de destino (Organización Panamericana de la Salud, 2017). Influyen en el ritmo, la inercia y los diferenciales de crecimiento demográfico entre los territorios, dada la selectividad de la migración por edad, nivel educativo y lugar de residencia, afectando el nivel educativo regional, las tasas de desempleo y los indicadores de pobreza (Álvarez et al., 2007). Las migraciones internas en Argentina han redistribuido población del campo a la ciudad y de las ciudades pequeñas, menores a 10. 000 habitantes, a las medianas y grandes, principalmente las capitales de provincias y sus respectivas áreas metropolitanas (Busso, 2007). Ha sido especialmente estudiado el periódo comprendido entre 1990 y 2001, cuando disminuye la proporción de migrantes internos, asociado a la ciclicidad y la recesión de la economía argentina, que afectó fuertemente a los principales centros urbanos de destino (Córdoba, Rosario, La Plata y Área Metropolitana de Buenos Aires) y a la flexibilidad y precariedad laboral que caracterizaron al mercado de trabajo. El resultado fue una menor migración interprovincial y una intraprovincial importante, especialmente en Chaco, Corrientes y Tucumán. La brecha de desigualdad originó situaciones de discriminación territorial negativa (bolsones y reductos de alta concentración de pobreza y desocupación), disminuyendo las oportunidades de desarrollo y dinámica social de la población más joven (bajo nivel de educación, sin empleo y sin chance de migración), instalando una heterogeneidad social-ciudadana que vulneró derechos de igualdad (Gat-

to, 2007). Estos estudios citados analizaron diversas fuentes de información, mayormente derivada de los censos nacionales y estadísticas provinciales, pero no de los padrones electorales y una articulación con las bases de datos nominales sería una contribución significativa multidisciplinaria.

Como se ha mencionado a lo largo de esta tesis, los padrones electorales son fuentes de datos de enorme valor. Estas bases masivas de apellidos comprenden más del 70 % de la población total del país, se actualizan con periodicidad y permiten georeferenciar la información. El objetivo de este estudio de caso fue conocer el alcance, intensidad y balance territorial de la migración dentro de Argentina, comparando los padrón electorales de los años 2001, 2015 y 2021, en conjunción con los datos del total poblacional y la tasa de crecimiento anual medio (ritmo al que la población aumenta o disminuye durante un período dado, debido al efecto combinado de natalidad, mortalidad y migración) que se obtuvieron de los censos 2001, 2010 y de valores estimados para 2020 (INDEC). Las 24 divisiones administrativas principales (23 provincias, un distrito federal) se agrupan aquí también en cinco regiones Centro (Buenos Aires, Córdoba, La Pampa, Santa Fe y Ciudad Autónoma de Buenos Aires), NOA (Catamarca, Jujuy, La Rioja, Salta, Santiago del Estero y Tucumán), Noreste o NEA (Corrientes, Chaco, Entre Ríos, Formosa y Misiones); Cuyo (Mendoza, San Juan y San Luis) y Patagonia (Río Negro, Neuquén, Chubut, Santa Cruz y Tierra del Fuego).

Para los cálculos de los índices μ (Karlin-McGregor) y m (Wright) se necesita conocer la cantidad de la población de la unidad en análisis (ver capítulo 1). Este dato se obtuvo del censo 2001 y de las proyecciones o estimaciones de población anuales para cada departamento, partido y comuna en los años 2010 y 2020.

3.5.2. Metodología

Al igual que en el caso de uso anterior, se implementaron las tuberías de datos presentadas en “Metodologías de extracción de información”. El procesamiento toma como entrada las versiones limpias y ordenadas de los padrones electorales, cuyas características principales son los datos individuales completos y la asociación de cada registro a algún departamento, con códigos apropiadamente estandarizados. A partir de estos datasets, la extracción de los apellidos es directa, y sirven como entrada o input para las funcio-

nes del paquete isonomía presentado en el Capítulo 2 ("La informática y la isonimia"). Así, se obtiene la "instantánea isonímica" para diferentes años en la línea temporal. Esta organización sistemática de los datos iniciales permite establecer secuencias de tareas específicas, que se ocupan de conformar segmentos poblacionales basados en la estructura administrativa del país.

Para las 23 provincias y la Ciudad Autónoma de Buenos Aires, se obtuvieron los datos del total poblacional y la tasa de crecimiento anual medio (ritmo al que la población aumenta o disminuye durante un período dado, debido al efecto combinado de natalidad, mortalidad y migración) a partir de los censos 2001, 2010 y de valores estimados para 2020 (INDEC). Una tarea específica se encargó de la extracción, transformación y carga de los datos de proyecciones de población. Como resultado, se obtienen tablas con identificadores por unidad administrativa, listas para combinarse con la información isonímica.

Con los padrones electorales y las proyecciones de población como insumo de entrada principal, el pipeline diseñado conformó un proyecto de minería de datos con tareas organizadas en categorías. El nivel o categoría mayor de análisis comprende el país como un todo. Las distintas tareas permitieron obtener los valores isonímicos involucrando todos los apellidos del total de electores. Este descriptivo nacional provee el contexto-base para cada año. Luego, la tarea se repite para el nivel regional, analizando además la evolución de la tendencia para el período bajo consideración. Con los resultados de las cinco regiones se elaboraron mapas coropléticos y gráficos de líneas con la comparativa 2001, 2015 y 2021. Por último, una serie de tareas ordenadas calculan la frecuencia y ocurrencia de cada apellido y permiten obtener los gráficos log-log para cada período.

El mismo análisis de estado isonímico y su evolución se repite para el nivel provincial y el departamental, con los mismos productos (lineplots, mapas coropléticos y gráficos log-log) para las 24 provincias y los 530 departamentos. Detrás de cada gráfico existe una representación tabular, por lo que los reportes resultan de gran utilidad cuando queremos describir relaciones intra-niveles para, por ejemplo, comparar los valores de isonimia de un departamento con respecto a los de la provincia que integra. Para el nivel departamental es posible aplicar análisis adicionales, debido al número de entidades que emerge en este nivel de granularidad. Con los 530 departamentos, es posible establecer vecindades y evaluar la autocorrelación espacial de todos los indicadores, para los distintos años. También fue

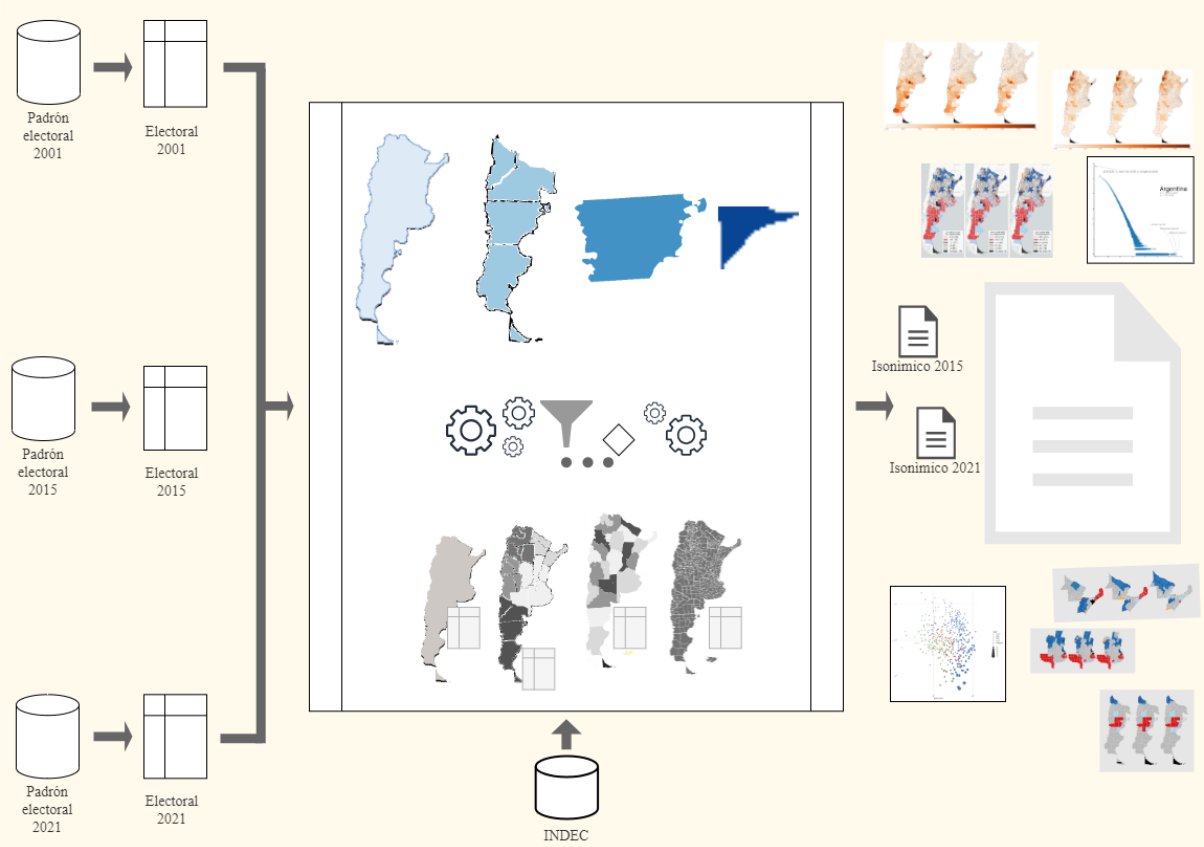


Figura 3.4: Pipeline de datos para el análisis de la migración a través de los apellidos.

posible calcular distancias isonímicas entre cada unidad, conformando cuatro matrices para las cuatro distancias presentadas en el capítulo anterior: Isonímica, Euclídea, Lasker y Nei. La Fig. 3.4 muestra las tareas del procedimiento general. La disponibilidad de nuevos padrones electorales conllevará la necesidad de llevar a cabo una tarea liviana adicional de extracción de datos, encargada de adecuar la información cruda proveniente del nuevo registro. Estas tareas se representan a la izquierda de la caja blanca de mayor tamaño, siendo, para este ejemplo presentado, tres adaptaciones de padrones electorales distintos. Además, se requerirá la adaptación de las actividades de procesamiento (aquellas representadas en el interior la caja blanca) las cuales pueden ser llevadas a cabo con gran facilidad gracias a esta disposición.

3.5.3. Resultados y conclusión

Tendencia

A lo largo de las dos décadas comprendidas en este análisis, la población incluida en los padrones incrementó su número de manera notable:

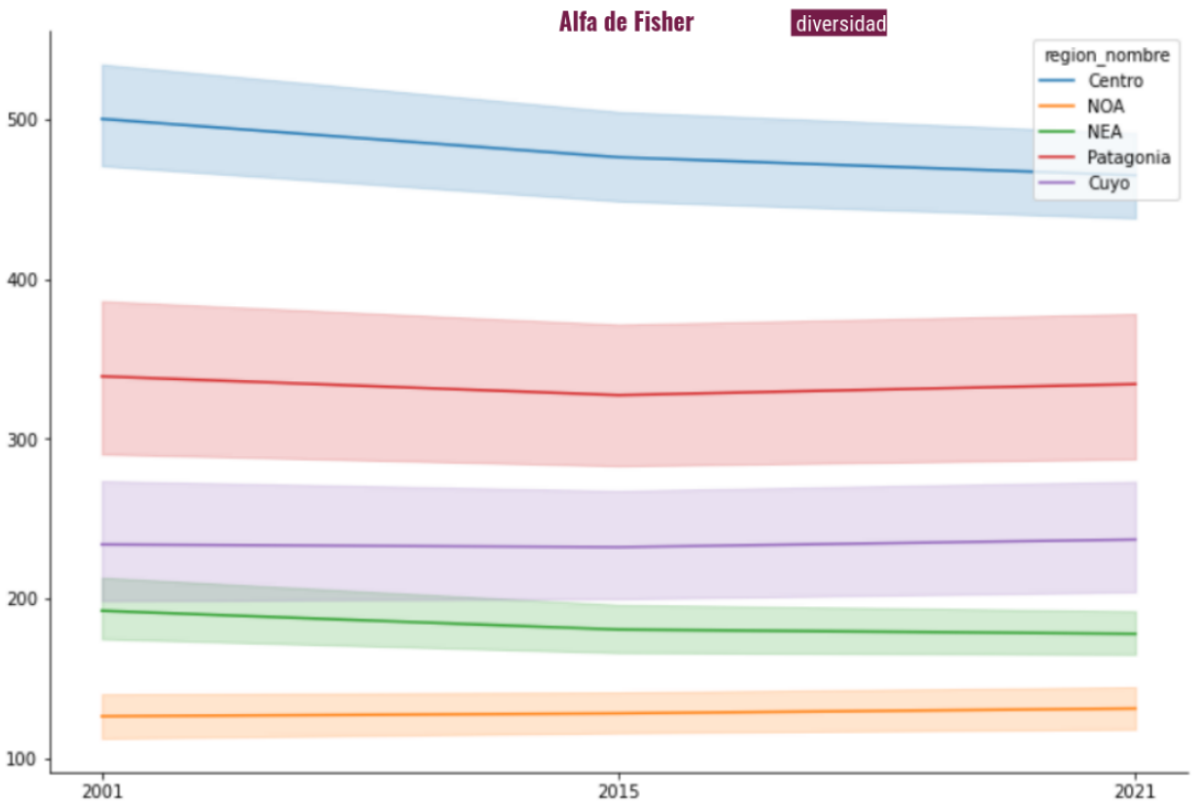


Figura 3.5: Producto del *pipeline Migración* que muestra la tendencia del índice α de Fisher o de diversidad de apellidos.

- 19.693.370 electores en el año 2001
- 30.530.194 en el año 2015
- 34.328.954 electores en 2021.

El mayor salto se dió entre el primer padrón y el segundo. La principal causa asociada es la promulgación e implementación de la Ley Nacional N° 22.864, que desde el año 2012 considera como electores a "(...) los argentinos nativos y por opción, desde los dieciséis años de edad" (PODER EJECUTIVO NACIONAL (P.E.N.), 2012) Sorprendentemente, este incremento en el N total no se vió reflejado en la cantidad de apellidos. Bien por el contrario, estos mostraron un claro descenso:

- Año 2001: 414.441 apellidos diferentes
- Año 2015: 373.709 apellidos
- Año 2021: 335.393 apellidos

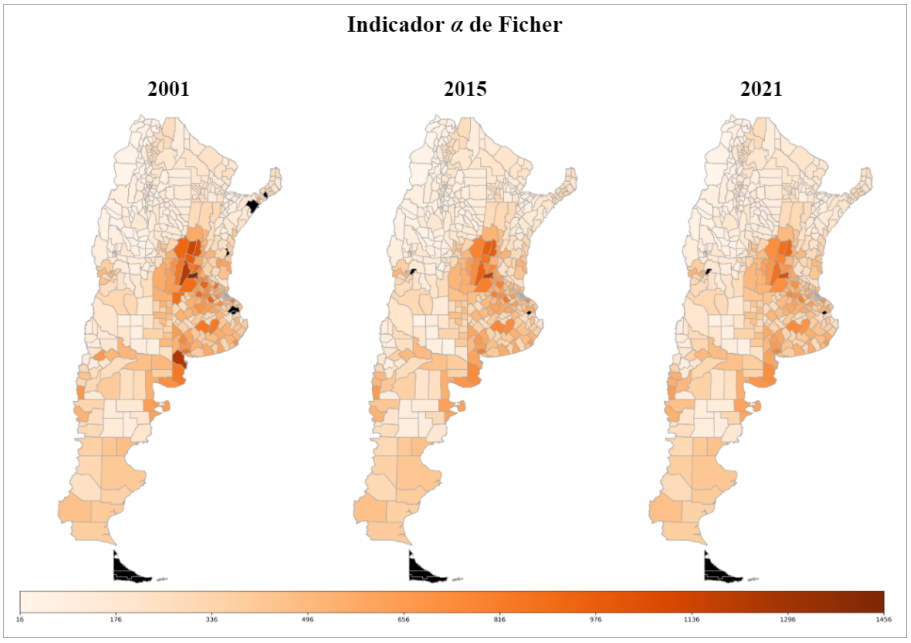


Figura 3.6: Producto del *pipeline Migración* que muestra los mapa de coropletas resumiendo los valores del índice α de Fisher para cada departamento de la Argentina, según los tres padrones electorales analizados.

Tal como se presentó en el capítulo 1, el indicador α de Fisher es un valor que representa la diversidad de la población, en este caso medida en términos de apellidos. En la Fig. 3.5 se muestra la tendencia de este índice a lo largo del período de dos décadas considerado. La línea central coloreada representa el valor promedio de cada región y el ancho es la variabilidad intrínseca entre provincias dentro de la misma región. Aunque mantiene siempre los valores más elevados, en términos absolutos hay una caída en la región Centro, fenómeno que no se registra en las otras regiones. Curiosamente, el Centro es el área del país de mayor densidad poblacional.

En la Fig. 3.6 se presentan los índices α por cada departamento en un mapa de coropletas. El gradiente de colores comprende desde valores altos (mayor diversidad, en naranja oscuro) hasta valores bajos (menor diversidad, en naranja claro). Aunque decrecen, los valores altos siguen concentrados en departamentos de las provincias de Córdoba, Santa Fe y Buenos Aires. La matriz productiva de nuestro país es históricamente un modelo agro productor y exportador. La región Centro se ha comportado como un foco de atracción de población, pasando en forma gradual de los ámbitos rurales a los grandes conglomerados urbanos. De estos movimientos del pasado se mantiene la variabilidad

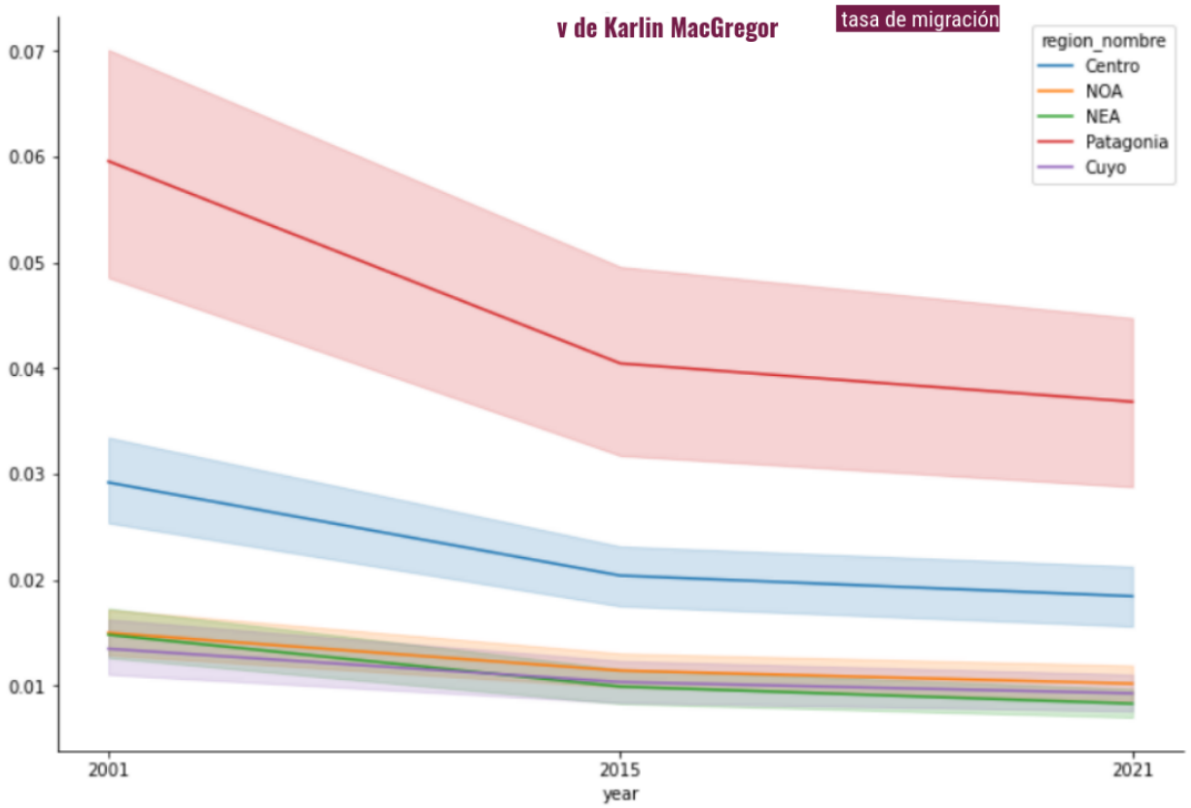


Figura 3.7: Producto del *pipeline Migración* que muestra la tendencia del indicador μ de Karlin-McGregor.

de nombres de familia. Sin embargo, observando las tendencias de los indicadores de migración, no son estas provincias las que mayor variación intercensal sufrieron.

La Fig. 3.7 resume los valores del indicador μ de Karlin-McGregor. Dado que se cuenta con tres padrones electorales abarcando distintos períodos de tiempo (catorce años en el primer intervalo y sólo seis en el siguiente), la escala resultante es muy abrupta. Si bien la tendencia general es decreciente, la migración como fenómeno demográfico es notablemente más alta en las provincias de la región patagónica que en cualquier otra parte del país.

En la Fig. 3.8 un nuevo mapa de coropletas es presentado, esta vez para resumir los valores de m de Wright por departamento. El gradiente de colores comprende naranja oscuro en el extremo derecho (valores altos), concentrados mayormente en departamentos de las provincias de La Pampa, Chubut y Santa Cruz, hasta naranja claro en el extremo izquierdo (valores bajos), concentrados en la región Noroeste.

Observar los resultados al menor nivel administrativo permite una evaluación precisa de los movimientos de las personas a través del territorio. Como se mencionó en párrafos

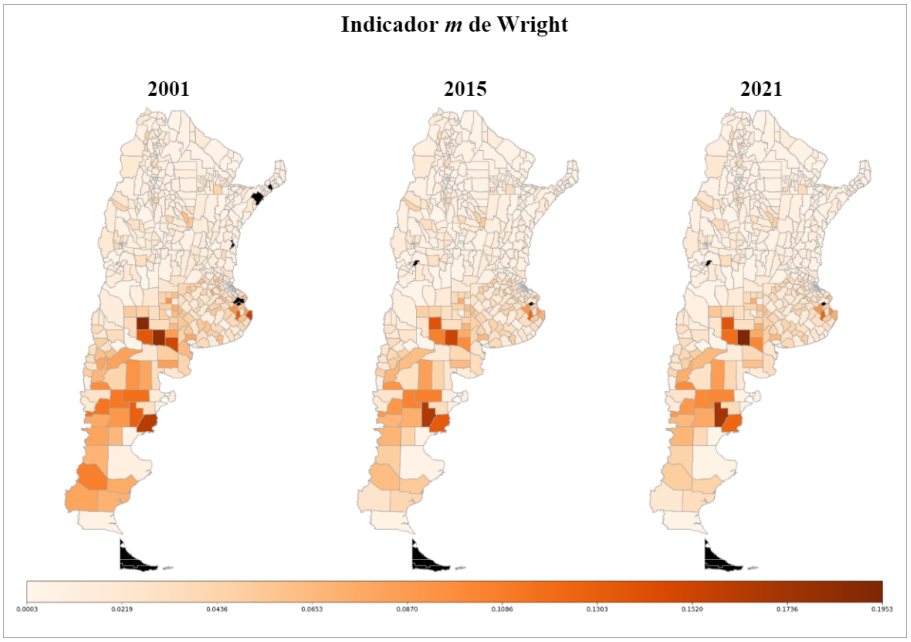


Figura 3.8: Producto del *pipeline Migración* que despliega mapa de coropletas resumiendo los valores del índice *m* de Wright para cada departamento de la Argentina, según los tres padrones electorales analizados.

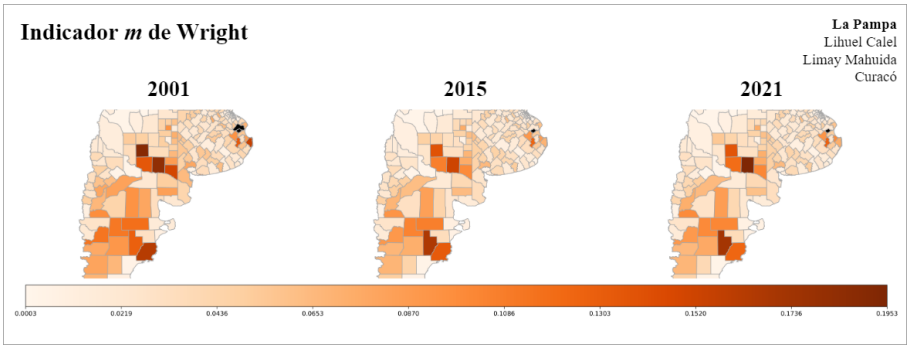


Figura 3.9: Departamentos que mantuvieron valores extremos del índice *m* de Wright en los tres padrones electorales analizados.

precedentes, la migración se da en dos sentidos: los individuos parten desde un lugar de origen y se desplazan hacia un nuevo lugar de destino. Los estadísticos calculados a partir de apellidos señalan la espacialidad de ambos fenómenos. Nos enfocaremos en los cinco departamentos que registraron los mayores valores en todo el país.

La provincia de La Pampa está dividida territorialmente en 22 departamentos. Entre ellos Lihuel Calel, Limay Mahuida y Curacó son contiguos, ubicándose en la región centro-sur. Desde el censo 2001 se destacan, además, por ser los de menor densidad poblacional. El cálculo de los índices de migración v y m a partir de apellidos es muy sensible a la variación en el tamaño de la población. Encontrar en la misma provincia los valores ex-

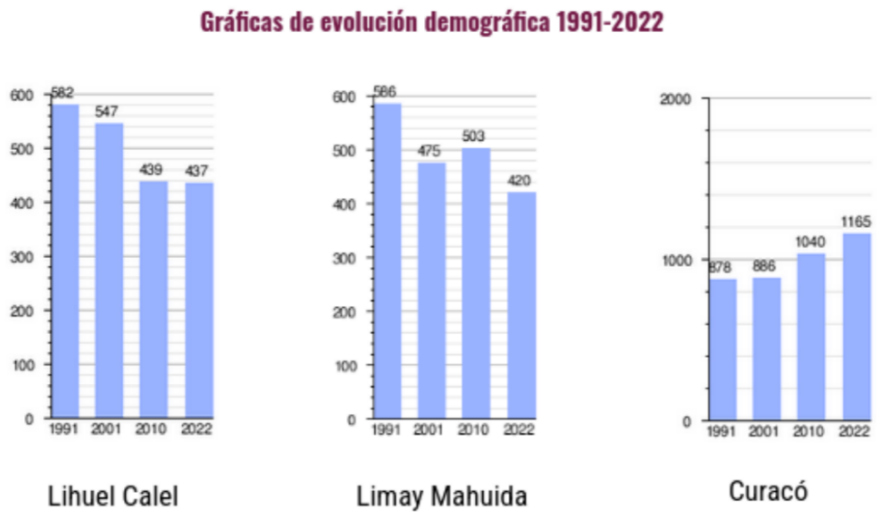


Figura 3.10: Variación en el tamaño poblacional intercensal en los tres departamentos de la provincia de La Pampa.

tremos de ambos índices está señalando una evolución demográfica fuertemente marcada por el desplazamiento de la población (ya sea como lugar de partida o de destino), la que constatamos luego con los datos definitivos intercensales, tal como se presentan en la Fig. 3.9.

La mayor variación intercensal para Lihuel Calel se dió entre el año 2001 y el año 2010, con un -20 % de población registrada como residente en el departamento. Desde entonces, el número se mantiene bajo y decreciente. En Limay Mahuida se registraron dos pulsos ascendentes y dos decrecientes, llegando a una densidad demográfica actual de 0,05 habitantes por kilómetro cuadrado. En Curacó, por el contrario, la tendencia se mantuvo creciente, con una variación de +12 % entre 2010 y 2022.

La provincia de Chubut, por su parte, está dividida en 16 departamentos, siendo Ameghino y Mártires contiguos, ubicándose en el extremo sureste. En ambos, la población está mayormente asentada en parajes rurales, pero el primero duplica al segundo en cantidad de personas y mostrando además una continua tendencia creciente, tal como se resume en la Fig. 3.11. En Mártires, la mayor variación intercensal se registró entre 2001 y 2010, pasando de 977 a 778 habitantes, lo que representa una reducción del 20,4 %. Esta cifra lo convierte en el menos poblado de los departamentos chubutenses y en el tercero menos poblado de toda la Argentina continental, tras los departamentos pampeanos de Limay Mahuida y Lihuel Calel.

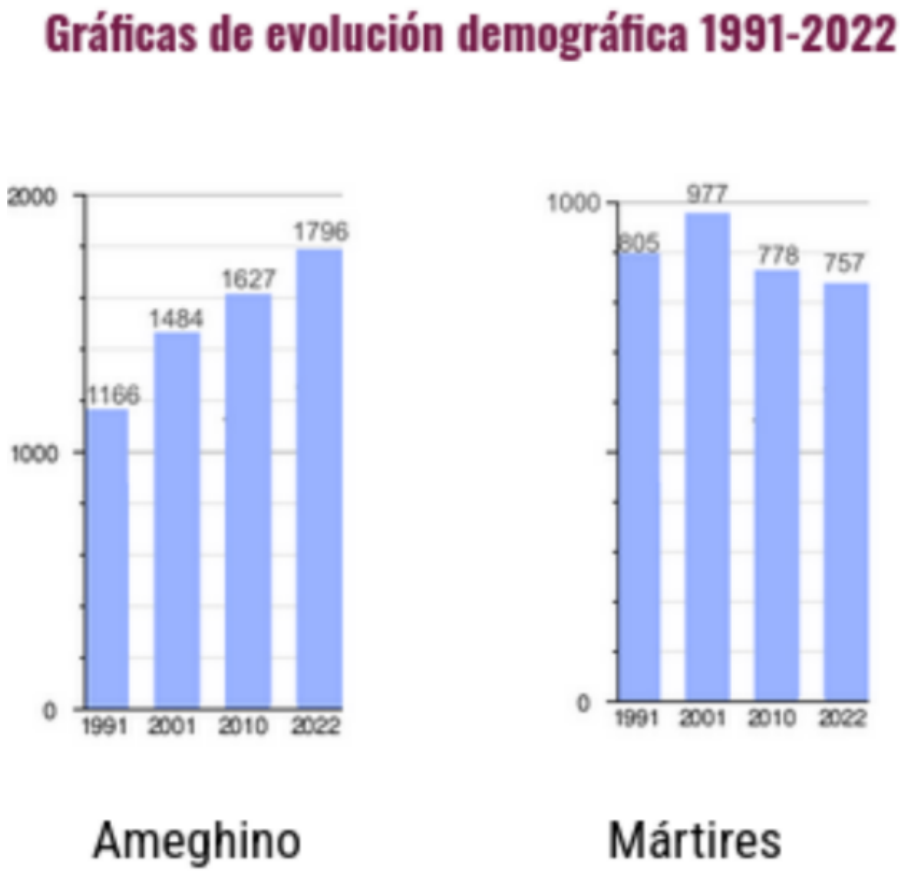


Figura 3.11: Variación en el tamaño poblacional intercensal en los dos departamentos de la provincia de Chubut.

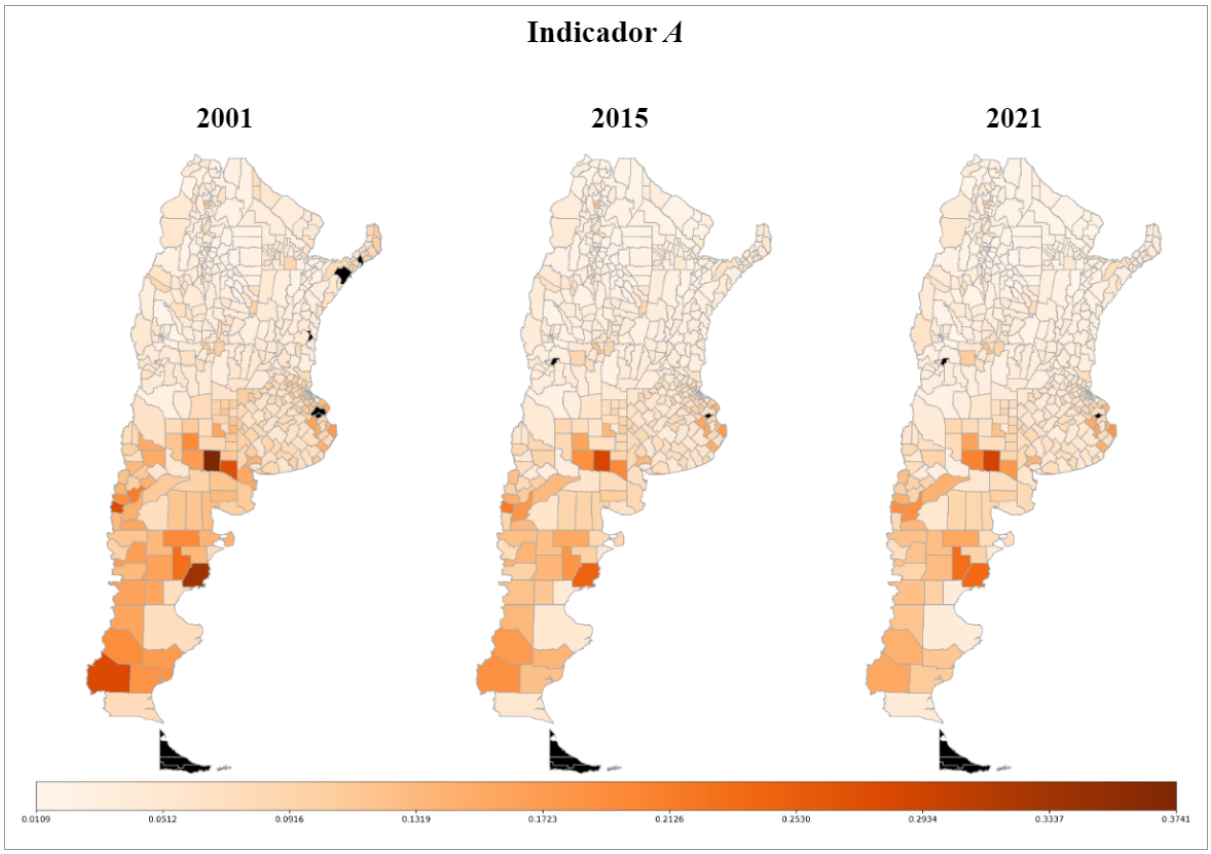


Figura 3.12: Producto del *pipeline Migración* que muestra la tendencia del indicador A para cada departamento de la Argentina según los tres padrones electorales analizados, con mapas de coropletas.

Como se presentó en el Capítulo 1, el indicador A representa el porcentaje de la población que es única portadora de su apellido. En la Fig. 3.12 se presentan los valores de este indicador A nivel departamental en un mapa de coropletas. Los colores fríos (naranja claro) representan las regiones del país donde este fenómeno es poco frecuente, es decir, allí donde cada apellido tiene al menos dos o más portadores. Los colores cálidos (naranja oscuro) representan la situación inversa. Nuevamente vemos en destaque varios departamentos de la región Patagónica y de la provincia de La Pampa. Estos valores altos pueden producirse inmigración. Las personas se establecen en una nueva residencia, aportando su nuevo apellido a la población pero hasta no tener descendencia no pueden transmitirlo y permanecen como únicos representantes de ese nombre de familia. El indicador A también puede ser alto por emigración. Si muchos portadores del mismo apellido se trasladan, la última persona residente en el lugar de origen queda la única representante del apellido. La emigración puede ser la causa principal de los valores altos de m de Wright y del indicador A en Lihuel Calel, Limay Mahuida (La Pampa) y Mártires (Chubut), mientras que la inmigración es la fuerza actuante en Ameghino (Chubut), así como en departamentos de Nuequén, Santa Cruz y la provincia de Buenos Aires.

El indicador B representa el porcentaje de individuos cuyo apellido se encuentra dentro de los 7 más frecuentes en una población determinada y suele tener un patrón espejado a la distribución del indicador A. Como puede apreciarse en la Fig. 3.13, los valores más altos se concentran en departamentos de la región Noroeste, especialmente en las provincias de Jujuy y Salta. Se interpreta como sedentarismo poblacional, al no recibir nuevos apellidos por migración.

Ocurrencias versus frecuencias

El pipeline isonimico y de migración propuesto genera rápidamente los gráficos log-log, tomando distintos conjuntos de personas según la jerarquía emergente. Estas gráficas tienen la ventaja de resumir no sólo la pendiente de la diferencia entre cantidad de portadores de los apellidos más frecuentes y los menos frecuentes, sino también señalar cuáles son dichos apellidos. Retomando la potencia de su clasificación por origen geográfico, étnico o lingüístico más probable, se obtiene una representación muy fiel de las diferentes corrientes migratorias de cada departamento, provincia o región. En la Fig. 3.14 se presentan los

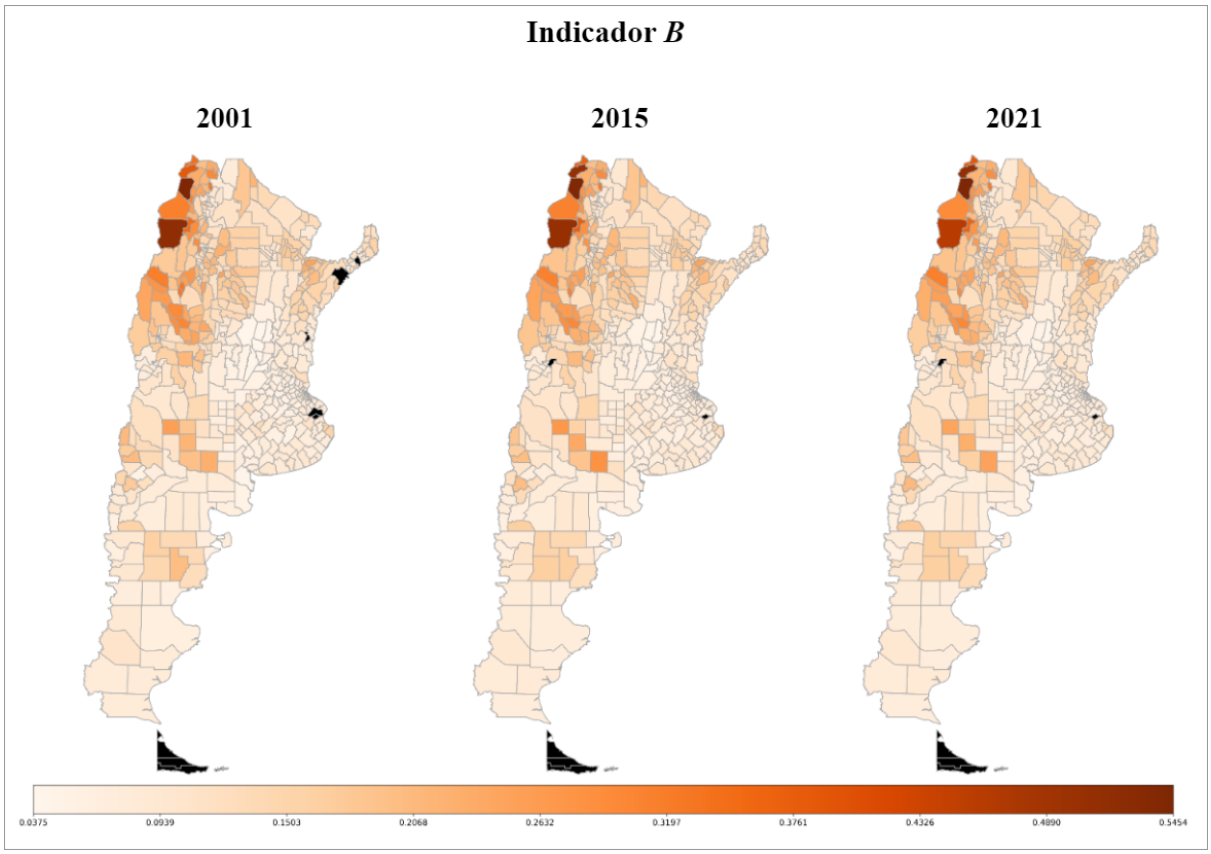


Figura 3.13: Producto del *pipeline Migración* que ilustra, con mapas de coropletas, los valores del para cada departamento de la Argentina, según los tres padrones electorales analizados.

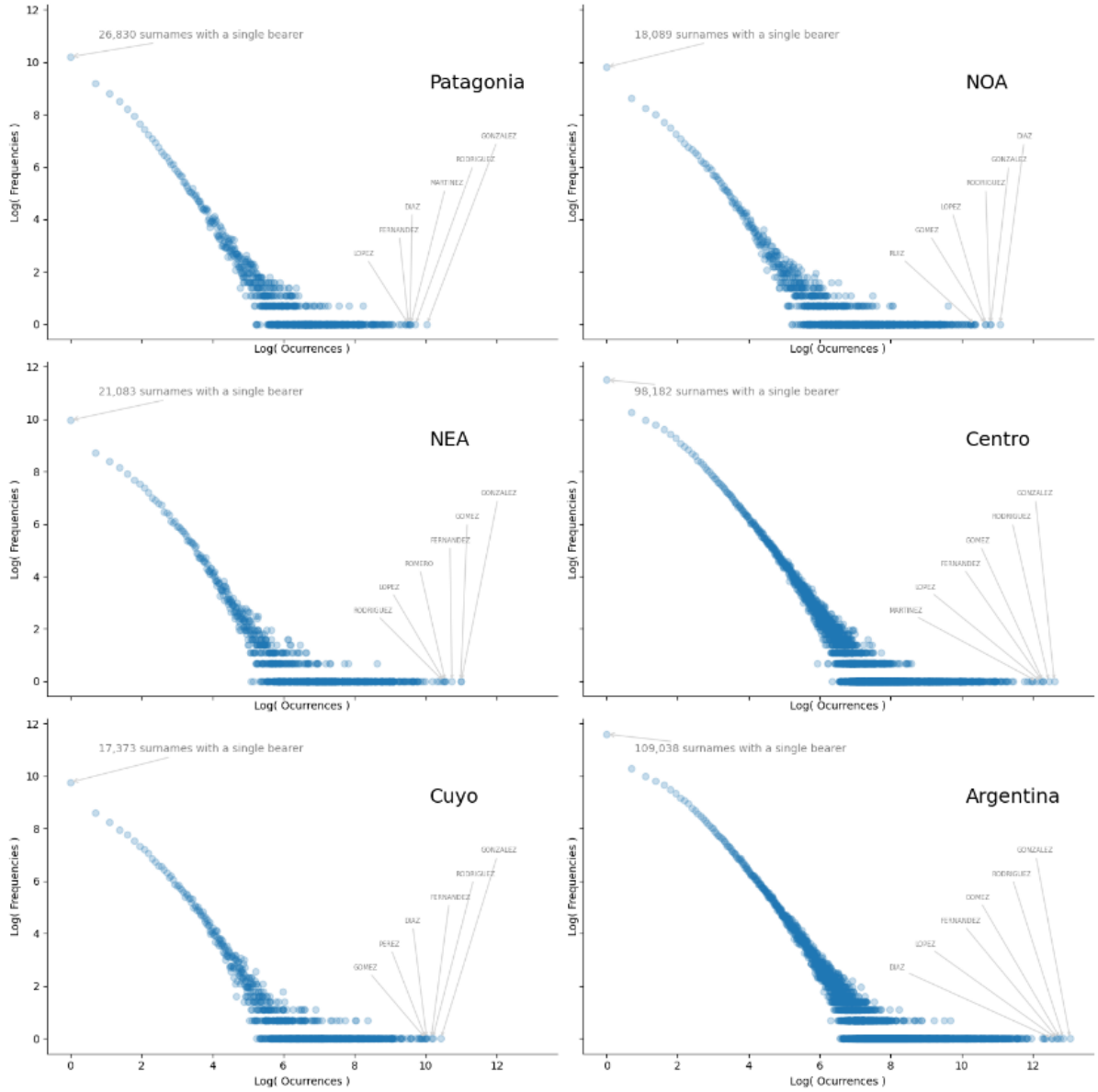


Figura 3.14: Producto del *pipeline Migración* que contiene los gráficos log-log para Argentina y sus cinco regiones.

gráficos log-log a partir del padrón del año 2015 para las cinco regiones y para Argentina como un todo. A este nivel, vemos que la región Centro es la que presenta mayor cantidad de apellidos con un único portador (98,182), este número es considerablemente menor en Cuyo (17,373). En todas las regiones y en el total del país, los seis apellidos más frecuentes son patronímicos de origen español. A nivel nacional, estos son: Gonzalez, Rodriguez, Gomez, Fernandez, Lopez y Diaz.

Este mismo análisis se replicó a nivel provincial y departamental. A diferentes niveles de agregación de los datos del padrón electoral, emerge la variabilidad isonímica producto de

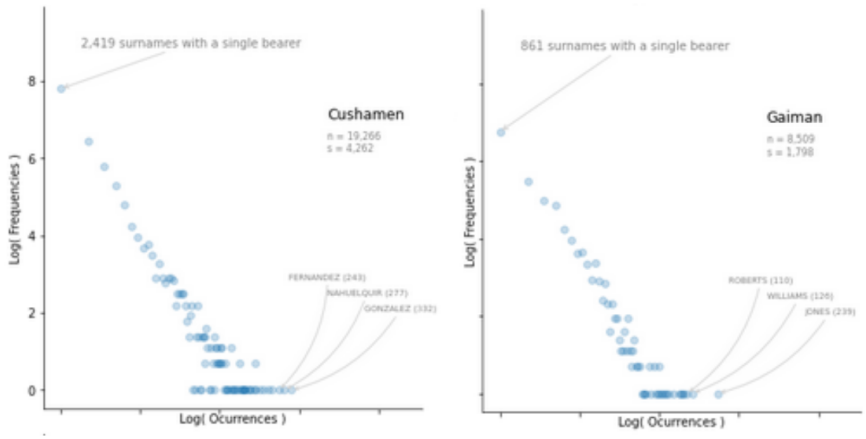


Figura 3.15: Producto del *pipeline Migración* con los gráficos log-log para los departamentos chubutenses de Cushamen y Gaiman.

los diferentes procesos de poblamiento y migración. A modo de ejemplo, en la Fig. 3.15, se presentan las gráficas log-log de dos departamentos de la provincia de Chubut con diferentes historias demográficas: Cushamen y Gaiman.

Al analizar por origen más probable los apellidos con mayor cantidad de portadores, Gonzalez vuelve a ser el más frecuente, pero en Cushamen también encontramos a Nahuelquir, apellido de origen mapuche, portado por 277 electores. En Gaiman, los tres apellidos más frecuentes son de origen galés: Jones, Williams y Roberts. Es el único departamento de todo el país en el que no aparecen patronímicos españoles encabezando la tríada de apellidos con mayores portadores.

El pipeline define tareas repetitivas que se aplican sobre diversas unidades administrativas, independientemente de su nivel (regional, provincial, departamental). Esto genera una amplia gama de resultados, cuya inspección requiere la disposición de tablas resumen. Es crucial confeccionar estas tablas de manera adecuada para que, al examinarlas, se fomente la investigación enfocada en los resultados obtenidos. Un ejemplo ilustrativo de esta situación se evidencia en la generación de gráficos log-log, donde la tarea se realiza en lotes para los 530 departamentos de Argentina (además del para el resto de unidades en los otros niveles). Como agregado, los resultados se amplían con la repetición del procedimiento para los datos de un nuevo año. La exploración de estos resultados sería bastante laboriosa sin disponer de algún recurso de resumen, cuya elaboración también debe implementarse en las tareas del pipeline. El código fuente de este proyecto

de datos se encuentra en el repositorio Github, accesible a través del siguiente enlace: github.com/LeoMorales/surname-migrations-pipeline;

Contar con una tubería de datos facilita extender y replicar toda la batería de análisis hacía diferentes períodos, conformando artefactos de software cómodamente adaptables a nuevos padrones electorales, en la medida en que éstos estén disponibles. Con estas representaciones más exhaustivas, se amplían los horizontes de la exploración, contextualización y comparación isonímica.

En el repositorio del código fuente del proyecto, se localiza la carpeta denominada reporte", la cual contiene el archivo resultante final, actualizado conforme a la ejecución más reciente del flujo de procesamiento. Este reporte final, contiene el extenso conjunto de resultados, tales como gráficos de tendencias, mapas coropléticos y gráficos de frecuencias de apellidos log-log para las divisiones. En este último apartado, se superan los 500 gráficos entre departamentos, provincias, regiones y país.

Capítulo 4

Epidemiología

4.1. Introducción

Saber cuántas personas fallecen cada año y cuáles son las causas de dichos decesos constituye un recurso fundamental para la planificación y elaboración de políticas públicas de salud. De suma importancia es también establecer las enfermedades más prevalentes, sus agentes infecciosos o vectores -si los hubiera-, así como su estacionalidad y los rangos de edades del segmento poblacional más susceptible o en mayor riesgo. Este conocimiento permite abordar la compleja interacción entre genomas y ambiente, que dan como resultado la emergencia y desarrollo de una enfermedad, así como discriminar el rol de las múltiples variables sociales, económicas y culturales que influyen en esa interacción.

Desde los últimos 150 años, los estudios sobre estadísticas mundiales de morbi-mortalidad han demostrado un cambio en los patrones de enfermedad, en las causas de muerte y en la distribución por edades de las defunciones (Foschiatti, 2011). Gradualmente se pasó de un periodo caracterizado por epidemias infecciosas y una alta mortalidad infantil, a uno donde hay una disminución de las llamadas "muertes evitables", en el que las enfermedades crónicas, degenerativas y externas tienen mayor prevalencia. A este cambio se le conoce bajo el nombre de transición epidemiológica.

La transición epidemiológica es el resultado de varios factores:

- Cambios demográficos: la reducción en mortalidad infantil conlleva a una reducción en las tasas de fertilidad.
- Como consecuencia, un mayor porcentaje de la población llega a la edad adulta y desarrollará enfermedades devenidas del proceso de envejecimiento.

- Cambios en los factores de riesgo: esto incluye cambios en la abundancia, distribución y/o virulencia de microorganismos patógenos, factores ambientales —frecuentemente causados por la actividad humana— que pueden causar enfermedades, y factores sociales y culturales, como por ejemplo estilo de vida y tipo de dieta.

- Prácticas de la medicina moderna: como la disminución de afección por ciertos agentes hasta la directa erradicación de enfermedades a través de la vacunación masiva o el descubrimiento de antibióticos y la extensión de los sistemas de salud (Soto-Estrada et al., 2016).

Las principales causas de muerte dejaron de ser las enfermedades infecciosas y parasitarias, dando lugar a un aumento de las Enfermedades No Transmisibles (ENT), como aquellas que afectan al aparato circulatorio, las neoplasias o las neurodegenerativas, como el mal de Alzheimer presentado en el Capítulo 3. El peso de la enfermedad avanzó de los grupos más jóvenes hacia los adultos y sobre todo los ancianos, por lo que padecer una enfermedad dejó de ser un proceso de corta duración, para comenzar a formar parte de toda una vida (Sanmartino, 2016).

En este último capítulo de tesis se presentan otras respuestas desde la ciencia de datos orientadas al estudio de patrones espaciales y la detección temprana de valores anómalos para ciertos fenómenos. La incorporación de la perspectiva espacial y sindémica hace contribuciones importantes a la comprensión de los procesos locales de salud-enfermedad.

El término "sindemia" ha sido introducido hace veinte años aproximadamente por profesionales de la medicina y la antropología, para caracterizar la interacción sinérgica producida entre dos o más enfermedades coexistentes y su exceso de carga de morbilidad resultante (Singer and Clair, 2003). Busca además conceptualizar la asociación que tiene esta ocurrencia simultánea de infecciones no transmisibles, con los factores ambientales sociales, culturales, económicos y físicos (Singer, 2009).

A continuación, se presentan tres casos de análisis con datos sanitarios de nuestro país, aplicando los estadísticos y metodologías ya resumidos. En el primero se busca comprender la tendencia y distribución de muertes por Enfermedades Poco Frecuentes (EPOF), según sexo, edad y grupo de enfermedad (según CIE-10), analizando todos los decesos entre los años 1997 a 2017. En el segundo se analiza un problema de salud pública tanto

de países desarrollados como en desarrollo: las muertes fetales. El objetivo es definir la variación espacial, temporal y las causas asociadas a este fenómeno en Argentina, compilando registros entre los años 1994 a 2019. Finalmente, en el tercero se realizó un estudio transversal para describir la frecuencia de hospitalizaciones de lactantes menores de un año de edad con bronquiolitis en la ciudad de Puerto Madryn, Chubut, durante 2017. También se estudió la distribución espacial de los casos y su relación con indicadores socioeconómicos, a fin de visualizar y comprender mejor los procesos subyacentes a la manifestación local de la enfermedad, mediante la creación de un mapa de vulnerabilidad de la ciudad.

4.2. Fuentes de datos utilizadas

El registro de defunciones es accesible y abierto, a través del portal web del Ministerio de Salud [1]. Los datos sistematizados y puestos a disposición comprenden el periodo 2005-2019. Con el fin de responder interrogantes más complejos, en etapas más avanzadas del trabajo académico, se obtuvo acceso a un nuevo listado de fallecimientos, esta vez abarcando el período desde 1991. Este corpus de datos, brindado para nuestros fines de investigación por la Dirección de Estadísticas e Información de la Salud, se destaca por un grado mayor de detalle. A diferencia de las opciones disponibles en el sitio web (en donde las muertes se agrupan por códigos y años), cada registro se corresponde con un deceso. El archivo está tabulado por cada año, con la codificación de las causas de fallecimientos según el estándar de la CIE en la versión 9 para el primer período (hasta el año 1996 inclusive) y en la versión 10 para el segundo conjunto de datos (año 1997 en adelante). El esquema definido para estructurar la información en cada archivo puede variar de un año a otro.

Se utilizó además una segunda fuente de datos: las proyecciones de población correspondientes al período 1991-2017. Estas proyecciones se derivaron de los informes que detallan las estimaciones poblacionales para el total del país desde 1950 hasta 2015 (según el Censo Nacional de Población, Hogares y Viviendas del año 1991). Esta información fue complementada con el informe análogo correspondiente al período 2010-2040 (según el Censo Nacional de Población, Hogares y Viviendas del año 2010).

[1] https://www.argentina.gob.ar/salud/deis/datos/defunciones

En junio de 2011 fue promulgada en la Argentina la Ley 26.689, que fomenta la prestación de atención integral a individuos afectados por EPOF y a sus familias, estableciendo así el Programa Nacional de Enfermedades Poco Frecuentes. Esta legislación impulsa una política pública orientada a facilitar el acceso a información actualizada y confiable, así como a abordar de manera integral las necesidades de las personas afectadas por este tipo de condiciones, con el objetivo de mejorar su calidad de vida. Mediante la Resolución Ministerial 641/2021 [2] se aprobó el "Listado de Enfermedades Poco Frecuentes". Este documento lista las EPOF, detallando el nombre de la enfermedad y el código ORPHA correspondiente. La red Orphanet reúne información sobre las enfermedades poco frecuentes. En su página web presenta un inventario de medicamentos huérfanos, directorio de asociaciones de pacientes, de profesionales e instituciones y de centros de consulta. También actualiza información sobre los laboratorios clínicos que ofrecen pruebas diagnósticas para enfermedades poco frecuentes y sobre proyectos, ensayos clínicos, registros y biobancos activos. Orphanet fue fundada en Francia en 1997 y tres años después se convirtió en un consorcio que abarca 40 países. Mantiene la nomenclatura Orphanet o códigos ORPHA (una sucesión de números), para mejorar la visibilidad de las EPOF en los sistemas de información sanitarios y de investigación. El listado de 5885 dolencias elaborado en 2023 por el Programa Nacional de Enfermedades Poco Frecuentes (Ministerio de Salud) está organizado siguiendo los códigos ORPHA.

Otros conjuntos de datos principales componen el cuerpo de datos crudos o insumo de entrada, cuyo detalle y fuente de procedencia fueron resumidos en los Capítulos 2 y 3:

- Capas geográficas para representar regiones, provincias y departamentos.
- Codificación de la jerarquía de unidades territoriales, especificada según la división administrativa de Argentina, para contextualizar los valores y tasas calculadas.
- Estadísticas Vitales sobre defunciones provistas por la Dirección de Estadísticas e Información de la Salud (DEIS).
- Sistema de Clasificación Internacional de Enfermedades (CIE) en sus versiones 9 y 10, también disponibles en el material provisto por la DEIS.

[2] https://www.boletinoficial.gob.ar/detalleAviso/primera/240777/20210212

- Indicadores demográficos provistos por el Instituto Nacional de Estadísticas y Censos (INDEC).
- La base de datos Mendelian Inheritance in Man (MIM) y su versión electrónica Online Mendelian Inheritance in Man, (OMIM), registro de conocimiento sobre genes humanos y sus enfermedades asociadas.

4.3. Caso 1 - Epidemiología de las Muertes por EPOF en Argentina

4.3.1. Introducción, fuentes de datos y objetivo

Según la Organización Mundial de la Salud, existen más de 6000 condiciones clínicas de baja prevalencia o EPOF. Una estimación del año 2018 calcula que el 8 % de la población mundial padece directamente alguna EPOF. En Argentina, esto representa 3 millones y medio de personas. Debido a su baja prevalencia, la especialización médica asociada, la disponibilidad de conocimiento y la oferta de atención resultan escasas o inadecuadas. Suele designarse al colectivo de pacientes como huérfanos de los sistemas de salud, ya que a menudo se les niega o dificulta el diagnóstico y tratamiento.

En marzo de 2023, el Ministerio de Salud (autoridad de aplicación de la Ley 26.689), a través del Programa Nacional de Enfermedades Poco Frecuentes, publicó la más reciente revisión del registro nacional de enfermedades. Las fuentes de información consultadas fueron testimonios de referentes provinciales, usuarias/os de dicho registro, asociaciones de pacientes y bibliografía específica. El listado incluye 5885 dolencias caracterizadas como poco frecuentes, de las cuales el 80 % posee un origen genético identificado, implicando cambios en uno o varios genes. Muchas de estas mutaciones pueden transmitirse de una generación a otra, lo que explica por qué ciertas EPOF tienen carácter hereditario. Se denominan enfermedades genéticas poco frecuentes. Sin embargo, la genética es sólo una pieza del rompecabezas. Los factores ambientales (dieta, tabaquismo, exposición a sustancias químicas), también pueden influir. Dichos factores pueden causar directamente una enfermedad o interactuar con factores genéticos para causar o aumentar la gravedad de la enfermedad. Otras EPOF son causadas por infecciones (bacterianas o víricas), alergias, o se deben a causas degenerativas, proliferativas o teratógenas (productos químicos,

radiación, etc). Para otras aún se desconoce la etiología. El 75 % de los casos se presenta en edad pediátrica.

Directamente relacionado al desafío de diagnosticarlas se encuentra la labor de medir su morbilidad y mortalidad. Un medio para este fin es el constituido por la Clasificación Internacional de Enfermedades, en sus versiones CIE-10 y CIE-11. Este nomenclador estandariza códigos (una combinación de letras y números) para fines estadísticos, permitiendo la comparación internacional. Presenta una división en capítulos, desde I a XXII, según el conjunto de enfermedades que agrupe y es la clasificación utilizada en los certificados de defunción. Dado que el 80 % de las enfermedades poco frecuentes posee un origen genético, otra herramienta de gran utilidad es la base de datos OMIM. Centrada en la relación feno-genotipo, es un compendio completo de variantes génicas y fenotipos asociados, disponible en forma gratuita y que se actualiza diariamente. Esta base de datos fue iniciada en 1960 como un catálogo de rasgos y trastornos de herencia de tipo mendeliana y hoy contiene información sobre todos los trastornos conocidos, incluyendo más de 16.000 genes. Finalmente, y de carácter más específico, la red Orphanet reúne información sobre las enfermedades poco frecuentes y fue el nomenclador utilizado por el Ministerio de Salud para el listado de 5885 dolencias.

Nuestro objetivo fue comprender la tendencia y distribución de las muertes por EPOF en Argentina, según sexo, edad y grupo de enfermedad. Las bases de datos, códigos y nomencladores mencionados buscan mantener una equivalencia, aunque existen discrepancias y redundancias que pueden inducir a errores. Se estandarizaron los datos de tal forma que todas las fuentes puedan interactuar entre sí.

4.3.2. Metodología

Nuevamente, al igual que hemos presentado en otros casos de uso de esta tesis, para la puesta en marcha del proyecto de datos sobre epidemiología de las EPOF, se procedió a la creación de un pipeline de datos.

En esta instancia, los parámetros generales del pipeline desempeñan un papel fundamental al posibilitar la variación de los códigos de enfermedades sujetas a estudio. Empleando un mismo procedimiento general, es posible especificar conjuntos de códigos y generar diferentes salidas como resultado. La parametrización de códigos entre los re-

gistros CIE, OMIM y ORPHA hizo posible discriminar las causas de fallecimiento en los registros de defunción y conocer cuáles pueden asociarse a una enfermedad poco frecuente, convirtiendo el pipeline específico EPOF en un pipeline general aplicable a cualquier enfermedad.

Describiendo el procedimiento, nos encontramos con tareas estructuradas conforme a la organización administrativa del país, al igual que en pipelines de casos de estudio anteriores. Antes de realizar las bifurcaciones basadas en el nivel administrativo, se aplicaron una serie de operaciones sobre los conjuntos de datos crudos, para conformar una base uniforme, común a todas las tareas en las cuatro capas administrativas: nacional, regional, provincial y departamental.

En primer término, con respecto a la demografía, se tomaron las proyecciones de población según los dos informes del INDEC mencionados arriba y se unificó en un registro completo para todo el período. Para aquellos años en donde la información se solapa, específicamente entre 2010 y 2017, se optó por mantener la estimación provista por el informe más reciente.

Segundo, para unificar los registros de fallecimientos debimos:

- Extraer los datos para los períodos 1991 al 2000, 2001 al 2014 y 2015 al 2017, los tres períodos con esquemas diferentes (definición de encabezados de las tablas).

- Limpiar las columnas para converger a un dataset con las columnas: *"provincia_id"*, *"department_id"*, *"codigo_defuncion"*, *"sex"*, *"year"*, *"age_in_years"*, *"age_group"*. Para todas las columnas debió utilizarse alguna operación de limpieza. Por ejemplo, se aplicaron conversiones de tipos para todas las columnas, se rellenaron con ceros los códigos de departamento y provincia, o en el caso de la edad se interpretaron dos columnas ("unidad_edad" y "valor_edad") para obtener la cantidad de años al momento del fallecimiento. Esta unificación se ilustra en la Fig. 4.1.

- Los códigos de unidad administrativa se obtuvieron utilizando el paquete isonímia y las bases estandarizadas pertenecientes a los proyectos mencionados en los capítulos anteriores.

En tercer término, se extrajeron 5885 dolencias del "Listado de Enfermedades Poco Frecuentes aprobado por el Ministerio de Salud", mediante el software Tábula. La curaduría

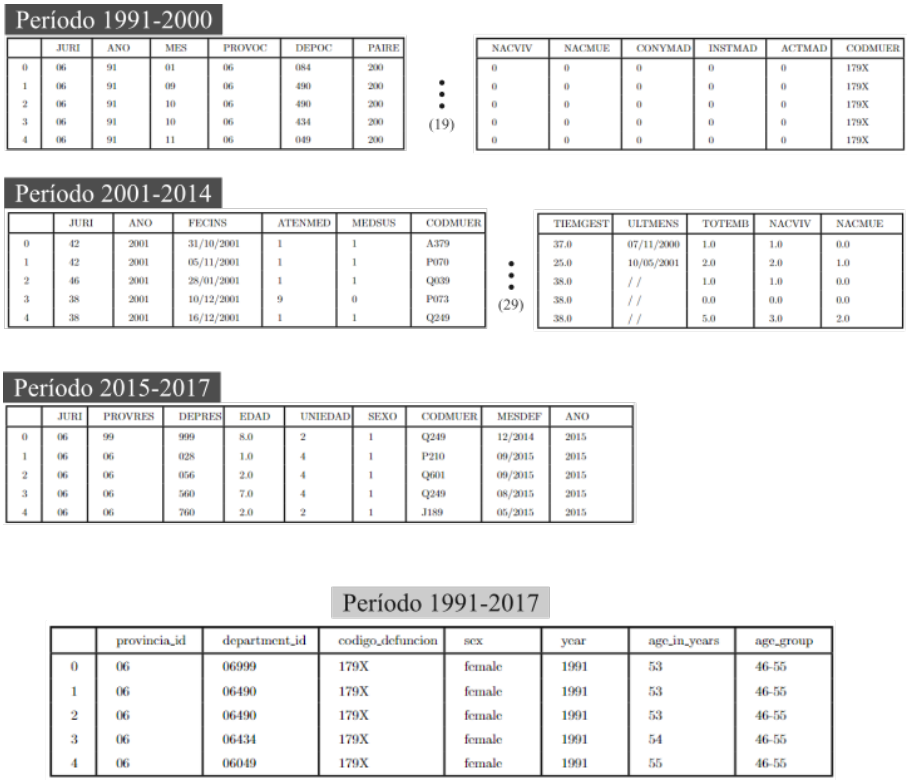


Figura 4.1: Extracción de los datos de fallecimientos y carga de un dataset unificado. En los dos primeros períodos, se muestran las cantidades de columnas omitidas entre paréntesis, 19 en el primer período y 29 en el segundo. El dataset final consta de 7 columnas.

se realizó con las herramientas de ciencia de datos de Python, para convertir cada código ORPHA a su correspondiente CIE-10. Se obtuvieron así 882 códigos CIE-10 que permiten interactuar con las bases de fallecimiento colectadas. Se tomó este conjunto de códigos únicos y se establecieron como parámetros generales del pipeline. Si bien la base de datos de defunciones abarca información desde el año 1991, se ha optado por trabajar en el período comprendido entre 1997 y 2017, dado que es en este lapso de tiempo donde se implementa la codificación CIE-10 en Argentina. Este intervalo ha sido seleccionado como el período de referencia para los análisis posteriores.

Con los datos base ya limpios y dispuestos para el entrecruzamiento, se procede a una segunda etapa para obtener los indicadores (tasas) que servirán de material para generar las visualizaciones de los reportes finales. Se recopilaron las cifras anuales de defunciones atribuibles a causas específicas, utilizando los códigos de enfermedades de interés definidos (establecidos en los parámetros generales del pipeline). Estos códigos se utilizaron como filtro para obtener un subconjunto a partir del conjunto general de fallecimientos y se

agregaron por unidad administrativa, grupo etario, sexo, año (o período) y codificación CIE-10.

Procediendo en el pipeline, utilizando estos dos conjuntos de datos base (el conjunto completo de defunciones y el conjunto de defunciones específicas), se llevaron a cabo las tareas de cálculo necesarias para determinar la mortalidad por causa específica por cada mil personas (fallecimientos EPOF / fallecimientos totales * 1000). Estos cálculos se realizaron sobre los registros de unidades en cualquiera de los cuatro niveles administrativos. En este punto, es muy importante modularizar las tareas de cálculo, ya que posibilita trabajar con maneras muy diferentes de agregar los datos.

En lo referido al corte temporal, se calcularon las tasas de mortalidad por causas específicas, en inglés Cause Specific Mortality Rates (CSMR), para todo el período (años 1997 al 2017), y para intervalos anuales y quinquenales. Al estar el pipeline parametrizado, el usuario cuenta con la posibilidad de indicar, a través de una estructura de datos Python, una conformación específica de intervalos de años que le interese y sobre esos datos se aplica el cálculo de la tasa (por ejemplo: bianual, trianual, etc. o también en intervalos irregulares, con períodos de distinta cantidad de años).

Con respecto a las edades, a cada registro se le asigna una marca de grupo etario según una estructura definida en el archivo de parámetros globales del pipeline, que mapea el rango de edades con su etiqueta correspondiente. Esto garantiza la flexibilidad de conformar grupos etarios según el problema que se esté analizando. Por ejemplo, para las EPOF nos interesa especificar cortes etarios precisos (durante los primeros años de vida y hasta la adolescencia), mientras que en el caso del Alzheimer esto no ocurre. Se puede así definir menor cantidad de grupos etarios, poniendo el foco en las edades más avanzadas.

En relación al sexo, la depuración de los datos condujo a registros etiquetados con una de las tres etiquetas: “male”, “female”, “not-defined”. En la elaboración de los informes, se realizaron filtrados de los registros que se corresponden con las dos primeras categorías, mientras que aquellos etiquetados como “no definido” fueron excluidos.

Como ha sido mencionado, la extensa cantidad de enfermedades que están descritas por los códigos CIE se encuentra organizada en capítulos. Por ejemplo, el capítulo IV incluye todas las enfermedades endocrinas, nutricionales y metabólicas, mientras que

el capítulo V resume los trastornos mentales y del comportamiento. Calcular las tasas diferenciales para cada grupo es de gran importancia en estos estudios epidemiológicos.

Estos cuatro atributos de los datos generan cuatro posibilidades de cálculo de las CSMR, que pueden además combinarse para obtener análisis más específicos. Así, se definieron tareas más completas en todos los niveles. Por ejemplo, se calcularon CSMR anuales por grupos etarios para las cinco regiones argentinas, o CSMR por quinquenio según grupo etario y sexo, para todos los departamentos del país, o la comparativa de CSMR anual por código CIE-10 por sexo para todo el país.

El código fuente de este proyecto de datos se encuentra en el repositorio Github, accesible a través del siguiente enlace: github.com/LeoMorales/epidemiology-pipeline.

4.3.3. Resultados y conclusión

En el periodo considerado se produjeron 826.232 muertes por EPOF (13 % del total de óbitos). La proporción mujer/varón fue de 0,75. Se crearon tareas de elaboración de resúmenes para la información calculada. La Tabla 4.1 muestra los números para todo el período general, sin agrupamiento alguno: fallecidos totales, fallecidos por causas específicas (EPOF) y las tasas por cada 1000.

Otra tarea resumen para el periodo entero contabiliza los fallecimientos según cada grupo etario y según sexo. Los números resultantes se muestran en la Tabla 4.2.

La generación de visualizaciones a partir de los datos computados implica el procesamiento de las diversas tablas de datos producidas en etapas anteriores. El objetivo es representarlas en alguna visualización, como la Fig. 4.2. Allí se resume la cantidad absoluta de fallecidos en todo el país, incluyendo todas las edades y causas, pero diferenciándolos por sexo. En la misma figura se muestran los óbitos por causas específicas relacionados a las EPOF. La ventaja de este tipo de gráficos es identificar rápidamente la evolución de la mortalidad, así como distinguir años específicos que hayan representado un cambio notable en la tendencia. El número absoluto de decesos aumenta como consecuencia del aumento de la población, registrando dos ciclos de incremento y decrecimiento para ambos sexos por igual: uno en el bienio 2007-2008 y otro en el bienio 2016-2017. Además, en el análisis de la mortalidad relacionada a las EPOF, se suma también un incremento notable en el año 1999, que afectó exclusivamente a los óbitos de individuos masculinos.

	REGIONES / PROVINCIAS	Numero fallecidos totales	Numero fallecidos EPOF	% / Total fallecidos	Tasa * 1000 individuos
0	**Centro**	4501839	591952	13.15	131.49
1	Buenos Aires	2604061	327009	12.56	125.57
2	CABA (*)	700390	93191	13.30	133.05
3	Córdoba	557125	77231	13.86	138.62
4	La Pampa	50255	8201	16.32	163.19
5	Santa Fe	590008	86320	14.63	146.30
6	**Cuyo**	406450	54863	13.50	134.98
7	Mendoza	256265	34388	13.42	134.19
8	San Juan	94211	11869	12.60	125.98
9	San Luis	55974	8606	15.37	153.75
10	**NEA**	665306	82813	12.45	124.47
11	Chaco	142222	17074	12.00	120.05
12	Corrientes	133325	16629	12.47	124.72
13	Entre Ríos	200486	27930	13.93	139.31
14	Formosa	65908	8257	12.53	125.28
15	Misiones	123365	12923	10.47	104.75
16	**NOA**	599310	61665	10.29	102.89
17	Catamarca	44000	4459	10.13	101.34
18	Jujuy	78717	8044	10.29	102.19
19	La Rioja	38742	4188	10.81	108.10
20	Salta	137929	13562	9.83	98.32
21	Stgo. del Estero	105720	10781	10.20	101.97
22	Tucumán	194202	20631	10.62	106.23
23	**Patagonia**	219404	34939	15.92	159.24
24	Chubut	57331	8480	14.79	147.91
25	Neuquén	53300	8739	16.39	163.96
26	Río Negro	75066	12080	16.09	160.92
27	Santa Cruz	25286	4090	16.17	161.75
28	TFAIAS (**)	8421	1550	18.40	184.06
29	**Argentina**	6392309	826232	12.92	129.25

Tabla 4.1: Resumen producto del *pipeline EPOF*. Muestra, por cada región, provincia y para el total del país, los óbitos registrados en el periodo 1997-2017, discriminando las muertes asociadas a Enfermedades Poco Frecuentes (EPOF), porcentajes y tasas de mortalidad por causas específicas o CSMR por sus siglas en inglés (Cause Specific Mortality Rates). (*) CABA: Ciudad Autónoma de Buenos Aires. (**) TFAIAS: Tierra del Fuego, Antártida e Islas del Atlántico Sur

	EDADES	Fallecimientos mujeres	Fallecimientos varones	Fallecimientos EPOF mujeres	Fallecimientos EPOF varones	Tasa por cada 1000 mujeres	Tasa por cada 1000 varones
0	≤ 5	103 417	132 086	18 835	23 502	182.12	177.92
1	6-15	18 204	26 322	2 017	2 307	110.79	87.645
2	16-35	80 003	193 706	6 351	10 352	79.384	53.441
3	36-45	79 836	132 476	10 522	15 815	131.79	119.38
4	46-55	160 341	278 965	28 256	49 875	176.22	178.78
5	56-65	295 480	533 133	52 613	106 233	178.05	199.26
6	66-75	522 689	798 790	78 281	138 374	149.76	173.22
7	76-85	893 283	845 465	91 296	105 120	102.20	124.33
8	85+	859 797	429 368	52 643	32 773	61.22	76.32

Tabla 4.2: Resumen discriminado por sexo biológico y grupos etarios de los óbitos totales registrados en el periodo 1997-2017, de las muertes asociadas a Enfermedades Poco Frecuentes (EPOF) y las tasas de mortalidad por causas específicas o CSMR por sus siglas en inglés (Cause Specific Mortality Rates)

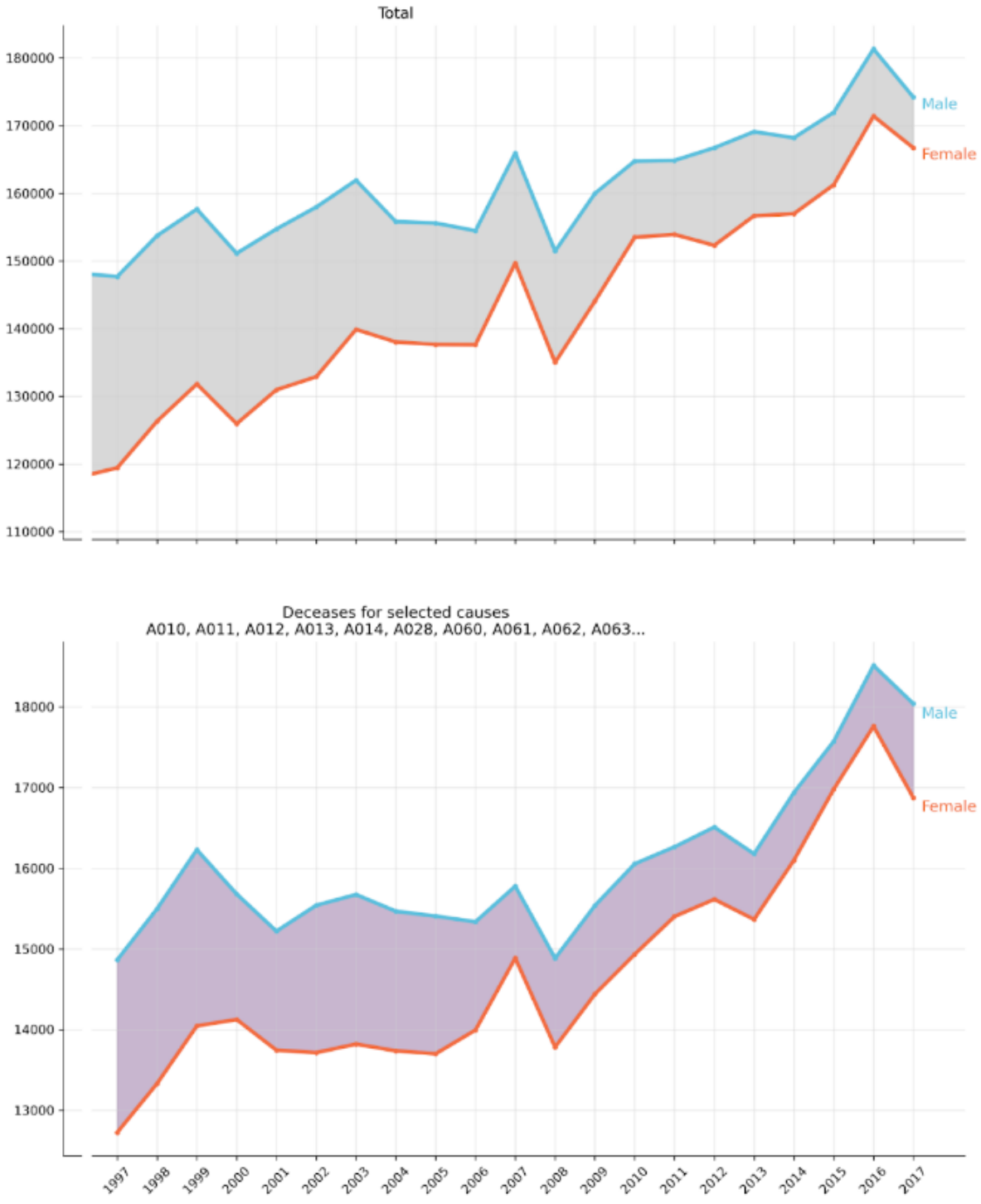


Figura 4.2: Producto del *pipeline EPOF* que ilustra los fallecimientos en cifras absolutas para todo el periodo. a) Total de decesos por todas las causas b) Total de decesos por causas específicas relacionadas a las enfermedades poco frecuentes. En ambas gráficas, la línea azul señala los óbitos masculinos y la línea naranja los femeninos.

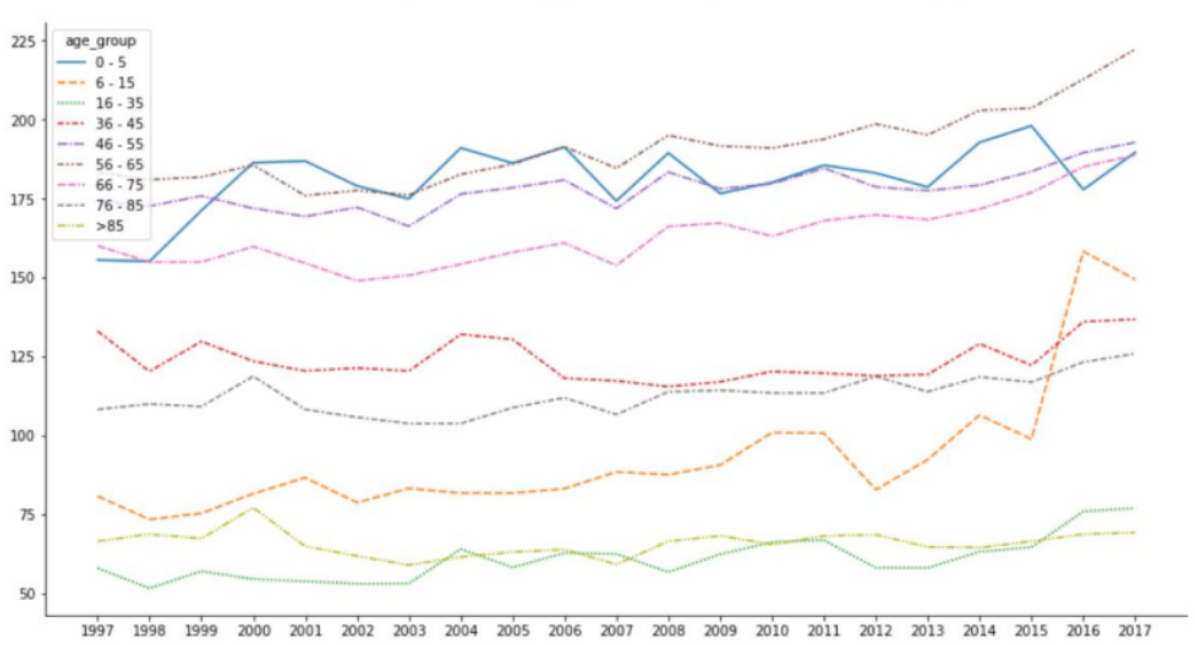


Figura 4.3: Producto del *pipeline EPOF* que ilustra la tendencia de las tasas de mortalidad por causas específicas (CSMR) en Argentina y por grupo etario.

Como se mencionó en la introducción de este caso, el 80 % de las Enfermedades Poco Frecuentes posee un origen genético identificado y el 75 % de los casos se presenta en edad pediátrica. En virtud de este conocimiento, resulta de gran interés analizar la tendencia de las CSMR para cada grupo etario, divididos en rangos de cinco años, desde el nacimiento hasta la edad más avanzada registrada en un certificado de defunción, tal como se resume en la Fig. 4.3. Las tasas más altas (191,7/1000 y 179,7/1000) se presentaron en los rangos etarios 56-65 y 0-5 años respectivamente.

Todas estas gráficas se muestran para la Argentina en general, pero el pipeline las replica para las cinco regiones, para las 24 provincias y para los 530 departamentos. Siendo tan amplia la cantidad de resultados, las tareas que generan tablas de resumen se proponen como medio para identificar valores de interés y consultar luego las visualizaciones específicas. En la Fig. 4.4, por ejemplo, se comparan par a par el número absoluto de decesos, ya sea por EPOF o no, según sexo y rango de edad.

Las CSMR se analizaron diferenciando los códigos de enfermedades según los capítulos del CIE-10 en los que estuvieran clasificadas. Los resultados por año se ilustran en la Fig. 4.5, ya sea para ambos sexos (Fig 4.5A), para mujeres (Fig. 4.5B) y para varones (Fig. 4.5C). Los grupos de enfermedades, según orden decreciente de sus tasas en todo el

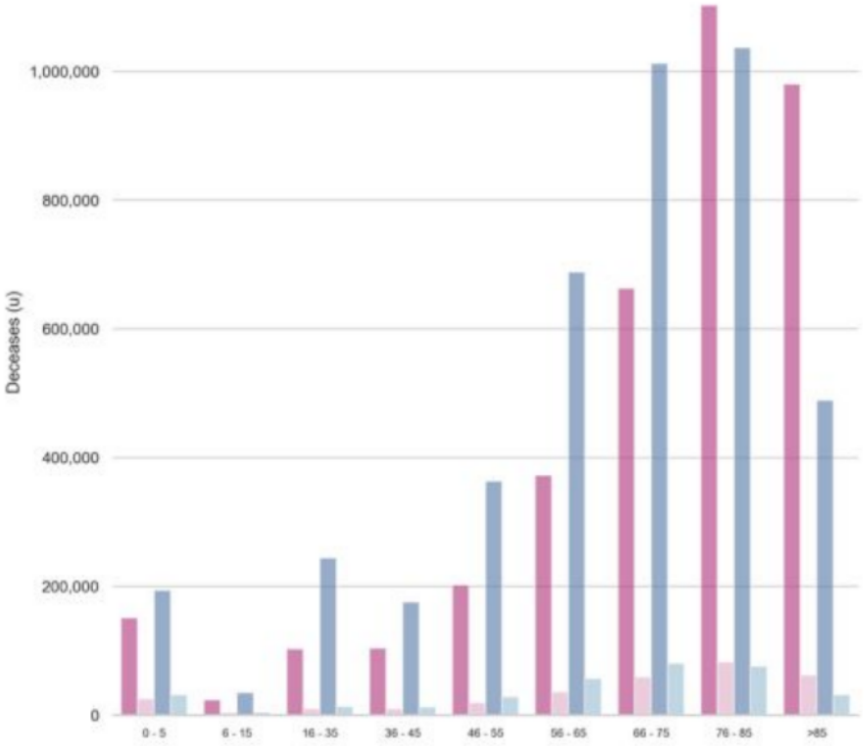


Figura 4.4: Producto del *pipeline EPOF* que ilustra los fallecimientos en el periodo según edad y sexo. El color rosa indica individuos femeninos y el azul masculinos. A la derecha de cada barra se grafica el número absoluto de decesos relacionados a alguna Enfermedad Poco Frecuente.

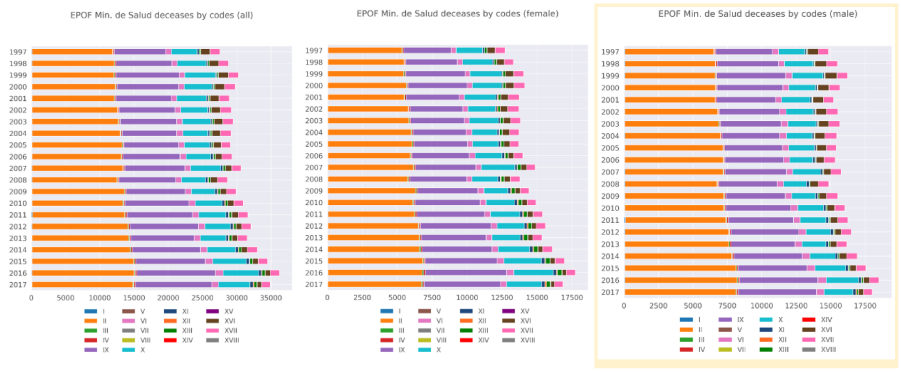


Figura 4.5: Producto del *pipeline EPOF* que ilustra las tasas anuales de mortalidad por causas específicas (CSMR) según los capítulos del CIE-10. 25A) fallecimientos en ambos sexos. (B): exclusivamente femeninos. (C): exclusivamente masculinos

periodo, fueron las neoplasias (90,5/1000), comprendidas en el Capítulo II de la CIE-10, las del aparato circulatorio (20/1000) en el Capítulo IX, digestivo (7,1/1000) en el Capítulo XI, malformaciones (3,4/1000) en el Capítulo XVII, en el periodo perinatal (3/1000) en el Capítulo XVI, respiratorio (2,4/1000) en el Capítulo X, neurológicas (2/1000) en el Capítulo VI, endocrinas (0,4/1000) en el Capítulo IV, de la sangre (0,3/1000) en el Capítulo III y osteomuscular (0,1/1000) en el Capítulo XIII, seguidas, con tasas muy menores, por los grupos de enfermedades clasificadas en los capítulos V (Trastornos mentales y del comportamiento), XII (Enfermedades de la piel y el tejido subcutáneo), XV (Embarazo, parto y puerperio), XIV (Enfermedades del aparato genitourinario) y VII (Enfermedades del ojo y sus anexos). El único año en que la tendencia cambia es en 2008, donde se registra una reducción general de todos los decesos, tal como se había discutido en las tendencias graficadas en la Fig. 4.2.

Las tasas de mortalidad por causas específicas anuales para cada provincia se visualizan utilizando un mapa de calor, como se muestra en la Fig. 4.6. Los valores más altos (representados por los colores cálidos), a lo largo de toda la serie temporal, se registraron siempre en las provincias patagónicas. En las provincias del NOA las tasas son bajas en comparación al resto del país (representadas por los colores fríos). En las provincias del NEA las tasas son variables y especialmente destacables en la provincia de Formosa. En las provincias del Centro, CABA tiene una tendencia creciente hacia el final del período analizado, mientras que Santa Fé y Córdoba son consistentemente altas para la región. Las explicaciones más plausibles a esta diferencia son, por un lado, la elevada densidad demográfica de la región, que vuelve más probable cualquier proceso de salud-enfermedad

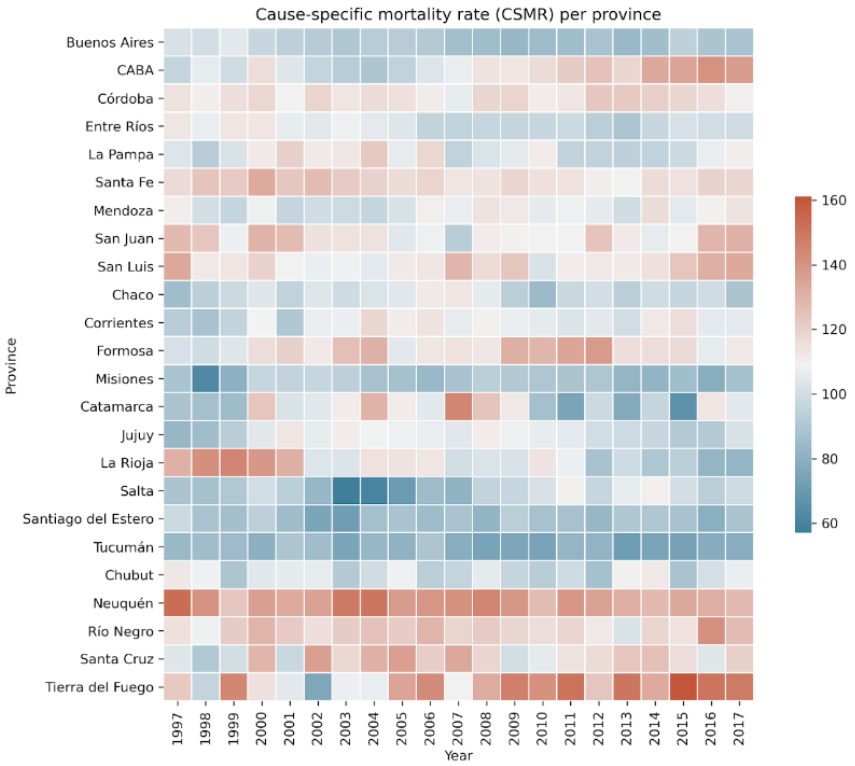


Figura 4.6: Producto del *pipeline EPOF* que ilustra las tasas de mortalidad por causas específicas anuales por provincia. Colores cálidos indican valores altos y colores fríos valores bajos

por mera probabilidad estadística. Por otro lado, en esta región se registra el índice per cápita más alto del país y las mayores concentraciones urbanas, con la consecuente ampliación en el área de servicios y asistencia. Aquí se localizan los hospitales y centros médicos especializados en dolencias específicas, a los que son derivados los pacientes de otras provincias. Al morir, el deceso se registra en la zona Centro, pero puede no tratarse de un caso originario de dicha región.

En la Fig. 4.7 se muestra la evolución del valor de CSMR por departamento a lo largo de todo el período, en tramos de cinco años. Este nivel de análisis permite discriminar la variabilidad de cada región o provincia. Los departamentos de Patagonia Norte registran las tasas más altas mientras que en el NOA siempre se advirtieron los valores más bajos, a excepción del período 2007-2011 en donde el departamento jujeño de Tilcara exhibió una tasa de mortalidad específica superior a la media nacional y muy disímil a lo ocurrido en el resto de la provincia. Sin embargo, en el periodo subsiguiente, se observó una estabilización de dicha tasa, alineándose con los valores característicos de los demás departamentos de su provincia.

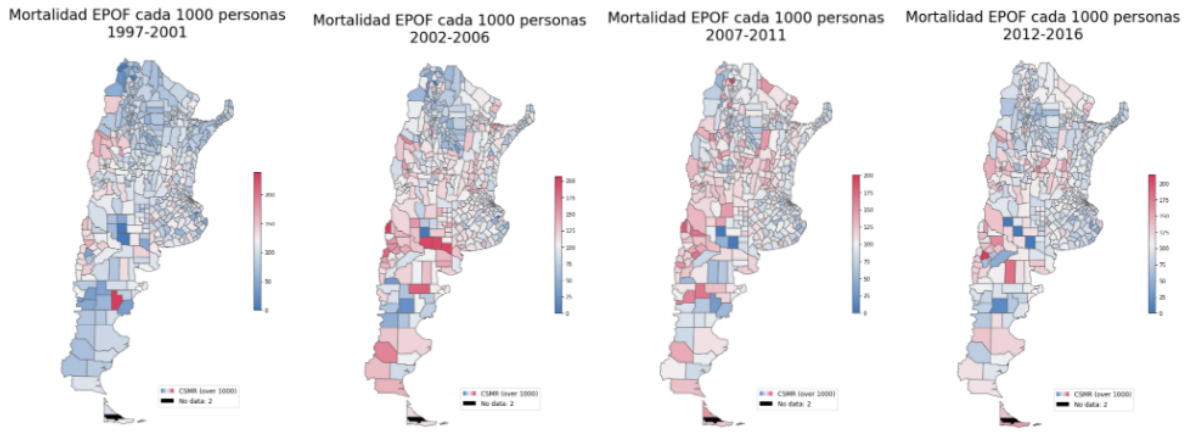


Figura 4.7: Producto del *pipeline EPOF* que ilustra las tasas de mortalidad por causas específicas (CSMR) para cada departamento en mapas de coropletas por quinquenio.

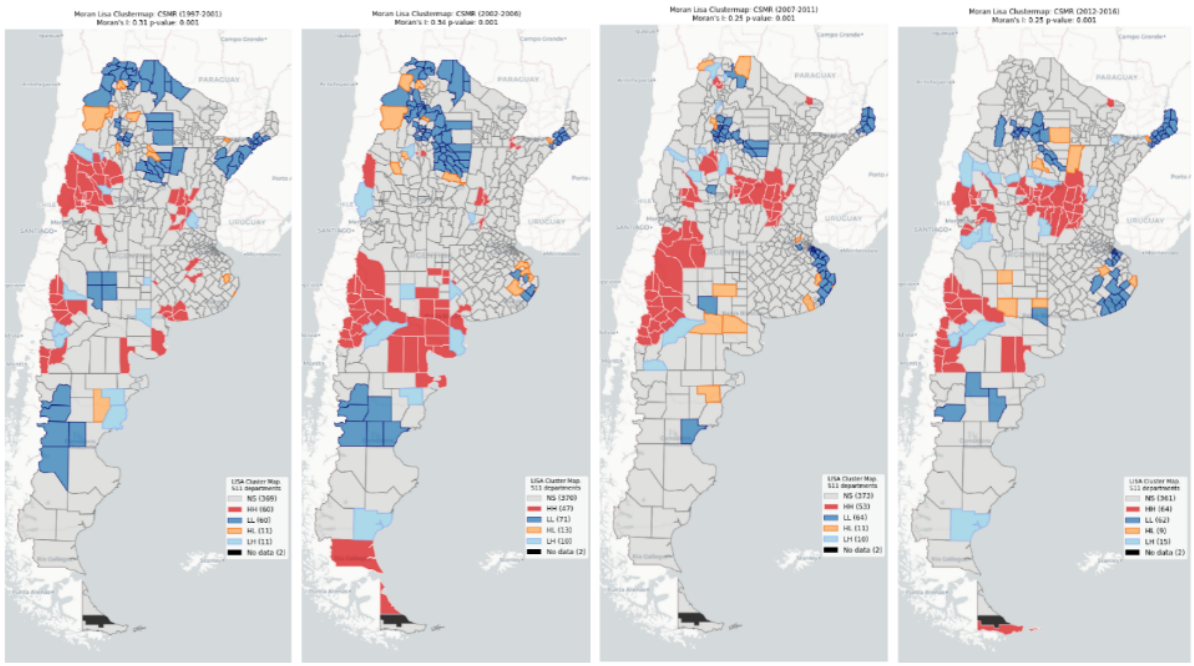


Figura 4.8: Producto del *pipeline EPOF* que exibe mapa de agrupamientos según Indice de Moran para las CSMR, por quinquenio.

Esta variabilidad fue la base para calcular el índice de autocorrelación de Moran, analizar la distribución espacial y detectar posibles agrupamientos. Este índice cuantifica la correlación entre la CSMR de un área específica y los valores correspondientes a las áreas circundantes. En la Fig. 4.8 se muestra la evolución de los clusters y outliers espaciales a lo largo de todo el período y en quinquenios, como en la figura anterior. Al principio del período, 60 departamentos se identificaron como hotspots y se encontraban distribuidos mayoritariamente en las provincias cuyanas de La Rioja y San Juan, en algunas de la Patagonia Norte (Neuquén y Río Negro) y en la provincia de Santa Fé de la región central. En los dos quinquenios siguientes se mantuvo esta tendencia en las tres regiones mencionadas, Cuyo, Centro y Patagonia, con menor identificación de hotspots, mientras que en el último quinquenio volvió a aumentar la cantidad de departamentos incluidos en estos clusters (64), incorporándose muchos de la provincia de Córdoba. Los departamentos del noroeste y la costa de la provincia de Buenos Aires se identificaron como coldspots para la mayoría de los períodos.

Este estudio sigue el espíritu de la Ley 26.689 (2011), que señala la necesidad de realizar en forma periódica estudios epidemiológicos que den cuenta de la prevalencia de las EPOF a nivel regional y nacional.

La calidad de vida de una persona que vive con EPOF se ve afectada por la falta o pérdida de autonomía debido a los aspectos crónicos, progresivos, degenerativos y frecuentemente potencialmente mortales de la enfermedad. Acarrean un alto grado de dependencia y de carga social, sanitaria y económica. Requieren, por lo tanto, de una atención biopsicosocial, que contemple la asistencia clínica especializada -en atención primaria y/o de alta complejidad- y los servicios sociales y apoyo psicológico para el paciente y su grupo familiar. A la ya mencionada dificultad para el acceso a un diagnóstico precoz, se suma la escasez de información. Estas características comunes de las afecciones y sus consecuencias definen a todas estas personas como un colectivo social vulnerable. Esperamos que los resultados presentados constituyan un insumo preliminar para definir políticas sanitarias y estrategias a favor de pacientes y familias. La Ley 26.689 señala la necesidad de realizar en forma periódica estudios epidemiológicos que den cuenta de la prevalencia de las EPOF a nivel regional y nacional. Entendemos que nuestro análisis es un ejemplo sobre los abordajes específicos y muy económicos que pueden desarrollarse,

trabajando de manera crítica con bases de datos masivas. Los resultados de la tasa de mortalidad estandarizada (SMR) por región permite realizar un seguimiento espacial de las muertes relacionadas a EPOF, un insumo preliminar e imprescindible para definir políticas sanitarias y estrategias a favor de quienes las padecen. Las asociaciones de pacientes y familias tienen un rol fundamental. Existen más de 60 de estas organizaciones civiles en Argentina. Dada su representación federal, funcionan en ocasiones como la única red de contención y asesoría real, aportando información y herramientas a las familias para acceder a coberturas, prácticas de salud y medicamentos.

4.4. Caso 2 - Epidemiología de las Muertes Fetales en Argentina: variación espacial y temporal

4.4.1. Introducción, fuentes de datos y objetivo

Se han propuesto diferentes definiciones de Muerte Fetal (MF), dependiendo del peso del feto o de la edad gestacional. Estos varían de un país a otro y no son equivalentes, lo que dificulta tener un estándar único aceptado internacionalmente. La Organización Mundial de la Salud (OMS) recomienda que se cuenten todas las muertes fetales, pero las comparaciones internacionales solo incluyen las muertes fetales al final del embarazo, con un peso mayor o igual 1000 gramos o una edad gestacional mayor o igual a 28 semanas (Blencowe et al., 2016). Según el Departamento de Estadísticas e Información en Salud del Ministerio de Salud de la Nación Argentina, la muerte fetal se define como "la que ocurre antes de la completa expulsión o extracción del producto de la concepción, independientemente de la duración del embarazo". La muerte se caracteriza porque, después de dicha separación, el feto no respira ni muestra otros signos de vida, como latidos del corazón, pulsaciones del cordón umbilical o movimientos efectivos de los músculos voluntarios [3]. Esta definición, también adoptada por la Organización Panamericana de la Salud (OPS) [4] es la que se seguirá en este trabajo.

En 2009, la tasa mundial estimada de muertes fetales por cada mil nacidos vivos fue de 18,9 y el 76,2 % de estos decesos ocurrieron en el sur de Asia y el África subsahariana (Cousens et al., 2011). Las mismas tendencias se mantuvieron una década

[3] Estadísticas vitales. Información básica año 2017. Ministerio de Salud de la Nación. Buenos Aires, Argentina, Diciembre 2018. ISSN: 1668-9054 Serie 5 N 61

[4] Lineamientos básicos para el análisis de la mortalidad. https://iris.paho.org/handle/10665.2/34492

después (Souza and Bahl, 2022). Las grandes disparidades en la incidencia de las muertes fetales se estructuran principalmente en torno al eje socioeconómico. A nivel mundial, el 98 % se produce en países de ingresos bajos y medios (Blencowe et al., 2016). Sin embargo, también se observan tasas elevadas en los países de altos ingresos entre los grupos vulnerables y las minorías étnicas desfavorecidas. Basándose en la CIE-10, la OMS publicó la Clasificación Internacional de Enfermedades de Mortalidad Perinatal (CIEMP) en 2016, que enumera la causa de la muerte perinatal utilizando códigos CIE-10, separados por el momento de la muerte, y las condiciones maternas asociadas (Organization et al., 2016).

La tragedia que representa la muerte fetal para las familias a menudo no se aborda en las agendas políticas ni en los programas de salud gubernamentales. Resulta relevante mejorar la obtención de datos como insumo básico para ampliar el alcance de intervenciones de probada eficacia para la supervivencia posnatal. El objetivo de este estudio retrospectivo eco-epidemiológico es analizar el comportamiento temporal y espacial de las muertes fetales y sus causas en Argentina desde 1994 hasta 2019.

Se recopilaron bases de datos de nacimientos y defunciones fetales producidas entre 1994 y 2019, organizadas según la estructura jerárquica administrativa del país (provincias y departamentos). Los datos fueron proporcionados por la Dirección de Estadísticas e Información de Salud y se obtuvieron de las actas de nacimiento y los certificados de defunción fetal. Dada la profundidad temporal de los registros, se utilizaron tablas de conversión para estandarizar los códigos en las bases de datos entre la transición de la codificación CIE-9 a la CIE-10 (desde el año 1996 en adelante). Las causas de muertes fetales se agruparon en seis conjuntos según las categorías propuestas por (Hoyert and Gregory, 2016):

- P00: feto afectado por condiciones maternas no relacionadas con el embarazo actual.
- P01: feto afectado por complicaciones maternas del embarazo.
- P02: feto afectado por complicaciones de placenta, cordón umbilical y membranas.
- P95: muerte fetal de causa no especificada.
- Q00-Q99: malformaciones congénitas, deformidades y anomalías cromosómicas.
- Otros: todas las demás causas posibles combinadas.

A su vez, se consideraron las cinco regiones geográficas, que comparten características ambientales similares: NOA (provincias de Catamarca, Jujuy, La Rioja, Salta, Santiago del Estero y Tucumán), NEA (provincias de Misiones, Formosa, Corrientes, Chaco y Entre Ríos), Cuyo (provincias de Mendoza, San Juan y San Luis), Centro (provincias de Buenos Aires, Córdoba, La Pampa, Santa Fe y Ciudad Autónoma de Buenos Aires) y Patagonia (provincias de Chubut, Neuquén, Río Negro y Santa Cruz). En la Fig. 4.9 se resume esta información, junto con las estimaciones demográficas para 2023, publicadas por el INDEC. La Región Centro tiene la mayor densidad de población, allí reside el 65,2 % de la población total del país. Por el contrario, en relación al tamaño de su territorio, la Región Patagónica presenta la menor densidad poblacional: 3,3 habitantes por kilómetro cuadrado.

Este estudio, al igual que los otros casos presentados en esta tesis, sigue los lineamientos éticos propuestos por el Ministerio de Salud de la Nación de Argentina, que eximen de la obtención del consentimiento informado a los estudios epidemiológicos que utilizan registros o información pública o disponible públicamente.

4.4.2. Metodología

Para identificar posibles tendencias en el tiempo, todo el período se dividió en intervalos de cinco años. Sobre la base de los registros de nacimientos y defunciones, se calcularon el número total o absoluto de Muertes Fetales (MFA) y las Tasas de Muertes Fetales (TMF), siendo TMF = número de muertes fetales / número total de recién nacidos *1000 independientemente del sexo/género, para todo el país, en las cinco regiones geográficas, y para cada provincia y sus departamentos. Para identificar cambios significativos en las TMF, utilizamos un análisis de regresión de puntos conjuntos o Joinpoint. Este método identifica los años en los que se produce un cambio de tendencia significativo y calcula el cambio porcentual anual (CPA) en las tasas entre dichos puntos en el tiempo. La significancia estadística se calculó considerando un valor p inferior a 0,05, mediante método de Monte Carlo. El índice de autocorrelación I_{Moran} se utilizó para determinar el patrón espacial de los distintos valores de las TMF, incluyendo 525 departamentos, utilizando el criterio de Rook o de contigüidad. A partir del Índice Global de Moran se calculó un mapa de Indicadores Locales de Asociación Espacial (LISA) siguiendo la metodología descrita por (Rey et al., 2023). Esto determina si un valor dado en un departamento y la media

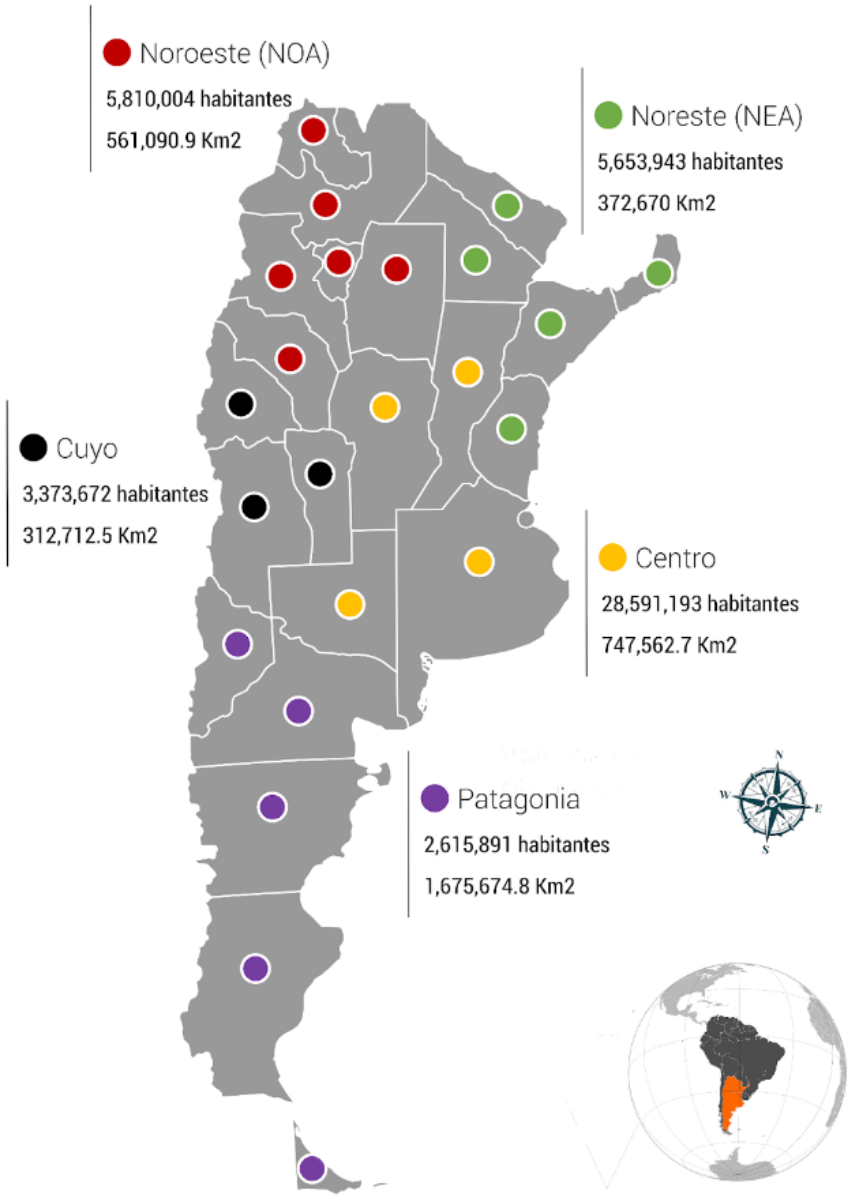


Figura 4.9: División administrativa de Argentina y sus cinco regiones geográficas, incluyendo población y área total en kilómetros cuadrados, según Censo Nacional 2022 (INDEC, 2023).

de sus vecinos son más similares (alto-alto o HH, bajo-bajo o LL) o diferentes (alto-bajo o HL, bajo-alto o LH) de lo que se esperaría por casualidad. La significancia se determinó con un nivel de confianza de 0,05 utilizando la prueba de Monte Carlo (999 permutaciones) bajo la hipótesis nula de distribución aleatoria. El análisis se realizó utilizando la biblioteca exploratoria de análisis de datos espaciales ESDA del paquete PySAL (Rey and Anselin, 2009), como se realizó en análisis previamente presentados. El código fuente de proyecto de datos, *pipeline MF*, se encuentra en el repositorio Github, accesible a través del siguiente enlace: github.com/LeoMorales/stillbirth-pipeline.

4.4.3. Resultados y conclusión

Las bases de datos documentaron 18.405.630 nacidos vivos de 1994 a 2019, con un total de 173.330 muertes fetales. Las TMF mostraron una tendencia a la disminución a lo largo del periodo. Las regiones de Cuyo y NEA destacan por tener TMF considerablemente mayores que el resto para los dos primeros quinquenios (1994-1998, 1999-2003) y por mostrar una disminución significativa posterior. Patagonia tuvo siempre las tasas más bajas, mientras que la región NEA fue la única área que mostró un aumento neto en las TMF al final del período de investigación. Para el país en su conjunto, esta tasa disminuyó 3,8 puntos.

El análisis de regresión de puntos conjuntos o Joinpoint destaca estas diferencias interregionales en las tendencias. La región NOA muestra una disminución estable en la mortalidad desde 2002, mientras que el NEA muestra un pico de incremento en el año 1999, después del cual comienza a disminuir hasta 2010. Resulta muy relevante el aumento notable en TMF en todas las regiones entre 1999 y 2003, ya que tuvo lugar en medio de una grave crisis socioeconómica en Argentina. La misma situación se produjo entre 2011 y 2012, siendo este aumento especialmente importante en la región NOA.

La Fig. 4.10 muestra el porcentaje de muertes fetales categorizadas por causa, tanto a nivel nacional como regional, durante cada período de cinco años. Para el conjunto de Argentina, P02 (complicaciones de la placenta, cordón umbilical y membranas), P95 (causa no especificada) y "otras causas" son las categorías con mayores porcentajes de ocurrencia. En conjunto, explican más del 80 % de las muertes (para todos los niveles de análisis). La prevalencia de muertes fetales por causas congénitas (Q00-Q99) se man-

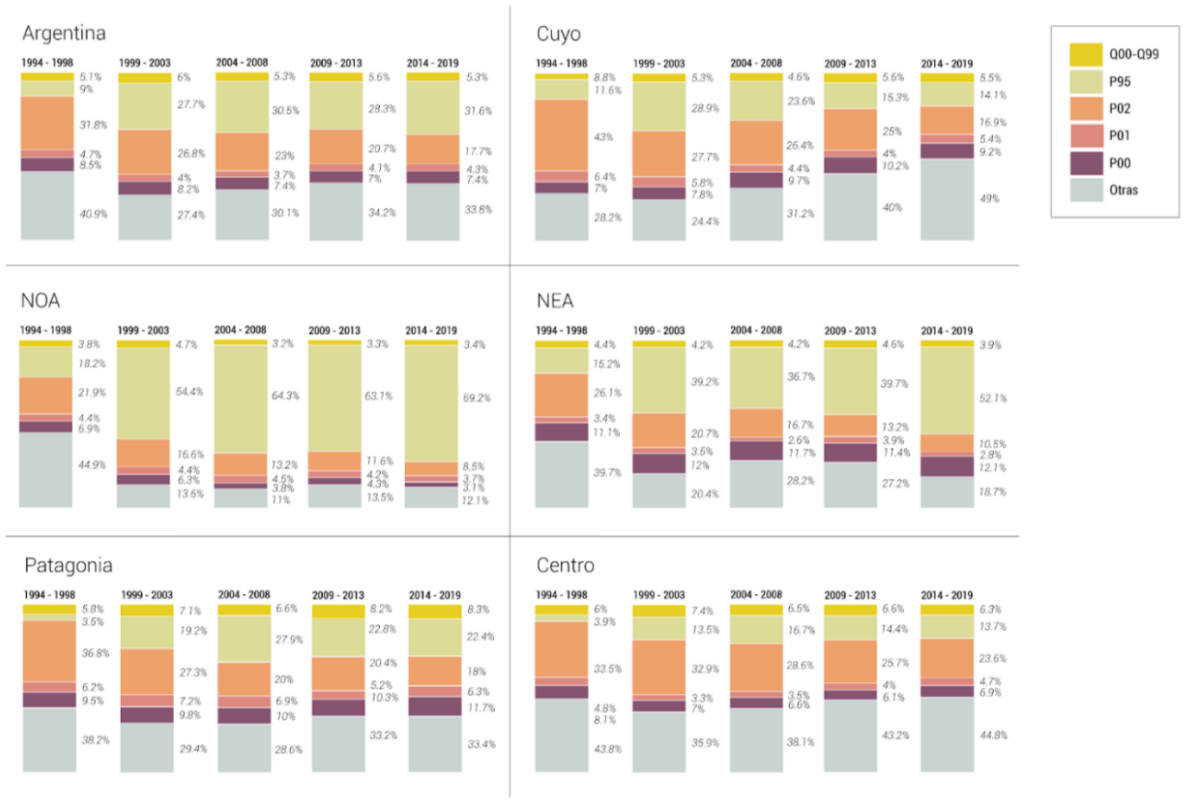


Figura 4.10: Producto del *pipeline MF* que presenta los gráficos de barras apiladas con porcentajes de muertes fetales categorizadas por causas en intervalos de cinco años tanto a nivel regional como nacional. P00: Feto afectado por condiciones maternas no relacionadas con el embarazo actual; P01: Feto afectado por complicaciones maternas del embarazo; P02: Feto afectado por complicaciones de placenta, cordón umbilical y membranas; P95: Muerte fetal por causa no especificada; Q00-Q99: Malformaciones congénitas, deformidades y anomalías cromosómicas; Otro: todas las demás causas posibles combinadas.

tiene estable y relativamente baja tanto a nivel nacional como dentro de cada región, registrando los porcentajes más altos en Centro y Patagónica. Las MF relacionadas con condiciones maternas (P00 y P01) también presentan en general una ocurrencia relativamente baja y estable. La mayor disminución se observó en las muertes fetales provocadas por complicaciones de placenta y cordón (P02).

Por otro lado, el número de MF clasificadas como no especificados (P95) muestra una tendencia general al alza, comenzando para todas las regiones alrededor del segundo período (1999-2003). Como se señaló anteriormente, coincidió con una grave crisis socioeconómica que afectó a todo el país. Esta tendencia ascendente es particularmente pronunciada en las regiones NOA y NEA, mientras que la región de Cuyo logró revertirse y reducirse a la mitad al final del período estudiado.

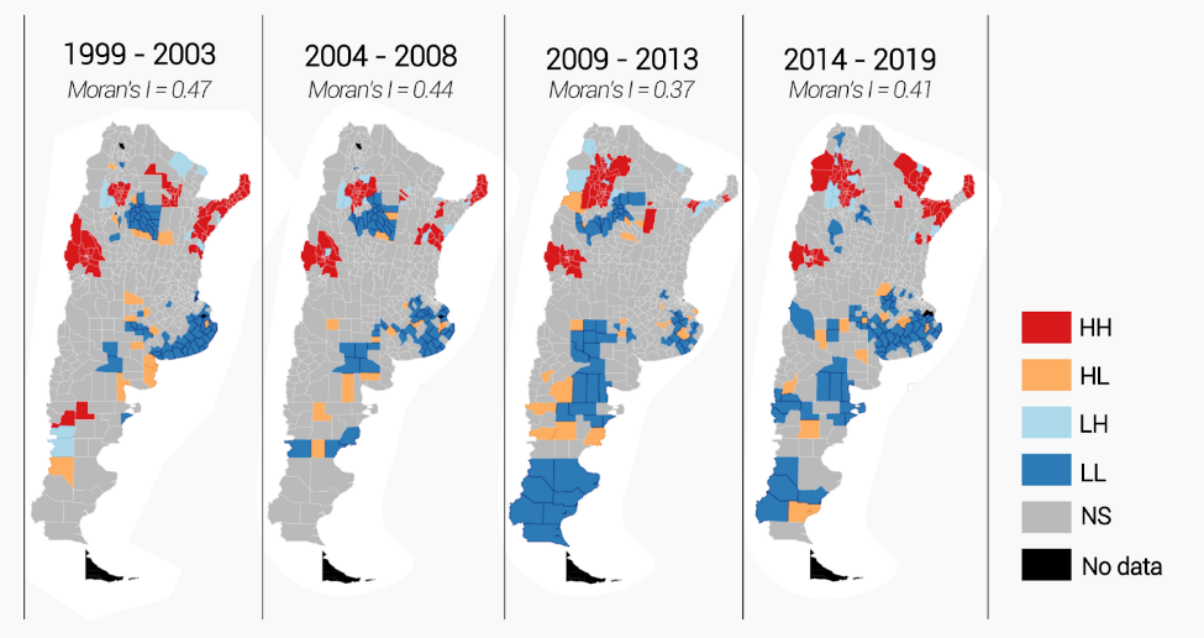


Figura 4.11: Producto del *pipeline MF* que presenta los mapas de agrupamientos de las TMF departamentales. En rojo: departamentos con TMF altas que están rodeados de vecinos con valores altos (HH o high-high). En naranja: departamentos con TMF altas pero rodeados de TMF bajas (HL o high low). En azul claro: departamentos con TMF baja pero rodeados de TMF altas (LH o low-high). En azul: agrupamientos con tasas bajas (LL o low-low). En gris, no significativo (NS).

La Fig. 4.11 resume los resultados del análisis espacial con las Tasas de Muertes Fetales entre los años 1999 a 2019, evidenciando una distribución no aleatoria del fenómeno. Los puntos calientes o hotspot, conformados por departamentos con altas tasas rodeadas de otros con altas tasas, son más frecuentes en algunas regiones del norte del país. Ciertas zonas de la Patagonia y la región central tienen puntos fríos o coldspot, que se definen como áreas con tasas bajas rodeadas de otras áreas con tasas bajas. Todos los análisis indicaron una autocorrelación espacial positiva y estadísticamente significativa. El valor del I_{Moran} para el período completo de 1999 a 2019 fue de 0,57.

Según los registros estudiados, en las regiones Noroeste y Nordeste las tasas de muerte fetal siguen siendo elevadas desde el principio de la serie temporal. El NOA incluso experimentó un aumento significativo entre 2000 y 2003, del que no se recuperó por completo en los años siguientes. Esto pone de relieve la necesidad de analizar el fenómeno en diferentes escalas espaciales y temporales, para evitar sesgos interpretativos. Claramente, la persistencia de altos niveles de MF en las provincias del norte está asociada con variables regionales y/o locales específicas. La tasa de mortalidad infantil (TMI) es un indicador

fiable del nivel de desarrollo de una población y resulta de la interacción entre las características de la población, los factores de riesgo o patógenos y el entorno social, físico y biológico (Behm, 2011). En Argentina, el NOA tiene la TMI más alta en comparación con la llamada Pampa Húmeda (Bolsi et al., 2005), y en 2010 era la segunda región con mayor TMI después del NEA (Mazzeo, 2014). Esto podría estar relacionado con factores ambientales, como la altitud: esta región incluye el piedemonte y ecorregiones andinas que alcanzan hasta 5000 metros sobre el nivel del mar. Un estudio de Chapur, Alfaro, Bronberg y Dipierri (Chapur et al., 2017) encontró que la mortalidad posneonatal (28-365 días después del nacimiento) estaba directamente relacionada con esta característica geográfica. La ecología de las zonas de gran altitud es compleja debido a los bajos niveles de oxígeno, las bajas temperaturas, la intensa radiación solar y la baja humedad ambiental. Estas variables impactan en la calidad de vida de la población, además de las dificultades económicas y estructurales para acceder a servicios de salud y/o información sanitaria. El NOA es una región con un bajo nivel de desarrollo económico y la tasa de pobreza más alta del país según el INDEC 2023 [5], lo que la coloca en desventaja comparativa con otras regiones y la haría particularmente vulnerable a altas tasas de MF.

Las mejoras en la atención obstétrica, las instalaciones del sistema de salud y el seguimiento temprano del embarazo pueden explicar la disminución de las muertes fetales asociadas con factores maternos (códigos P00-P02, Fig. 4.10) durante el período de estudio. Estos cambios permiten partos más controlados conociendo las posibles complicaciones. Al mismo tiempo, se ha producido un aumento de causas no especificadas en el certificado de defunción (código P95). Probablemente esto se deba a la falta de información sobre muchas muertes, lo que requeriría un examen post mortem más detallado. En cuanto a las muertes causadas por malformaciones congénitas, los porcentajes se mantuvieron similares en todos los periodos. Este fenómeno requiere un mayor análisis, ya que ha tendido a aumentar en las últimas décadas en diferentes países y también en Argentina, como expresión de la transición epidemiológica. Esta transición se puede definir como el cambio de una fase dominada por enfermedades infecciosas a otra dominada por enfermedades crónico-degenerativas (Omram, 2001, Rosano et al., 2000, Bronberg et al., 2021). En

[5]Informes técnicos / Vol. 7, n° 63. ISSN 2545-6636

18 países con datos confiables, las anomalías congénitas representan una mediana del 7,4 por ciento de las muertes fetales (Blencowe et al., 2016).

Gracias a la colaboración con el Departamento de Estadística e Información en Salud, se creó una base de datos integral. La investigación presenta un análisis preliminar de las tendencias. A pesar de la importante heterogeneidad observada destacamos que las tasas de muertes fetales muestran una tendencia secular negativa, especialmente en las regiones más desarrolladas de Argentina (Cuyo, Centro y Patagonia). Argentina ha implementado políticas a nivel nacional que han demostrado ser altamente beneficiosas para la salud pública. Una de esas medidas es el enriquecimiento de la harina con ácido fólico, destinado a prevenir defectos congénitos del tubo neural. Como resultado, ha habido una reducción significativa de las muertes fetales causadas por estos defectos (Calvo and Biglieri, 2008, Bronberg et al., 2023). Dada la continua disminución de las tasas de fertilidad en los últimos años (DEIS, 2020), es crucial mantener y ampliar los esfuerzos para minimizar la mortalidad fetal en todo el país, específicamente en las regiones y provincias más afectadas.

4.5. Caso 3 - Bronquiolitis

4.5.1. Introducción, fuentes de datos y objetivo

Las Infecciones Agudas del Tracto Respiratorio Inferior o Infecciones Respiratorias Agudas Bajas (IRAB) son una de las principales causas de muerte a nivel mundial, con más de 4 millones de fallecimientos anuales (Internacionales, 2017). Dentro de este grupo se incluyen la bronquiolitis y la neumonía, particularmente relevantes debido a su impacto significativo en la morbi-mortalidad de los pacientes infantiles (de Pediatría and Subcomisiones, 2015). En Argentina, las IRAB continúan siendo una importante causa de morbi-mortalidad (MINSAL, 2010). Anualmente, alrededor de 500 a 600 niños menores de 5 años fallecen por infecciones respiratorias, situándose como la tercera causa de muerte después de las muertes perinatales y las anomalías congénitas y cromosómicas. En 2005, las Infecciones Respiratorias Agudas Bajas (IRAB) fueron responsables de 68.605 altas hospitalarias en pacientes menores de 5 años, representando el 21,5 % del total de altas en ese rango de edad (N = 318.347). Por su parte, las IRAB fueron la segunda causa más común de hospitalización en este grupo, superadas solo por las complicaciones durante el

período perinatal 22,06 %) y por encima de otras enfermedades infecciosas en conjunto (11,2 % de los casos) [6].

La bronquiolitis se refiere generalmente al primer episodio de sibilancias en lactantes menores de 12 meses (Meissner, 2016). Su causa principal es el virus respiratorio sincitial, en inglés Respiratory Syncytial Virus (RSV), que tiene patrones estacionales globales influenciados por el clima. En el hemisferio sur, la epidemia de RSV ocurre entre marzo y junio, disminuyendo de agosto a octubre (Obando-Pacheco et al., 2018). La bronquiolitis demanda importantes recursos sanitarios, especialmente en casos de hospitalización. El escenario se tornó más complejo con el advenimiento de la pandemia de COVID-19, que ha alterado la epidemiología de las infecciones respiratorias agudas en niños, con brotes más pequeños y retrasos en las temporadas de RSV.

Argentina muestra una tendencia descendente en la tasa de mortalidad por enfermedades respiratorias en menores de 5 años, pero persisten disparidades espaciales y demográficas. Puntualmente en Puerto Madryn (Chubut, Patagonia Norte), las condiciones ambientales que aumentan el riesgo de virus respiratorios ocurren entre otoño e invierno. Las temperaturas extremas más frías en la ciudad se registran durante los meses de junio a julio (Cannizzaro, A., & Nuñez de la Rosa, D., 2020). A nivel nacional, según estadísticas del año 2018, la mayoría de las muertes infantiles en menores de 5 años fueron causadas por IRAB. Más del 50 % de estas muertes se registraron entre junio y septiembre, lo que destaca la importancia de este período crítico para la salud infantil. Los estudios sindémicos han señalado la naturaleza biosocial de las enfermedades respiratorias, donde interactúan factores ambientales, sociales y económicos. Dentro de este contexto, se observa que las desigualdades sociales pueden conducir a condiciones de vida adversas que favorecen la agregación de enfermedades infecciosas y no infecciosas. La infección por RSV se ha asociado con coinfecciones bacterianas y una mayor gravedad de la enfermedad en individuos susceptibles. Esto incluye factores socioeconómicos, geográficos, culturales y genéticos, lo que aumenta la necesidad de registros y evaluaciones localizados (Nair et al., 2013). Variables demográficas como la densidad de población, el tabaquismo paterno en el hogar, la presencia de un hermano y los antecedentes de hospitalización influyen en la transmi-

[6] Ministerio de Salud de la Nación. Sistema Estadístico de Salud. (2007). Egresos de Establecimientos Oficiales por Diagnóstico. 2005. Serie 11 – Número 1. Buenos Aires, Dirección de Estadísticas e Informacion de Salud. https://www.argentina.gob.ar/sites/default/files/serie11nro1.pdf

sión de virus respiratorios (Pitzer et al., 2015). Su naturaleza y efectos pueden ser muy específicos y requerir análisis locales. En este contexto, el hacinamiento en los hogares es un factor de riesgo de enfermedades respiratorias reconocido mundialmente. Sin embargo, faltan investigaciones sobre su impacto en los ingresos hospitalarios por IRAB entre niños menores de cinco años en Argentina. La aglomeración, urbanización y migración pueden contribuir al crecimiento económico, pero existe el riesgo de desigualdad espacial a corto y mediano plazo. La desigualdad espacial abarca la distribución desigual de recursos y servicios, incluyendo atención sanitaria, bienestar, servicios públicos, ingresos e infraestructuras. La distribución espacial de estas características puede examinarse en función de la proximidad, la distancia, la agrupación y concentración. En este escenario, comprender y medir las diferencias espaciales y sus tendencias puede ser fundamental para desarrollar políticas y estrategias que reduzcan la morbilidad, la mortalidad y mejoren el acceso a los recursos sanitarios.

Los objetivos de esta investigación fueron determinar la incidencia en los ingresos hospitalarios de infantes menores de un año diagnosticados con bronquiolitis en Puerto Madryn y analizar la distribución de este fenómeno en el espacio urbano, georeferenciando los domicilios y correlacionando con variables socioeconómicas. Se buscó facilitar la comprensión y la visualización de los factores subyacentes a la manifestación local de la enfermedad a través de la elaboración de un mapa que identifique las áreas vulnerables dentro de la ciudad. Se diseñó un análisis de tipo transversal que abarcó a la totalidad de los pacientes que recibieron el alta médica por bronquiolitis en el hospital público de Puerto Madryn durante todo el año 2017, de enero a diciembre. Puntualmente, se tuvieron en cuenta los pacientes con diagnóstico de bronquiolitis aguda. El protocolo utilizado fue evaluado y aprobado por el Comité Interdisciplinario de Docencia e Investigación del Hospital Público Zonal "Dr. Andrés R. Isola" (15 de diciembre de 2017).

Tres conjuntos de datos crudos se utilizaron en este trabajo, los casos de bronquiolitis, la información socioeconómica y las capas geográficas de la ciudad.

Los casos se obtuvieron del archivo del hospital y para cada registro se recopilaron los siguientes datos: fecha de nacimiento, edad en meses, fecha de ingreso y alta, domicilio, diagnóstico primario de alta con su correspondiente código CIE-10 alfanumérico J21 y si hubo necesidad de oxígeno.

Sumado a estas variables y como segundo conjunto de datos de interés, se consideraron los factores socioeconómicos para los radios censales de la ciudad. La pobreza es la exclusión o limitación en el acceso a determinadas condiciones materiales como consecuencia de la carencia de recursos necesarios. A pesar del carácter multidimensional y complejo que puede tener, la pobreza es una condición en la cual una o más personas tienen un nivel de bienestar inferior al mínimo necesario para la sobrevivencia. Uno de los métodos para medirla es el índice de Necesidades Básicas Insatisfechas (NBI), que evalúa un conjunto de necesidades estructurales relacionadas con vivienda, educación, salud, infraestructura pública, etc. Este método califica como población en pobreza a aquella que tiene al menos una necesidad básica insatisfecha y como pobres extremos a los que presentan dos o más indicadores en esa situación (Barneche et al., 2010).

En relación a la condición de salud analizada, se consideró como medida más sensible el hacinamiento (relación entre el número de dormitorios respecto al número total de habitantes de cada vivienda). Operativamente, se considera que existe hacinamiento crítico en una vivienda cuando hay más de tres personas por habitación. Esta información provino del Instituto Nacional de Estadísticas y Censos [7] aplicación mediante del método de Feres & Mancero (Feres and Mancero, 2001).

Por último, se acondicionó una capa de polígonos con los bordes de los radios censales de Puerto Madryn.

4.5.2. Metodología

En las tareas de adecuación de los datos, la primera y más importante fue la georreferenciación de los casos de bronquiolitis. Se tomó la dirección de la casa del paciente y se obtuvo su latitud y longitud utilizando la interfaz Python del servicio de geocodificación de Google (Google Maps Geocoding API) [8]. Como resultado, se obtuvo una nueva capa geográfica, conformada de puntos en el espacio territorial de Puerto Madryn. Esta capa se combinó con la de polígonos de la ciudad, determinando la cantidad de casos por cada radio censal a partir de una operación geográfica (de inclusión). Posteriormente a partir de los datos socioeconómicos agregados por identificador de unidad censal, se asignó el

[7] INDEC – Instituto Nacional de Estadísticas y Censos. (2010). Necesidades Basicas Insatisfechas en la República Argentina. https://www.indec.gob.ar/indec/web/Nivel4-Tema-4-47-156

[8] Google. (2022). Google maps geocoding API. https://developers.google.com/maps/documentation/geocoding/overview

porcentaje de hogares hacinados a cada radio como un nuevo atributo en el mapa de polígonos.

Una tarea de visualización se encargó de aplicar una segmentación sobre el rango del porcentaje de hacinamiento de cada radio censal y de adjudicar un patrón de colores para ilustrar la intensidad de cada categoría. Además, se dibujaron sobre esta capa los casos, tanto admisiones como re-admisiones.

En cuanto a las tareas de análisis espacial, dado que se definió en término de polígonos y atributos, se reutilizaron los procedimientos utilizados en casos de uso previo, solo que se trabajó en un nivel mucho más enfocado: el municipal. Así, utilizando las funciones ya implementadas sobre la biblioteca de análisis exploratorio de datos espaciales PySAL se calcularon tanto el I_{Moran} Global univariante sobre los casos de bronquiolitis como el I_{Moran} bivariante local (BILISA) entre los casos de bronquiolitis y el hacinamiento. El primero se calculó con el fin de probar la hipótesis nula de distribución espacial aleatoria de los casos en toda la ciudad y el segundo para medir la fuerza y la dirección de la relación espacial entre los dos atributos.

El código fuente de este caso de uso se encuentra en el repositorio Github, accesible a través del siguiente enlace: github.com/LeoMorales/pi-bronquiolitis-pipeline.

4.5.3. Resultados y conclusión

Un total de 120 menores de 12 meses dados de alta del hospital cumplieron con los criterios de inclusión y fueron seleccionados para este estudio. La mediana de edad fue de 4,45 meses. La estancia media en internación fue de 7,30 días. Catorce pacientes (11,66 %) tuvieron que ser reingresados en el hospital, 7 de ellos en un plazo incluso inferior al mes tras su alta. De los 120 casos hospitalizados, 100 (83,33 %) vivían en barrios que previamente estaban registrados como hogares con hacinamiento en las Encuestas Permanentes de Hogares de 2001 y 2010. Además, la mayoría de los casos que requirieron reingreso (12 de 14) vivían en estos barrios.

Tanto la distribución espacial de los casos como la presencia de hogares hacinados a lo largo de los radios censales no son aleatorias. Los valores de NBI exhiben un patrón de valores altos en el oeste y noroeste de la ciudad, como se ilustra en la Fig. 4.12, mientras que el sur y especialmente el oriente de la ciudad tiene el menor porcentaje de hacinamiento

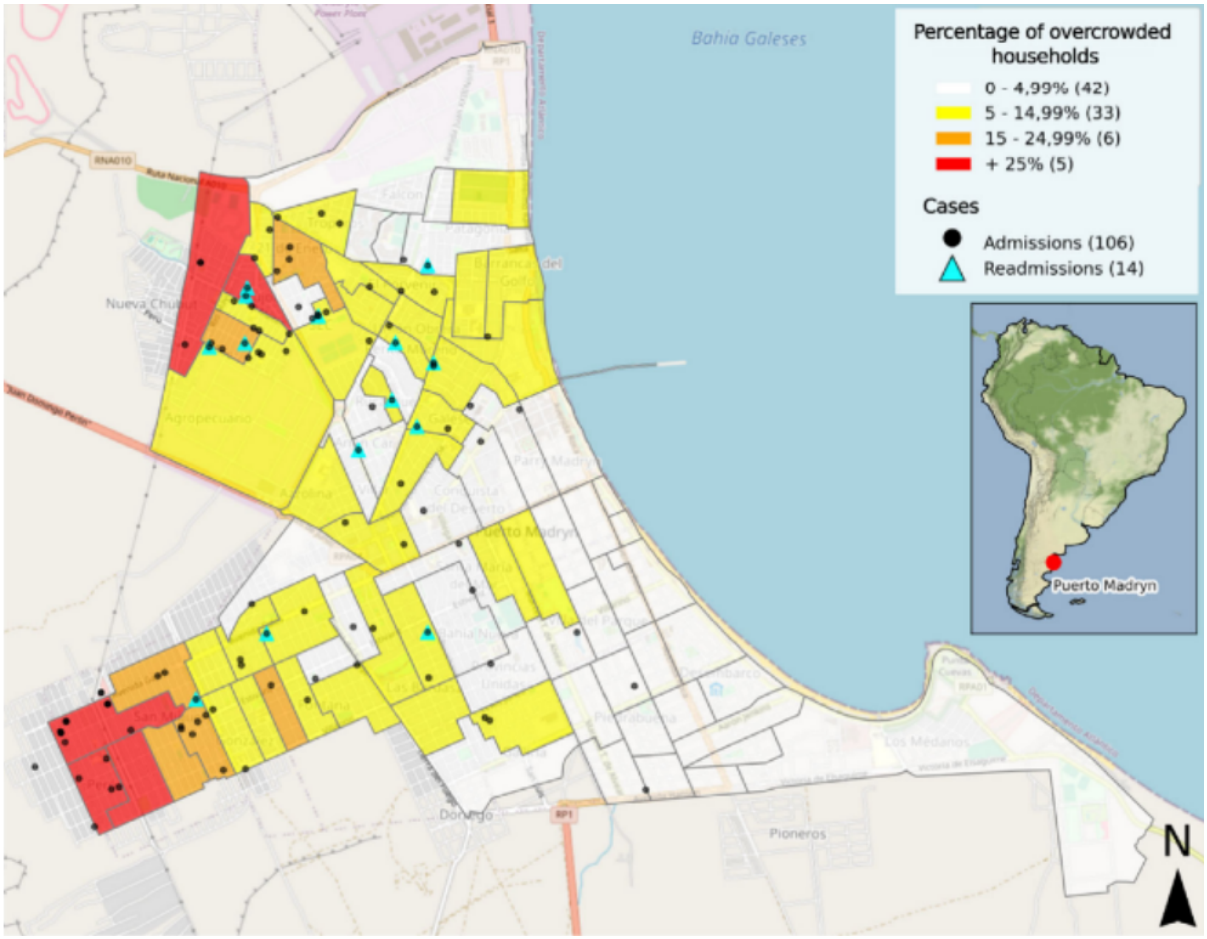


Figura 4.12: Producto del *pipeline Bronquiolitis* que muestra un mapa de Puerto Madryn con los radios censales coloreados según el porcentaje de hogares hacinados. Entre paréntesis se muestra el número de radios censales que entran en cada categoría. Los puntos negros y los triángulos celestes representan los casos de bronquiolitis georreferenciados (ingresos y readmisiones, respectivamente) en 2017.

y casi ningún caso de bronquiolitis. Este patrón espacial emergente fue estadísticamente significativo. Obtuvimos un I_{Moran} global univariado positivo de 0,31 (p=0,001).

Se identificaron importantes puntos calientes y fríos espaciales en el noroeste y sureste de la ciudad, respectivamente como se muestra en la Fig. 4.13.

Obtuvimos un coeficiente I_{Moran} bivariado global de 0,45, que fue estadísticamente significativo según las permutaciones de Monte Carlo (valor p=0,001). El coeficiente alto y positivo indica que valores similares de las dos variables tienden a agruparse en el espacio. El diagrama de dispersión de Moran (Fig. 4.13A) proporciona una representación visual de las asociaciones espaciales. En el eje X están los valores estandarizados de casos de bronquiolitis para cada unidad espacial (polígono) y en el eje Y el desfase espacial de los valores de hacinamiento. Los puntos en los cuadrantes superior derecho (o alto-alto=HH) e inferior izquierdo (o bajo-bajo=LL) indican una asociación espacial positiva, con valores superiores e inferiores, respectivamente, que la media de la muestra. Los cuadrantes inferior derecho (o alto-bajo=HL) y superior izquierdo (o bajo-alto=LH) contienen observaciones con asociación espacial negativa (estas observaciones tienen poca similitud con sus vecinas). La Fig. 4.13B muestra la distribución de valores generados al realizar permutaciones aleatorias en la ubicación de los datos en el espacio y calcular su valor I_{Moran}. Después de 999 permutaciones, se obtiene una distribución de referencia (valor medio en azul) y se compara con el valor I_{Moran} observado (en rojo), que se encuentra en el extremo positivo de esta distribución. La Fig. 4.13C muestra cada polígono en el mapa (radios censales) coloreado según su ubicación en el cuadrante correspondiente en el diagrama de dispersión de Moran, indicando qué unidades tienen autocorrelación estadísticamente significativa entre estas variables. El patrón espacial emergente confirma la dependencia espacial entre los casos de bronquiolitis y el hacinamiento.

La migración, como fenómeno biológico y social complejo que involucra a individuos (migrantes y sus familias) y sociedades, es particularmente relevante en este contexto local. Entre 1970 y 2010, Puerto Madryn multiplicó por trece su población, de 6,100 a casi 80,000 habitantes. Este dramático crecimiento poblacional ha cambiado la estructura demográfica y generado varios problemas colaterales, como la concentración de la pobreza urbana y la reestructuración desigual de los servicios públicos, entre otros aspectos socioeconómicos y ambientales (Kaminker, 2015). Este proceso tiene muchas dimensiones

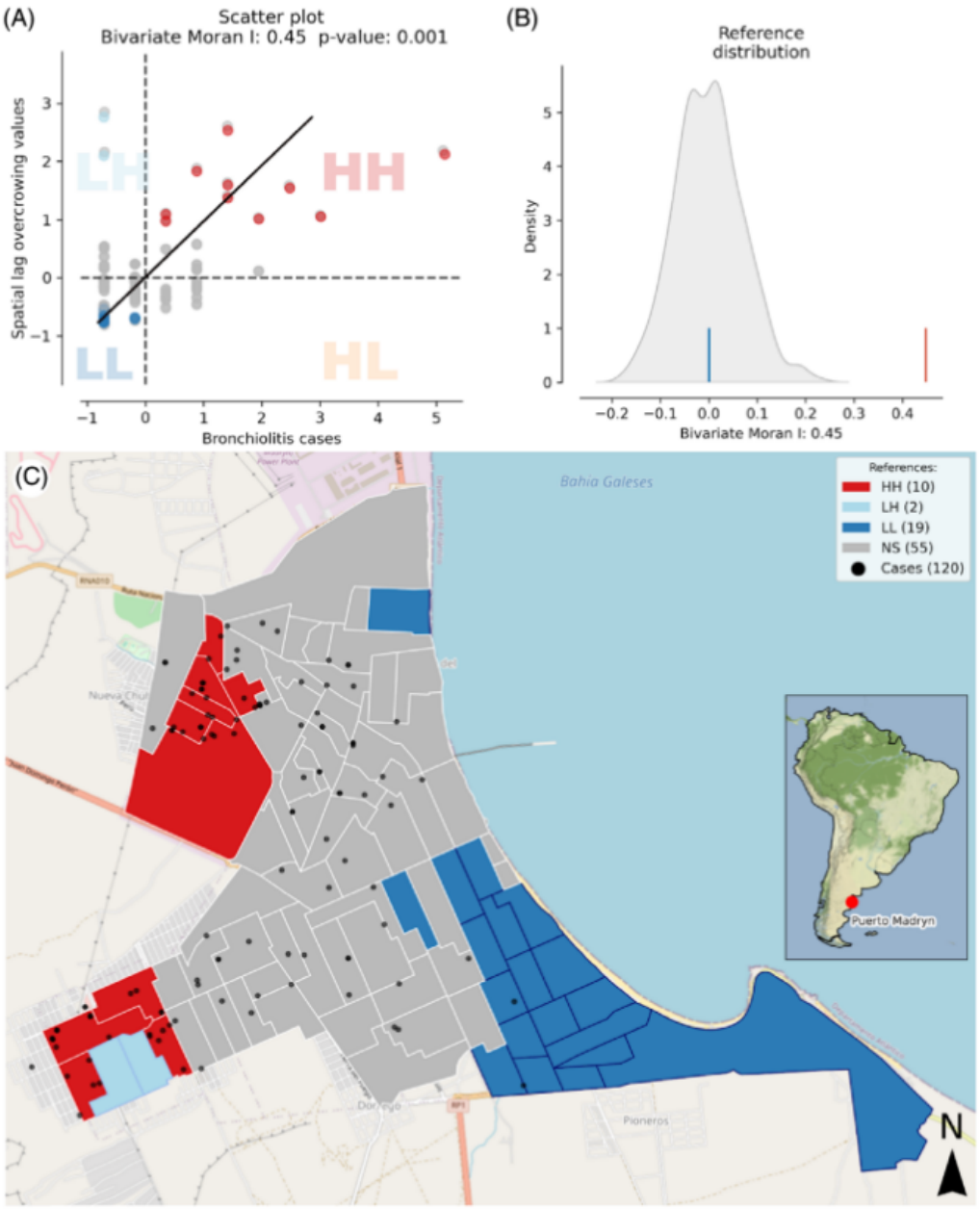


Figura 4.13: Producto del *pipeline Bronquiolitis* con el resultados del análisis de Moran bivariante. En (A) se muestra el diagrama de dispersión de Moran bivariante para casos de bronquiolitis y hacinamiento, mostrando cuatro posibles estados: en rojo, valores altos de desfase espacial rodeados de otros valores altos (HH), en azul, valores bajos rodeados de valores bajos (LL) y las valores atípicos espaciales, HL (naranja), LH (azul claro) y no significativo o NS (gris). En (B) se muestra la distribución de referencia para la I_{Moran} bivariante con su valor medio en azul y el valor de I_{Moran} observado en rojo (0,45). En (C) se presenta el mapa de BILISAs. Los puntos negros representan casos de bronquiolitis (número total entre paréntesis). Los rectángulos coloreados representan polígonos correspondientes a los estados y claves de color mencionados en (A) (entre paréntesis, el número de polígonos o radios censales que entran en cada una de estas categorías).

e implicaciones, incluido el aumento del hacinamiento. Análisis demográficos anteriores (Kaminker, 2016) han demostrado que la práctica de expansión residencial de Puerto Madryn, llamada urbanización acelerada, ha segregado estructuralmente a los hogares de bajos ingresos. Según el Instituto Nacional de Estadística y Censos, con base en información recopilada en 2010, el 9,42 % de la población total de Puerto Madryn vivía en hogares con una o más NBI (INDEC, 2015). La planificación urbana que se ha llevado a cabo en la ciudad no ha tenido en cuenta cómo la gestión de los recursos ha contribuido a la producción y reproducción de desigualdades.

La producción de mapas de vulnerabilidad a nivel de distrito utilizando registros nacionales disponibles, basados en encuestas de población y vivienda, es una fuente útil de información que se puede cotejar con datos epidemiológicos georreferenciados para facilitar la visualización, identificación y análisis de áreas prioritarias. El gradiente social en salud y enfermedad que existe dentro y entre los países es un desafío global (Marmot, 2005). Este gradiente social involucra dimensiones de justicia, moral y ética y está determinado en gran medida por desigualdades en salud debido a diferencias sistemáticas e innecesarias (Braveman, 2019). Nuestra estrategia fue un análisis retrospectivo que combinó información a nivel poblacional con datos a nivel individual. Hasta donde sabemos, este es el primer estudio en la Patagonia argentina que combina registros hospitalarios, ubicación del hogar de los pacientes, bienestar del hogar y acceso a las necesidades básicas. Dadas las características de los centros de salud locales, las tasas de natalidad conocidas, la estacionalidad del Virus Sincitial Respiratorio y las características climáticas de la estación fría en la región patagónica, estudiar este conjunto de variables (cuyo comportamiento anual es esperado y conocido) combinado con herramientas SIG son útiles para describir conglomerados de casos y condiciones concomitantes dentro de la ciudad. Es importante realizar estudios locales porque las comunidades siempre son únicas en sus procesos de salud y enfermedad. Los hallazgos actuales pueden tener implicaciones importantes para el desarrollo y la implementación de intervenciones de salud más efectivas. El hacinamiento tiene un impacto mensurable y significativo en la incidencia de hospitalización de lactantes con bronquiolitis. El mapeo espacial de enfermedades y sus factores de riesgo es una herramienta prometedora para mejorar nuestra capacidad de comprender la compleja relación entre la salud humana y los factores socioeconómicos y ambientales. Esto permitirá

determinar si existen otros déficits estructurales que afecten a los grupos de población más desfavorecidos (que además son los más dependientes del sistema sanitario público).

Capítulo 5

Conclusiones

Hemos estudiado y desarrollado métodos computacionales eficientes que han diversificado el estudio de la estructura poblacional en Argentina. Utilizamos registros de apellidos, que son accesibles y practicamente de costo nulo y los combinamos con otras bases de fuentes muy diversas. La informática puesta a disposición del método isonímico, nos permitió explorar la subestructuración poblacional de nuestro país en términos de consanguinidad, parentesco y migración. Combinando esta información con datos sociodemográficos, médico-sanitarios y económicos, identificamos posibles aislados poblacionales y estudiamos la dinámica de ocupación del espacio y sus factores asociados.

A lo largo de los capítulos que componen esta tesis se presentaron diferentes proyectos desarrollados para el procesamiento automático de grandes volúmenes de datos de muy diversa naturaleza, en cuyas tareas se incluyen aquellas encargadas de la georreferenciación, cálculo y análisis de diferentes estadísticos y la elaboración de visualizaciones efectivas. Estos procesos tienen la ventaja contener tareas modulares, de fácil mantenimiento y potencialmente paralelizables, en donde el producto de cada una puede ser un insumo para otro análisis posterior. Así mismo, en etapas exploratorias, estos productos sirven para encauzar y agilizar las pruebas, tanto las que podrían haber estado diseñadas en un etapa preliminar como las que se generan como consecuencia de estudiar los patrones emergentes en las representaciones visuales parciales que son resultado de tareas de datos previas. El código fuente de todas las herramientas informáticas desarrolladas se encuentra accesible a través de repositorios de código abierto. Además, se comparten bases de datos de información isonímica argentina, actualizadas para los períodos estudiados y que abarcan prácticamente la totalidad de la población y el territorio. Aquellas bases que han sido combinadas y adaptadas a partir de bases públicas también se dejan a disposición

de aquellas personas interesadas. Los resultados generados tienen un gran potencial de aplicación, en particular en el sector salud, ya que permiten la identificación de regiones geográficas cuyas características histórico-demográficas los hacen susceptibles de mayores prevalencias de enfermedades autosómicas recesivas o en riesgo para malformaciones congénitas. Para estas regiones pueden proponerse programas de implementación y divulgación de información en Salud Pública, tendientes a una asesoría sanitaria, detección y atendimiento temprano. También son útiles en estudios de casos de seguimiento de patologías sensibles.

Los métodos aplicados sobre bases de datos de salud que han sido empleados en este estudio, han evidenciado una notable flexibilidad. Por ejemplo, se ha constatado su capacidad para adaptarse y analizar problemáticas distintas, como la prevalencia y la tendencia a lo largo del tiempo de los Defectos de Cierre del Tubo Neural (DTN). Esto se llevó a cabo mediante el examen de los registros de muertes fetales en Argentina entre los años 1994 y 2019, proporcionados por la Dirección de Estadísticas e Información de la Salud (DEIS). Muchos de las DTN se pueden prevenir mediante la ingesta periconcepcional de ácido fólico, que reduce la prevalencia de los defectos de cierre entre un 50 % y un 70 % dependiendo de la población. En Argentina, la Ley Nro 25630, promulgada en 2002 y regulada en 2003, establece la fortificación obligatoria de la harina de trigo comercializada localmente con hierro, ácido fólico y vitaminas. Para conocer la prevalencia de muertes fetales relacionadas con anencefalia y mielomeningocele, se calcularon los porcentajes de DTN en relación con las malformaciones congénitas y las muertes. Asimismo, se determinó la tasa de DTN (definida como el número de muertes fetales por DTN por cada 10,000 nacidos vivos) para toda Argentina, independientemente del sexo resgistrado en el certificado de defunción. El análisis de la tendencia secular se llevó a cabo mediante un modelo de Poisson. Los resultados obtenidos revelaron que la proporción de DTN con respecto a las muertes fetales fue del 1,32 en general. En 1994, las DTN representaban el 34,7 % de las muertes fetales debidas a malformaciones congénitas y el 1,7 % del total de muertes fetales. En contraste, en 2019, estos porcentajes disminuyeron significativamente a un 9,4 % y un 0,5 %, respectivamente. Además, se observó una tendencia secular negativa en las DTN ($p < 0{,}05$). El riesgo de muerte fetal por anencefalia y mielomeningocele experimentó una disminución del 67 % y 51 %, respectivamente ($p < 0{,}05$), entre 2005 y

2019 en comparación con el periodo anterior a la fortificación de la harina de trigo. Los hallazgos indicaron una reducción significativa en el riesgo de muertes fetales asociadas con DTN, especialmente anencefalia, en Argentina a lo largo del período de estudio, con la mayor disminución observada durante la implementación obligatoria de la fortificación de la harina. En consonancia con el segundo caso de estudio del Capitulo 4 de esta tesis, se destaca la importancia de incluir las muertes fetales en la vigilancia de los DTN, ya sea de manera independiente o en conjunto con otros resultados del embarazo, para monitorear eficazmente el impacto de las medidas preventivas. La significancia de estos último hallazgos enunciados se vio potenciada por la disponibilidad de las herramientas de software especializadas que han sido desarrolladas en esta investigación. Estos artefactos para el análisis en los casos de salud abordados han demostrado gran capacidad de adaptación para la resolución de nuevos problemas.

Este trabajo destaca la contribución del desarrollo ingenieril en nuestra disciplina, específicamente en la creación de software, mediante la producción de diversas herramientas. Entre ellas se incluyen las basadas en proyectos de análisis datos: para el estudio de la corriente migratoria del Volga, para el análisis de migraciones internas en Argentina a través de los apellidos, y para el estudio epidemiológico de enfermedades poco frecuentes, muertes fetales y bronquiolitis. Además, se destaca el desarrollo de un paquete del lenguaje de programación *Python* para el análisis isonímico y una aplicación web diseñada para explorar la estructura demográfica basada en apellidos de la Argentina. En el futuro, planeamos continuar investigando métodos de la ciencia de datos, del aprendizaje automático y la visualización de la información para descubrir patrones de incidencia de enfermedades estacionales y de etiología conocida. También buscaremos continuar el trabajo de identificación de aquellas relaciones que existan entre datos de salud y condiciones de vida, a través de la creación de reportes visuales interactivos efectivos.

Anexos

Anexo A

Publicaciones

A continuación se listan las contribuciones publicadas en revistas científicas y aquellas en proceso de publicación.

Journal of Biosocial Science (2024), page 1 of 14
doi:10.1017/S002193202400018X

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Volga German surnames and Alzheimer's disease in Argentina: an epidemiological perspective

Arturo Leonardo Morales[1,2,3,4†], Marcelo Isidro Figueroa[5,6†], Pablo Navarro[1,2,3,4], Estela Raquel Chaves[7], Anahí Ruderman[1,4], José Edgardo Dipierri[4,5] and Virginia Ramallo[1,4]

[1]Instituto Patagónico de Ciencias Sociales y Humanas (IPCSH), Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas, Puerto Madryn, Argentina, [2]Laboratorio de Ciencias de las Imágenes, Departamento de Ingeniería Eléctrica y Computadoras, Universidad Nacional del Sur, Bahía Blanca, Argentina, [3]Departamento de Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco, Trelew, Argentina, [4]Programa de Referencia y Biobanco Genómico de la Población Argentina (PoblAr), Buenos Aires, Argentina, [5]Unidad de Genética, Hospital Materno Infantil Doctor Hector Quintana, San Salvador de Jujuy, Argentina, [6]Instituto de Ecorregiones Andinas, Universidad Nacional de Jujuy-CONICET, San Salvador de Jujuy, Argentina and [7]Instituto de Biología de la Altura, Universidad Nacional de Jujuy, San Salvador de Jujuy, Argentina
Corresponding author: Virginia Ramallo; Email: ramallo@cenpat-conicet.gob.ar

(Received 12 January 2024; revised 24 February 2024; accepted 4 April 2024)

Abstract
The N141I variant (PSEN1 gene) is associated with familial forms of early-onset Alzheimer's disease (AD) in descendants of Volga Germans, whose migration to Argentina is well documented. As a proxy for geographic origin, surnames can be a valuable tool in population studies. The 2015 Argentine Electoral Registry provided geographic data for 30,530,194 individuals, including 326,922 with Volga German surnames. Between 2005 and 2017, the Ministry of Health recorded 4,115,216 deaths, of which 17,226 were attributed to AD and related causes. The study used both diachronic and synchronic data to identify patterns of territorial distribution and co-spatiality, using Moran's I and generalised linear model statistics. The frequency of surnames of Volga German origin accounts for 43.53% of the variation in deaths from AD and three clusters of high non-random frequency were found. Almost 150 years later, people descending from the Volga migration remain highly concentrated and may have a different risk of developing AD. The identification of spatial patterns provides reliable guidance for medical research and highlights the importance of specific health policies for particular populations.

Keywords: surname analyses; epidemiology; Volga migration-descended; Alzheimer's disease

Introduction

Migration and health

There is a well-established link between population mobility, migration, and the epidemiology of certain diseases (Gushulak and MacPherson, 2006). While communicable diseases have traditionally been the focus of attention, the increasing global importance of migration has led to a renewed interest in other aspects of population health, in particular non-communicable diseases that are linked to genetic characteristics.

By combining various analytical methods, available data on Alzheimer's disease (AD) can be used to identify determinants of health disparities in AD and their impact at the individual,

[†]Arturo Leonardo Morales and Marcelo Isidro Figueroa contributed equally to this manuscript.

community, and societal levels (Akushevich *et al.*, 2023). The determinants may include ethnic, gender, and geographic factors.

Alzheimer's disease and Volga Germans

AD can be either familial or sporadic. The familial form is autosomal dominant and has an early onset, occurring in individuals under 65 years of age. It accounts for only 1–5% of cases and is caused by mutations in the genes PSEN1 (presenilin 1) (MIM *104311) and PSEN2 (presenilin 2) (MIM *600759). PSEN1 is associated with Alzheimer's disease type 3 (MIM 607822) and PSEN2 with AD type 4 (MIM 606889). Late-onset AD is classified as occurring in patients over the age of 65. The sporadic form, which accounts for 95% of cases, has no known genetic cause (Andrade-Guerrero *et al.*, 2023).

In 1988, Bird *et al.* reported on presenile AD in five families in the United States. The disease was confirmed by autopsy and inherited as an autosomal dominant trait, present in both males and females for several generations (Bird *et al.*, 1988). All five families were descended from immigrants known as Volga Germans who arrived in the United States between 1870 and 1920. In 1992, the authors studied 28 families of Volga German descendants with AD. Eighteen families originated from the villages of Frank and Walter, near Saratov, Russia. The authors suggested that the cause of AD may be a founder effect. However, a common affected ancestor could not be identified (Bird *et al.*, 1992).

The diaspora of Volga Germans

To understand the context of the migration of Germans to Russia, it is necessary to consider the Seven Years' War. This international conflict over the control of colonies in North America and India began in 1756 and ended in 1763. It involved several empires, kingdoms, and other political structures established in territories that are now nation-states, including Germany. Catherine II, the Empress of the Russian Empire, aimed to populate the border regions near Asia with settled European peasants. In 1763, she promoted a series of benefits for migrants to make the colonisation of the Volga attractive to the post-war weakened German population. These benefits included freedom of religion, temporary tax exemption, interest-free loans, internal self-government, and permanent exemption from military conscription. Between 1764 and 1769, settlers primarily from Hesse, a state in Germany with its current capital in Wiesbaden, arrived on the lower Volga near the city of Saratov. During this time, 104 colonies were established, with a total population of 22,246 inhabitants (Pohl, 2009).

In 1864, Tsar Alexander II made amendments to the original agreement. By 1874, German immigrants were required to register for military service. This event marked the beginning of the diaspora, particularly in the United States, Canada, Australia, and Brazil (Pohl, 2009). In 1878, 1,100 Volga Germans arrived in Argentina. They settled in the provinces of Entre Ríos, Buenos Aires, Santa Fe, Chaco, and Córdoba, where numerous rural colonies were founded that still exist today. Immigration to Argentina continued until the First World War AufderHeide, 2006. Estimates from descendants' associations suggest that there are 2,000,000 Volga Germans in the country. Individual migration stories are well documented and accessible through the websites of the descendant societies and/or their social networks, including Alemanes del Wolga en Argentina (http://www.alemanesdelwolga.com.ar), Centro Cultural Argentino Wolgadeutsche, Federación Argentina de Descendientes de Alemanes del Volga (http://fadadav.org.ar), and Asociación Argentina de Descendientes de Alemanes del Volga Unser Licht.

This database contains family registers, a list of surnames, and information on the colonies in which they settled.

https://doi.org/10.1017/S002193202400018X Published online by Cambridge University Press

Epidemiology of Alzheimer's disease in Argentina

Research on AD has been concentrated on particular geographic regions and groups, as evidenced by studies conducted by Larraya *et al.* (2004), Melcon *et al.* (2010), Méndez *et al.* (2018), Itzcovich *et al.* (2020), García and Comesaña (2021), and Dalmasso *et al.* (2023). Currently, there is no national institutional registry of individuals diagnosed with AD. However, death certificates can serve as a valuable source of data. The Department of Health Statistics and Information (DEIS) is responsible for preparing periodic statistical reports that include an official registry of all deaths, their immediate causes, and any associated or pre-existing conditions. The International Classification of Diseases (ICD-10) is used to calculate morbidity and mortality statistics. The quality of this information is validated and serves as a critical input, although there may be variations between years or regions.

The aim of this study is to contribute to the epidemiology of AD by analysing the spatial distribution of all deaths related to AD, the spatial distribution of surnames of Volga German origin, and the association between these two phenomena in the country.

Materials and methods

Data sources

Argentina, with a population of 46,654,581 according to the National Institute of Statistics and Censuses (INDEC, 2023), is the second largest country in South America. It is divided into 23 provinces, one federal district, and 529 minor subdivisions known as departments. The provinces are further grouped into five geographical regions (refer to Fig. 1). Northwest (comprising the provinces of Catamarca, Jujuy, La Rioja, Salta, Santiago del Estero, and Tucumán), Northeast (Corrientes, Chaco, Formosa, and Misiones), Cuyo (Mendoza, San Juan, and San Luis), Central or Pampean (Buenos Aires, Córdoba, Entre Ríos, La Pampa, Santa Fe, and the Autonomous City of Buenos Aires), and Patagonia (Río Negro, Neuquén, Chubut, Santa Cruz, and Tierra del Fuego). The Central/Pampean region has the highest population density and income levels. In the 19th and 20th centuries, it was the most common destination for transcontinental migrations.

Since 2012, all Argentine citizens over the age of 16 have the right to vote, and the electoral roll is compiled annually. For this study, the 2015 Electoral Registry and national death records from 2005 to 2017 were used, as they share the same territorial data organisation. We analysed publicly available anonymous information. According to Argentine legislation, ethical approval was not required in these cases.

The Surnames of the Volga Germans

Surnames are sociocultural variables that result from historical and cultural processes. They are an important resource in bioanthropology and human population genetics. The linguistic and/or geographic origin of surnames can serve as a proxy for ethnicity in demographic and spatial analyses (Mateos, 2014; Albeck *et al.*, 2017). A list of surnames was compiled using the information available on the websites of the aforementioned associations of Volga German descendants. The study compared the 50 most common Volga surnames (VSs) in Argentina to the 50 most common surnames in the State of Hesse, Germany, as provided by Forebears (https://forebears.io/germany/hesse#surnames), and the 50 most common German surnames (excluding the State of Hesse) based on telephone user records of 30,000,000 individuals used by Rodriguez-Larralde *et al.* (1978). To ensure accuracy, a conservative approach was adopted for the comparative analysis, due to the potential for incorrect surname recording during migration in the 19th century. The threshold for homonymy was set at the difference of a diacritical mark or accent, such as a dieresis.

https://doi.org/10.1017/S002193202400018X Published online by Cambridge University Press

Figure 1. Map of Argentina's major administrative divisions and regions.

Using the Bulsarapp application (Morales *et al.*, 2021), surnames of Volga origin found in the 2015 electoral register were georeferenced. The frequency of these surnames was calculated by determining the number of individuals with VSs per 1000 voters (VS*1000) at the departmental, provincial, and regional levels.

Deaths due to Alzheimer's disease

Although the death certificate is an official document signed by a medical professional, it may contain inaccuracies if data on the general state of health prior to death is not well known or would require forensic examinations to be conclusive. Therefore, we have used a broad criterion to select the following ICD 10 codes: G30.0 (AD with early onset, usually before the age of 65), G30.1 (AD with late onset, usually after the age of 65), G30.8 (other AD), G30.9 (Alzheimer disease, unspecified), and G31 (other degenerative diseases of nervous system, not elsewhere classified).

https://doi.org/10.1017/S002193202400018X Published online by Cambridge University Press

The DEIS (2023) provided the time series of the number of deaths reported in Argentina from 2005 to 2017. Specific death rates (SDRs) related to AD were calculated per 1000 deaths (AD*1000) at departmental, provincial, and regional levels for the entire period.

Statistical and spatial analysis

To assess the impact of VS on SDRs, we used a generalised linear model with mixed effects. We employed three distribution types: Poisson (Mod. Poiss), Negative Binomial (Mod. BN), and zero-inflated Poisson (Mod. ZIP). We diagnosed the goodness of fit using the Q-Q statistic and assessed the model's relative quality using the Akaike information criterion (AIC). We considered the political division (departments within provinces) as a random term. The study calculated the incidence rate ratio by exponentiating the regression beta. Furthermore, the Nagelkerke coefficient of determination was used to assess the proportion of the model's variance explained.

The study utilised the Global Moran Index and Local Indicator of Spatial Autocorrelation to analyse the spatial distribution of VSs and SDRs, as well as the combination of both variables. Spatial autocorrelation is a statistical method used to determine whether a group of entities, such as regions, provinces, and departments, and their attributes, such as VS and SDR values, exhibit clustered, sparse, or random patterns across a territory. The analysis can aid in identifying spatial relationships and patterns that may not be immediately apparent (Anselin, 1995). Significance was determined at the 0.05 confidence level using the Monte Carlo test (999 permutations) under the null hypothesis of no spatial association. The analyses were conducted using GeoDa 1.14 software and the GeoPandas and PySAL function libraries for the Python programming language.

Results

Table 1 compares the 50 most common VSs in Argentina, Hesse (HS), and Germany (GS). Argentina and HS share 24% of the same surnames, while only 14% of the family names are common to all three lists.

According to the National Institute of Statistics and Censuses (INDEC, 2023), Argentina's total population in 2015 was 43,131,966. The electoral roll for that year had 30,530,194 registered voters with 373,709 different surnames, representing over 70% of the country's population. Out of these, 326,922 individuals (1.22%) had a VS, which accounts for a total of 1,109 surnames with this origin.

Table 2 displays the geographic distribution of VS carriers. The province of La Pampa, located in the Central region, had the highest frequency (VS*1000), while San Juan (Cuyo region) had the lowest. Additionally, Table 2 summarises death certificate data, which shows that between 2005 and 2017, there were 4,115,216 recorded deaths in Argentina, with 17,226 (4.19%) attributed to or associated with AD. At the provincial level, La Pampa had the highest SDR, while the province of Santa Cruz had the lowest. Of all deaths related to Alzheimer's, 68% were women and 31% were men. In the remaining cases, the biological sex of the deceased could not be determined. The most common code in the database was G30.9 (AD, unspecified), appearing in 85% of certificates. For death certificates with code G30.0 (early-onset AD), the mean age was 73.82 for women (SD 14.89) and 68.23 for men (SD 12.01).

All spatial analyses indicate positive autocorrelation with p-values below 0.001. Figure 2 displays choropleth maps that clearly visualise the variability in VS and SDR by department. Several departments in the Central-Pampean region have the highest concentration of voters with surnames of Volga origin (see Fig. 2a). The maps show notably low SDR values in the departments of the NWA, NEA, and southern regions of the country (Fig. 2b). After calculating the bivariate Moran's I for the two data sources (VS and SDR), three clusters with high and non-random values (Moran's I = 0.19) were identified. Hot spots are located in the Central/Pampean region and in northern Patagonia, with high specific mortality rates related to AD and a high frequency of VSs.

https://doi.org/10.1017/S002193202400018X Published online by Cambridge University Press

Table 1. Comparison between the 50 most frequent Volga surnames in Argentina, the State of Hesse (HS), and Germany (GS). The surnames that are common between Argentina and HS are highlighted in yellow, those common between Argentina and GS are highlighted in red, and the surnames that are common in all databases are shown in grey

Argentina	State of Hesse	Germany
MULLER	MÜLLER	HANSEN
COLMAN	SCHMIDT	KELLER
SCHMIDT	SCHNEIDER	LANG
SCHNEIDER	SCHÄFER	FRANKE
WAGNER	BECKER	FRANK
FRANK	WEBER	BRANDT
WEBER	WAGNER	JUNG
MAYER	FISCHER	LORENZ
MEYER	SCHMITT	ALBRECHT
FISCHER	HOFMANN	VOGEL
KLOSTER	KOCH	HAHN
BECKER	WOLF	WINKLER
SMITH	KLEIN	FRIEDRICH
WALTER	JUNG	FUCHS
KUHN	HARTMANN	WEISS
ROTH	BAUER	REGNER
KLEIN	HOFFMANN	MAJER
JACOB	RICHTER	MAYER
SCHULZ	WERNER	GÜNTHER
MILLER	MÖLLER	HUBER
SCHWINDT	ROTH	SCHUBERT
BAUER	SCHWARZ	KÖNING
HOFFMANN	BRAUN	KAISER
HANSEN	SCHULZ	SCHMID
KOCH	MEYER	SCHOLZ
KELLER	KÖHLER	HERRMANNN
BRAUN	KAISER	WALTER
GRAFF	NEUMANN	KÖHLER
VOGEL	WALTER	SCHMITT
WEISS	ZIMMERMANN	BRAUN
FRITZ	LANG	LEHMANN
RUPPEL	HAHN	HOFMANN
SCHWAB	JÄGER	SCHULZE
FUHR	KELLER	MEIER

(*Continued*)

https://doi.org/10.1017/S002193202400018X Published online by Cambridge University Press

Table 1. (*Continued*)

Argentina	State of Hesse	Germany
HEIT	FRIEDRICH	SCHMITZ
HUBER	SIMON	MÜLLER
BERGER	FUCHS	PETERS
KESSLER	STEIN	WERNER
REGNER	SCHRÖDER	HARTMANN
BROWN	DIEHL	ZIMMERMANN
MEIER	KRAFT	WOLF
MAIER	SAUER	KRAUSE
ZIMMERMANN	HERRMANN	LANGE
ULRICH	BECK	SCHWARTZ
GETTE	MICHEL	KLEIN
RIEDEL	WINTER	BAUER
SCHEFFER	KÖNIG	KRÜGER
LANG	KRÄMER	SCHÄFER
KLOSTERBECKER	LUDWIG	NEUMANN

Conversely, cold spots, or areas with low values, are located in the Northwest and Northeast regions (Fig. 2c). No significant spatial associations were found in the rest of the country.

The Q-Q and residual plots for Poisson, Negative Binomial, and zero-inflated Poisson models are shown in Fig. 3. The AIC is a measure of the quality of data usage in a model that penalises complexity. A model with a lower AIC is considered better than one with a higher AIC. The Negative Binomial distribution model had the lowest AIC values (Table 3) and suggests that for each unit of variation of VS, the SDR increases by 0.4%. The high frequency of VSs alone explains 43.53% of the observed variability of SDR, according to Nagelkerke R square.

Discussion

Diagnosing AD requires a combination of clinical evaluation, neuroimaging, and biomarkers. However, economic constraints often limit comprehensive studies, leading to underrepresentation or a lack of knowledge about certain mutations' frequency. Molecular studies of genes associated with this disease are infrequent in Argentina. In their study, *Bird et al.* (1988) described families that shared a single N141I mutation in the PSEN2 gene. Although more than 10 additional mutations in PSEN2 have been reported, the N141I mutation has only been found in families of Volga origin, suggesting its specificity to this population group (Blauwendraat *et al.*, 2016). Llibre-Guerra *et al.* (2020) conducted a meta-analysis in Latin America to identify pathogenic variants of autosomal dominant AD, including an investigation into the presence of N141I. Twenty-four variants were detected in 3,583 individuals at risk, mostly of European ancestry and typically attributable to founder effects. The frequency of these variants was higher in Colombia, followed by Puerto Rico and Mexico. A meta-analysis of 47 countries that reported variants in at least one of the three genes APP, PSEN1, and PSEN2 showed that the N141I variant is only found in Argentina, Germany, and the United States (Dehghani *et al.*, 2021).

https://doi.org/10.1017/S002193202400018X Published online by Cambridge University Press

Table 2. Frequency of Volga surnames (VSs) and specific death rates (SDRs) by regions, provinces, and the entire country

Region	Provinces	Voters 2015	Volga surnames (VSs) carriers	VS*1000	Number of Deaths 2005–2017	Deaths related to Alzheimer's diseases	Specific death rates (SDRs)
Central/ Pampean	Buenos Aires	11,384,389	142,590	12.53	1,677,007	5,258	3.13
	Ciudad Autónoma de Buenos Aires	2,541,076	20,750	8.17	422,702	1,129	2.67
	Córdoba	2,645,525	13,766	5.20	360,432	2,148	5.96
	Entre Ríos	979,546	46,768	47.74	128,128	736	5.74
	La Pampa	262,030	14,769	56.36	32,465	338	10.41
	Santa Fe	2,552,338	30,694	12.03	378,409	2,795	7.39
	Total	**20,364,904**	**269,337**	**13.23**	**2,999,143**	**12,404**	**4.14**
Northwest (NWA)	Catamarca	278,151	361	1.30	28,986	82	2.83
	Jujuy	478,463	477	1.00	51,760	101	1.95
	La Rioja	250,537	368	1.47	25,779	45	1.75
	Salta	885,984	1,351	1.52	92,080	271	2.94
	Santiago del Estero	648,777	1,102	1.70	70,362	132	1.88
	Tucumán	1,079,057	1,519	1.41	127,163	515	4.05
	Total	**3,620,969**	**5,178**	**1.43**	**396,130**	**1,146**	**2.89**
Cuyo	Mendoza	1,307,278	2,634	2.01	167,537	1,271	7.59
	San Juan	504,837	487	0.96	60,880	185	3.04
	San Luis	334,603	1,093	3.27	37,218	183	4.92
	Total	**2,146,718**	**4,214**	**1.96**	**265,635**	**1,639**	**6.17**
Northeast (NEA)	Chaco	815,907	8,359	10.25	92,655	268	2.89
	Corrientes	765,271	2,972	3.88	85,889	257	2.99
	Formosa	392,863	2,238	5.70	43,792	268	6.12
	Misiones	787,588	19,489	24.75	82,463	310	3.76
	Total	**2,761,629**	**33,058**	**11.97**	**304,799**	**1,103**	**3.62**

(*Continued*)

Table 2. (*Continued*)

Region	Provinces	Voters 2015	Volga surnames (VSs) carriers	VS*1000	Number of Deaths 2005–2017	Deaths related to Alzheimer's diseases	Specific death rates (SDRs)
Patagonia	Chubut	388,934	3,208	8.25	38,443	194	5.05
	Neuquén	436,081	3,397	7.79	36,851	332	9.01
	Río Negro	474,634	5,979	12.60	50,768	373	7.38
	Santa Cruz	220,278	1,481	6.72	17,353	19	1.09
	Tierra del Fuego	116,042	1,070	9.22	6,094	16	2.63
	Total	**1,635,969**	**15,135**	**9.25**	**149,509**	**934**	**6.25**
Argentina	**Total**	**30,530,189**	**326,922**	**10.71**	**4,115,216**	**17,226**	**4.19**

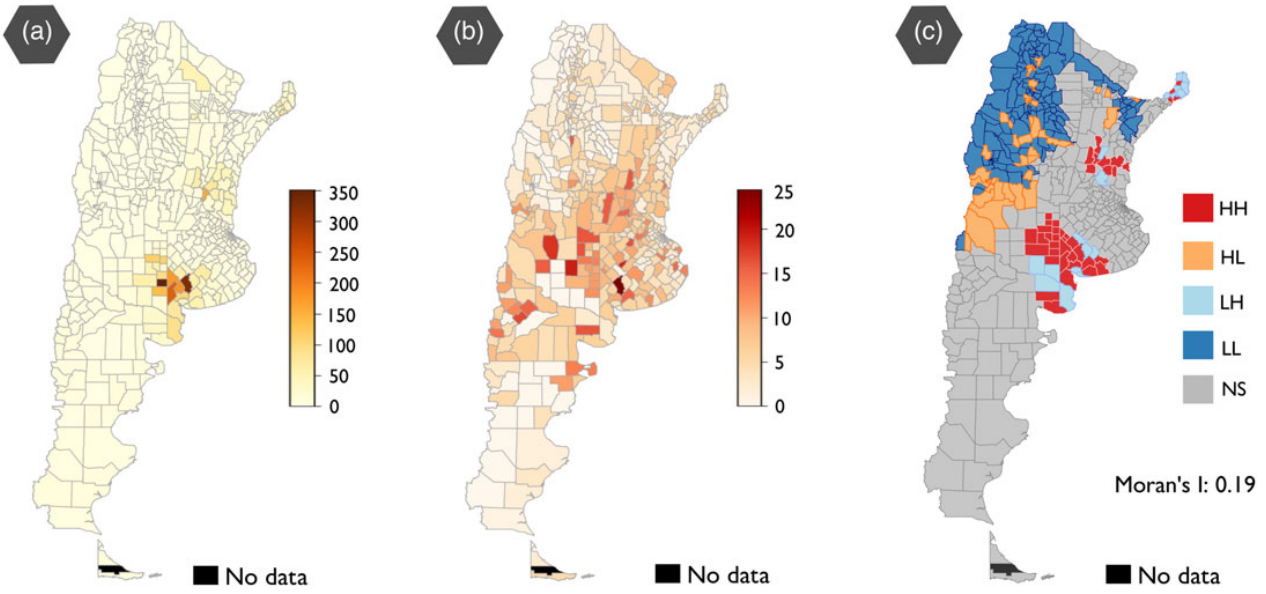


Figure 2. Choropleth map of frequency by departments of Volga surnames (a), and specific death rates related to AD (b). LISA maps illustrate the geographic clustering of both variables (c). Departments with high values that are surrounded by neighbours with high values are coloured in red (HH). High values surrounded by low values are coloured in orange (HL). Departments with low values surrounded by high values are coloured in light blue (LH). Low values clusters or cold spots are coloured in blue (LL). In grey, non-significant (NS).

In 2010, Yu *et al.* identified a haplotype based on six single-nucleotide polymorphisms that cover the PSEN2 gene. The frequency of this haplotype was 0.64 in Volga Germans carrying the N141I mutation, compared to 0.26 in Volga Germans without the mutation and 0.13 in Europeans typed by the Centre d'Etude du Polymorphisme Humain. The study suggests that the N141I mutation in PSEN2 may have occurred before the emigration from the Hesse region to Russia. The discovery of families with this mutation living in Argentina and Germany suggests the possibility of additional cases sharing this common ancestry (Yu *et al.*, 2010; Muchnik *et al.*, 2015). The mutations responsible for familial forms have known biochemical consequences that are likely to be at the root of sporadic AD. Early interventions can delay or even prevent dementia in asymptomatic individuals and families at risk, as well as slow progression in those with symptoms (Bateman *et al.*, 2011).

A study conducted at the beginning of the 21st century in Germany provided estimated values on the prevalence and incidence of dementing illness through large-scale epidemiological and meta-analyses (Bickel, 2000). The prevalence of dementia is higher in the states of Baden-Württemberg, Bavaria, Lower Saxony, North Rhine-Westphalia, and Hesse. This region is also the homeland of the Volga Germans. However, in recent times, there has been a change in this trend, particularly due to the increased frequency of cases of late-onset AD in female patients (Ziegler and Doblhammer, 2009; Lange *et al.*, 2017). These differences are primarily attributed to women's longer life expectancy, as advanced age remains the greatest risk factor for AD (Chêne *et al.*, 2015). Of all the registered deaths due to Alzheimer's in Argentina, 68% were women and 31% were men. A significant difference in the distribution of SDR and VS was observed between the Central/Pampean, Cuyo, and Patagonia regions compared to the NWA region. The NWA region comprises various environments, including the Andean foothills, where populations live at altitudes above 2,500 metres. Genomic studies indicate that the NWA populations have the highest proportion of Central Andean ancestry component (Muzzio *et al.*, 2018; Luisi *et al.*, 2020). Meanwhile, individuals from the province of Misiones (NEA) have the highest proportion of Central/Northern European ancestry. This aligns with the historical record of settlement of Polish, German, Danish, and Swedish colonies in this province (Luisi *et al.*, 2020).

https://doi.org/10.1017/S002193202400018X Published online by Cambridge University Press

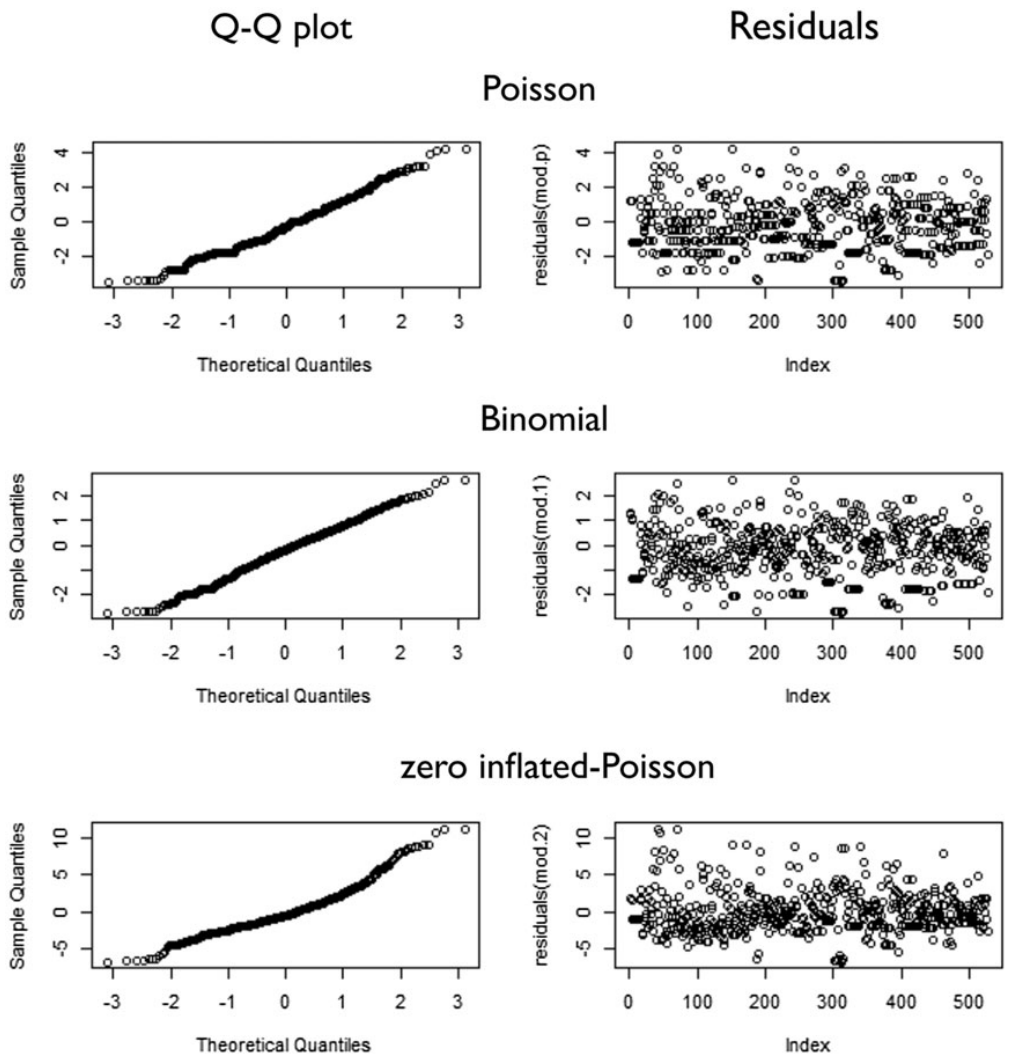


Figure 3. Q-Q and residuals plots for Poisson, Negative Binomial, and zero-inflated Poisson models.

Geographical factors and the biological history of a population can also have an impact on the risk of developing dementia (Alzheimer's Association, 2020). The migration of Volga Germans to Argentina should be distinguished from the migration of other German groups. According to Germanic sources, the massive migration took place between 1883 and 1890. However, the majority of the contingents that left through the Port of Bremen until 1870 went to the United States, Canada, and Brazil. Argentina only became a destination of interest after 1870. German migrants left from ports in Hannover and Bremen, and mainly came from cities in northern Germany, but also including the Duchy of Hesse (de Flachs, 1994). According to the Argentina National Censuses, German migrants represented a minority, comprising no more than 2.35%, 1.70%, and 1.14% of the total population in the 1869, 1895, and 1914 censuses, respectively. The immigrants were mainly concentrated in the same provinces as the Volga Germans, but they also settled in other destinations in Argentina. For example, 2.3% of the German immigrants registered in the aforementioned censuses resided in the NOA region and the majority of them were in the province of Tucumán. Volga migration remained stable in the provinces of Buenos Aires and La Pampa (Central/Pampean region). As shown in Table 3, the highest SDR in the country was recorded in the province of La Pampa (10.41). For this calculation, all deaths were considered according to ICD codes G30.0 (early-onset AD), G30.1 (late-onset AD), G30.8 (other AD), G30.9 (AD, unspecified), and G31 (other degenerative diseases of the nervous system). The rate was also

https://doi.org/10.1017/S002193202400018X Published online by Cambridge University Press

Table 3. Akaike information criterion (AIC) values for each model. Df, degrees of freedom

Model	df	AIC
Poisson (Mod. Poiss)	3	2563,192
Binomial (Mod. BN)	4	2445,948
zero-inflated Poisson (Mod. ZIP)	5	2458,591

calculated specifically for codes G30.0 and G30.9 to better weigh the impact of deaths from early-onset AD and to account for the possibility of under-diagnosis. Once again, the province of La Pampa had the highest SDR (9.61), with 483 deaths in the period analysed.

German surnames display considerable lexical, phonological, and morphological variation, which is reflected in their distribution across different regions (Dräger and Schmuck, 2009). The comparison of the 50 most common surnames shows that the Volga Germans in Argentina are more closely related to the population of Hesse than to any other German state. This makes them excellent markers for analysing large databases. Several dialects are spoken in Hesse. This linguistic diversity is also reflected in the specificity of family names. According to Rodriguez-Larralde *et al.* (1978), who based their research on the Lasker isonymic distance between German cities, this state is part of a cluster in southern Germany. The region is home to speakers of Rhine Franconian (including Hessian), Alemannic, and Bavarian (Konig, 1978). In Russia, the Volga German colonies were linguistic isolates with limited bilingualism and a situation of internal diglossia. Hipperdinger (2017) suggests that this sociolinguistic characteristic led to the replacement of Russian with Spanish in Argentina, a situation that lasted until the second half of the twentieth century. The VSs maintained their distinctiveness in both contexts.

Conclusion

AD is a significant health problem, especially due to the ageing population. Familial forms of AD represent a small percentage of cases but are critical to study. This requires a comprehensive understanding of the biology of the population. The combination of historical and official documents, such as electoral rolls, census data, and health information, can be used to analyse population dynamics. Our study uses a methodological approach that provides a coherent view of the spatial distribution of deaths from AD in Argentina and its relationship to migration processes by combining diachronic (surnames) and synchronous (death certificates) data. Nearly 150 years after the establishment of the first colonies, the population descended from the Volga migration remains highly concentrated in the southeastern departments of La Pampa Province and the southwestern departments of Buenos Aires Province. Within the Central Region, these two areas are contiguous. This demographic behaviour has health consequences. These departments are included in a statistically significant cluster with a high frequency of surnames of Volga origin and high SDRs from AD. Tracing surnames by origin is a cost-effective method for distinguishing structures within a seemingly homogeneous social group. These analyses provide a reliable basis for guiding patient recruitment in medical research and reducing sampling error by identifying where and when a pre-existing genetic pattern is likely to persist. This approach may be useful for describing complex migration scenarios in other Latin American countries undergoing similar population processes.

Acknowledgements. We express our gratitude to Dr. Rolando González-José and the authorities of the Cámara Nacional Electoral Argentina for providing us with access to the 2015 Electoral Register.

Funding statement. This work was supported by Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), grant PIP 2021–2023 number 11220200101902.

https://doi.org/10.1017/S002193202400018X Published online by Cambridge University Press

Competing interests. Authors declare no competing interests.

Ethical standard. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

References

Akushevich I, Kravchenko J, Yashkin A, Doraiswamy PM, Hill CV, and **Alzheimer's Disease and Related Dementia Health Disparities Collaborative Group** (2023) Alzheimer's Disease and Related Dementia Health Disparities Collaborative Group. Expanding the scope of health disparities research in Alzheimer's disease and related dementias: Recommendations from the "Leveraging Existing Data and Analytic Methods for Health Disparities Research Related to Aging and Alzheimer's Disease and Related Dementias" Workshop Series. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* **15**(1), e12415. https://doi.org/10.1002/dad2.12415

Albeck ME, Alfaro EL, Dipierri JE, and Chaves ER (2017) Los apellidos de Salta en el siglo XXI: origen geo-lingüítico, diversidad y frecuencia. *Andes* **28**(2), 2–21. https://www.redalyc.org/articulo.oa?id = 12755958008.

Alzheimer's Association (2020) 2020 Alzheimer's disease facts and figures. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* **16**(3), 391–460. https://doi.org/10.1002/alz.12068

Andrade-Guerrero J, Santiago-Balmaseda A, Jeronimo-Aguilar P, Vargas-Rodríguez I, Cadena-Suárez AR, Sánchez-Garibay C, Pozo-Molina G, Méndez-Catalá CF, Cardenas-Aguayo MD, Diaz-Cintra S, Pacheco-Herrero M, Luna-Muñoz J, and Soto-Rojas LO (2023) Alzheimer's disease: an updated overview of its genetics. *International Journal of Molecular Sciences* **24**(4), 3754. https://doi.org/10.3390/ijms24043754

Anselin L (1995) Local indicators of spatial association—LISA. *Geographical Analysis* **27**, 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

AufderHeide E (2006) Representations of German-Speaking Exiles and Immigrants in Argentina [Bowling Green State University/OhioLINK]. URL: http://rave.ohiolink.edu/etdc/view?acc_num = bgsu1162858784

Bateman RJ, Aisen PS, De Strooper B, Fox NC, Lemere CA, Ringman JM, Salloway S, Sperling RA, Windisch M, and Xiong C (2011) Autosomal-dominant Alzheimer's disease: a review and proposal for the prevention of Alzheimer's disease. *Alzheimer's Research & Therapy* **3**(1), 1. https://doi.org/10.1186/alzrt59

Bickel H (2000) Dementia syndrome and Alzheimer disease: an assessment of morbidity and annual incidence in Germany. *Gesundheitswesen* **62**(4), 211–218. https://doi.org/10.1055/s-2000-10858

Bird TD, Lampe TH, Nemens EJ, Miner GW, Sumi SM, and Schellenberg GD (1988) Familial Alzheimer's disease in American descendants of the Volga Germans: probable genetic founder effect. *Annals of Neurology* **23**(1), 25–31. https://doi.org/10.1002/ana.410230106

Bird TD, Nemens EM, Nochlin D, Sumi SM, Wijsman EM, and Schellenberg GD (1992) Familial Alzheimer's disease in Germans from Russia: a model of genetic heterogeneity in Alzheimer's disease. Heterogeneity of Alzheimer's Disease. In *Research and Perspectives in Alzheimer's Disease*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-46776-9_13

Blauwendraat C, Wilke C, Jansen IE, Schulte C, Simón-Sánchez J, Metzger FG, Bender B, Gasser T, Maetzler W, Rizzu P, Heutink P, and Synofzik M (2016) Pilot whole-exome sequencing of a German early-onset Alzheimer's disease cohort reveals a substantial frequency of PSEN2 variants. *Neurobiology of Aging*, **37**, 208.e11–208.e17. https://doi.org/10.1016/j.neurobiolaging.2015.09.016

Chêne G, Beiser A, Au R, Preis SR, Wolf PA, Dufouil C, and Seshadri S (2015) Gender and incidence of dementia in the Framingham Heart Study from mid-adult life. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* **11**(3), 310–320. https://doi.org/10.1016/j.jalz.2013.10.005

Dalmasso MC, de Rojas I, Olivar N, Muchnik C, Angel B, Gloger S, Sanchez Abalos MS, Chacón MV, Aránguiz R, Orellana P, Cuesta C, Galeano P, Campanelli L, Novack GV, Martinez LE, Medel N, Lisso J, Sevillano Z, Irureta N, Castaño EM, Montrreal L, Thoenes M, Hanses C, Heilmann-Heimbach S, Kairiyama C, Mintz I, Villella I, Rueda F, Romero A, Wukitsevits N, Quiroga I, Gona C, Lambert JC, Solis P, Politis DG, Mangone CA, Gonzalez-Billault C, Boada M, Tàrraga L, Slachevsky A, Albala C, Fuentes P, Kochen S, Brusco LI, Ruiz A, Morelli L, and Ramírez A (2023) The first genome-wide association study in the Argentinian and Chilean populations identifies shared genetics with Europeans in Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*. 10.1002/alz.13522. Advance online publication. https://doi.org/10.1002/alz.13522

de Flachs MCV (1994) Emigraciones transoceánicas. Los alemanes en América. 1850-1914. El caso argentino. *Cuadernos de Historia Contemporánea* **16**, 65.

Dehghani N, Bras J, and Guerreiro R (2021) Erratum to: How understudied populations have contributed to our understanding of Alzheimer's disease genetics. *Brain* **144**(8), e69. https://doi.org/10.1093/brain/awab028

Dirección de Estadísticas e Información de la Salud – DEIS (2023) Bases de datos de defunciones. URL: https://www.argentina.gob.ar/salud/deis/datos/defunciones (accessed on 02nd April 2023).

https://doi.org/10.1017/S002193202400018X Published online by Cambridge University Press

Dräger K, and Schmuck M (2009) The German Surname Atlas Project—Computer-based surname geography, Mainz. Available from: https://www.germanistik.uni-mainz.de/files/2015/01/Dr%C3%A4ger-Schmuck2009.pdf.

García MJ, and Comesaña A (2021) Prevalence of neurocognitive disorders in a rural area of Argentina. *Revista de la Facultad de Ciencias Medicas* **78**(4), 347–352.

Gushulak BD, and MacPherson DW (2006) The basic principles of migration health: population mobility and gaps in disease prevalence. *Emerging Themes in Epidemiology* **3**, 3. https://doi.org/10.1186/1742-7622-3-3

Hipperdinger YH (2017) Las lenguas inmigratorias en la Argentina: El caso de los alemanes del Volga. *Sociedad y Discurso* **30**, 92–114. https://doi.org/10.5278/ojs.s%20&%20d.v0i30.1851

Instituto Nacional de Estadísticas y Censos de la República Argentina – INDEC (2023) Estadísticas de población por año y región. URL: https://www.indec.gob.ar/indec/web/Nivel3-Tema-2-41 (accessed on 15th November 2023).

Itzcovich T, Chrem-Méndez P, Vázquez S, Barbieri-Kennedy M, Niikado M, Martinetto H, Allegri R, Sevlever G, and Surace EI (2020) A novel mutation in PSEN1 (p.T119I) in an Argentine family with early- and late-onset Alzheimer's disease. *Neurobiology of Aging* **85**, 155.e9–155.e12. https://doi.org/10.1016/j.neurobiolaging.2019.05.001

Konig W (1978) *dtv-Atlas zur deutschen Sprache* (First Edition). Deutscher Taschenbuch Verlag, Munich.

Lange L, Schulte T, Dittmann B, and Hildebrandt H (2017) 2 Regionale Verteilung der Demenz sowie Inanspruchnahme vor und nach Erstdiagnose. *Monitor Versorgungsforschung* 41–46. https://doi.org/10.24945/MVF.05.18.1866-0533.2098

Larraya FP, Grasso L, and Marí G (2004) Prevalence of dementia of the Alzheimre type, vascular dementia and other DSM-IV and ICD-10 dementias in the Republica Argentina. *Revista Neurologica Argentina* **29**, 148–153.

Llibre-Guerra JJ, Li Y, Allegri RF, Mendez PC, Surace EI, Llibre-Rodriguez JJ, Sosa AL, Aláez-Verson C, Longoria EM, Tellez A, Carrillo-Sánchez K, Flores-Lagunes LL, Sánchez V, Takada LT, Nitrini R, Ferreira-Frota NA, Benevides-Lima J, Lopera F, Ramírez L, Jiménez-Velázquez I, Schenk C, Acosta D, Behrens MI, Doering M, Ziegemeier E, Morris JC, McDade E, and Bateman RJ (2020) Dominantly inherited Alzheimer's disease in Latin America: genetic heterogeneity and clinical phenotypes. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* **17**(4), 653–664. https://doi.org/10.1002/alz.12227

Luisi P, García A, Berros JM, Motti JMB, Demarchi DA, Alfaro E, Aquilano E, Argüelles C, Avena S, Bailliet G, Beltramo J, Bravi CM, Cuello M, Dejean C, Dipierri JE, Jurado Medina LS, Lanata JL, Muzzio M, Parolin ML, Pauro M, Paz Sepúlveda PB, Rodríguez Golpe D, Santos MR, Schwab M, Silvero N, Zubrzycki J, Ramallo V, and Dopazo H (2020) Fine-scale genomic analyses of admixed individuals reveal unrecognized genetic ancestry components in Argentina. *PloS One* **15**(7), e0233808. https://doi.org/10.1371/journal.pone.0233808

Mateos P (2014) *Names, Ethnicity and Populations. Series: Advances in Spatial Science*. Berlin and Heidelberg, Germany: Springer. ISBN: 978-3-642-45412-7

Melcon CM, Bartoloni L, Katz M, Del Mónaco R, Mangone CA, Melcon MO, and Allegri RF (2010) Propuesta de un Registro centralizado de casos con Deterioro Cognitivo en Argentina (ReDeCAr) basado en el Sistema Nacional de Vigilancia Epidemiológica. *Neurología Argentina* **2**(3), 161–166.

Méndez PC, Calandri I, Nahas F, Russo MJ, Demey I, Martín ME, Clarens MF, Harris P, Tapajoz F, Campos J, Surace EI, Martinetto H, Ventrice F, Cohen G, Vázquez S, Romero C, Guinjoan S, Allegri RF, and Sevlever G (2018) Argentina-Alzheimer's disease neuroimaging initiative (Arg-ADNI): neuropsychological evolution profile after one-year follow up. *Arquivos de neuro-psiquiatria* **76**(4), 231–240. https://doi.org/10.1590/0004-282x20180025

Morales AL, Navarro P, Cintas C, González-José R, Ramallo V, and Delrieux C (2021) Bulsarapp: interactive visual analysis for surname trend exploration. *IEEE Computer Graphics and Applications* **42**, 28–33.

Muchnik C, Olivar N, Dalmasso MC, Azurmendi PJ, Liberczuk C, Morelli L, and Brusco LI (2015) Identification of PSEN2 mutation p.N141I in Argentine pedigrees with early-onset familial Alzheimer's Disease. *Neurobiology of Aging* **36**(10), 2674–7.e1. https://doi.org/10.1016/j.neurobiolaging.2015.06.011

Muzzio M, Motti JMB, Paz Sepulveda PB, Yee MC, Cooke T, Santos MR, Ramallo V, Alfaro EL, Dipierri JE, Bailliet G, Bravi CM, Bustamante CD, and Kenny EE (2018) Population structure in Argentina. *PLoS One* **13**(5), e0196325. https://doi.org/10.1371/journal.pone.0196325

Pohl JO (2009) Volk auf dem Weg: transnational migration of the Russian-Germans from 1763 to the present day. *Studies in Ethnicity and Nationalism* **9**(2), 267–286.

Rodriguez-Larralde A, Barrai I, Nesti C, Mamolini E, and Scapoli C (1998) Isonymy and isolation by distance in Germany. *Human Biology* **70**(6), 1041–1056.

Yu CE, Marchani E, Nikisch G, Müller U, Nolte D, Hertel A, Wijsman EM, and Bird TD (2010) The N141I mutation in PSEN2: implications for the quintessential case of Alzheimer disease. *Archives of Neurology* **67**(5), 631–633. https://doi.org/10.1001/archneurol.2010.87

Ziegler U, and Doblhammer G (2009) Prävalenz und Inzidenz von Demenz in Deutschland–Eine Studie auf Basis von Daten der gesetzlichen Krankenversicherungen von 2002. *Das Gesundheitswesen* **71**(05), 281–290.

Cite this article: Morales AL, Figueroa MI, Navarro P, Chaves ER, Ruderman A, Dipierri JE, and Ramallo V (2024). Volga German surnames and Alzheimer's disease in Argentina: an epidemiological perspective. *Journal of Biosocial Science*. https://doi.org/10.1017/S002193202400018X

Epidemiology of Fetal Deaths in Argentina (1994-2019): spatial and temporal variation

Anahí Ruderman [1] †
Email: ruderman@cenpat-conicet.gob.ar
Arturo Leonardo Morales [1 2 3] †
Email: lmorales@cenpat-conicet.gob.ar
José Edgardo Dipierri [4]
Email: jedjujuy@gmail.com
Jorge Iván Martínez [4]
Email: jorgemartinez@inbial.unju.edu.ar
Maria Soledad Silva [5]
Email: msolsil824@gmail.com
Soledad De Azevedo [1]
Email: deazevedo@cenpat-conicet.gob.ar
Damián Leonardo Taire [1 6]
Email: dtaire@cenpat-conicet.gob.ar
Virginia Ramallo [1]
Email: ramallo@cenpat-conicet.gob.ar

[1] Instituto Patagónico de Ciencias Sociales y Humanas (IPCSH), Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas, Puerto Madryn, Argentina
[2] Laboratorio de Ciencias de las Imágenes, Departamento de Ingeniería Eléctrica y Computadoras, Universidad Nacional del Sur, Bahía Blanca, Argentina.
[3] Departamento de Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco, Trelew, Argentina.
[4] Instituto de Biología de la Altura, Universidad Nacional de Jujuy – Instituto de Ecorregiones Andinas (Universidad Nacional de Jujuy-CONICET), San Salvador de Jujuy, Argentina.
[5] Servicio de Neonatología, Hospital Zonal “Dr. Andrés R. Isola”, Puerto Madryn, Argentina.
[6] Departamento de Neumonología Pediátrica, Hospital Zonal “Dr. Andrés R. Isola”, Puerto Madryn, Argentina.
† These authors contributed equally to this work

Corresponding author
Virginia Ramallo
Instituto Patagónico de Ciencias Sociales y Humanas (IPCSH), Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas, Puerto Madryn, Argentina.
Email: vramallo@cenpat-conicet.gob.ar

Data availability statement
Data supporting the findings of this study are available from the corresponding author upon request.

Funding statement
No specific grant from any public, commercial or not-for-profit funding agency was received for this research.

Conflict of interest disclosure
All authors declare that they have no conflicts of interest.

Ethics approval statement
Formal approval from the local medical ethics committee was not required due to the retrospective nature of the study.

Permission to reproduce material from other sources
Only open and public databases were used, collected by government agencies under the Ministry of Health and the National Institute of Statistics and Censuses of Argentina.

ABSTRACT
Background: Fetal deaths (FD) remain a public health problem in both developed and developing countries. Global and national information on FD is scarce.
Objective: To analyze the spatial and temporal variation of FD and its causes in Argentina between 1994-2019.
Methods: Data on FD and births were obtained from the Ministry of Health of Argentina. The rate of FD (FDR) and the percentages of causes (classified according to ICD-10) were calculated

geographically at the level of regions and departments and for periods of five years. Joint Point model, Moran index and Local Indicators of Spatial Association (LISA) were used to describe temporal and spatial variation.

Results: There is a negative secular trend in FDR, which is more pronounced and significant in the more developed regions of the country. FD due to complications of the placenta, umbilical cord and membranes tends to decrease as FD due to non-specified causes increases, especially in less developed regions.

Conclusions: A large spatial and temporal heterogeneity of FDRs and their causes was observed. Although measures of proven health effects have been taken at the national level, it is evident that other regional and/or local variables, particularly socio-economic differences as in access to the health system, are associated with the persistence of high FDR values and unspecified causes of FD in the northern region of Argentina.

KEYWORDS

Maternal and child health, secular trend, socio-economic determinants, spatial epidemiology

1 | INTRODUCTION

Different definitions of fetal deaths (FDs) have been proposed, depending on the weight of the fetus or gestational age. These vary from country to country and are not equivalent, making it difficult to have a single internationally accepted standard. The World Health Organization (WHO) recommends that all stillbirths should be counted, but international comparisons only include stillbirths in late pregnancy, weighing ≥1000 g or ≥28 weeks (Blencowe et al., 2021). According to the National Health Ministry's Department of Statistics and Health Information (DEIS, Dirección de Estadísticas e Información de Salud, Ministerio de Salud de la Nación Argentina), fetal death is defined as "that which occurs before the complete expulsion or extraction of the product of conception from the mother's body, regardless of the duration of the pregnancy". Death is characterized by the fetus not breathing or showing other signs of life, such as heartbeat, umbilical cord pulsation or effective movement of voluntary muscles, after such separation (DEIS, 2017). This definition, also adopted by the Pan American Health Organization (PAHO) (2017), is the one that will be followed in this work.

In 2009, the estimated global rate of fetal deaths per thousand live births was 18,9, with 76,2% of FDs occurring in South Asia and sub-Saharan Africa (Cousens et al., 2011). In 2019, 2 million babies were stillborn (a baby who dies after 28 weeks of pregnancy, but before or during birth), with more than three-quarters of these still occurring in sub-Saharan Africa and south Asia (Souza & Bahl, 2022). Inadequate efforts to decrease stillbirths in maternal and child health programs and a lack of precise, inclusive, and applicable information, particularly in high-risk regions, have hindered progress in addressing this issue (Kiguli, Kirunda, Nabaliisa, & Nalwadda, 2021).

The incidence of FD shows large disparities, which are mainly structured around the socio-economic axis. Globally, 98% are produced in low- and middle-income countries (Blencowe et al., 2016). However, high rates are also observed in high-income countries among vulnerable groups and disadvantaged ethnic minorities.

The tragedy that fetal death represents for families is often not addressed in political agendas or government health programs. However, there are some institutional initiatives in international health organizations. The "Global Strategy for Women's, Children's and Adolescents' Health 2016-30" plan (UN, 2015b) emphasizes the importance of FD prevention, considering as a central part of high-quality health care for women and children (Henzell et al., 2016; Flenady et al., 2016). In the same regard, the Lancet Ending Preventable Stillbirths Series Study Group was established in 2011 to help prevent newborn and maternal deaths. However, more action is needed on fetal deaths in terms of advocacy, policy formulation, monitoring and research. In the absence of a consistent and massive protocol of postmortem examinations (such as autopsy and/or placental examination, see Scalise, Cordasco, Sacco, Ricci, & Aquila, 2022), the most important and urgent action is to improve data recording (Frøen et al., 2016).

Based on ICD-10, WHO published the International Classification of Diseases for Perinatal Mortality (ICD-PM) in 2016, which lists the cause of perinatal death using ICD-10 codes, separated by time of death, with maternal conditions contributing to perinatal death (WHO, 2016a). Over time, classification systems have become more complex and the success of their application depends on the availability of detailed clinical information and laboratory investigations, which are not always available in all national databases (Aminu, Bar-Zeev, & van den Broek, 2017). The objective of this study was to analyze the temporal-spatial pattern and causes of FDs that occurred in Argentina from 1994 to 2019.

2 | SUBJECTS & METHODS

Study design and setting

This is a retrospective eco-epidemiological study that collected information on the Argentine population from birth and FD databases between 1994 and 2019. Argentina is a federal nation comprising twenty-three provinces, which are further divided into 525 departments for administrative purposes. In addition, the country has one autonomous city, Buenos Aires. Argentina can be divided into five geographical regions (Figure 1) with similar environmental characteristics: NorthWest (or NWA, includes the provinces of Catamarca, Jujuy, La Rioja, Salta, Santiago del Estero and Tucumán), North-East (NEA, provinces of Misiones, Formosa, Corrientes, Chaco and Entre Ríos), Cuyo (provinces of Mendoza, San Juan and San Luis), Center (provinces of Buenos Aires, Córdoba, La Pampa, Santa Fe and Autonomous City of Buenos Aires) and Patagonia (provinces of Chubut, Neuquén, Río Negro and Santa Cruz). The Center Region has the highest population density, with 65.2% of the country's total population residing there, as per the 2023 demographic estimates released by the National Institute of Statistics and Censuses (INDEC). On the contrary, in relation to the size of its territory, the Patagonia Region has the lowest population density: 3.3 inhabitants per square kilometer.

Figure 1: Administrative division of Argentina and its five geographical regions, including population and total area in square kilometers, as reported by the National Census 2022 (INDEC, 2023).

Cohort

We included all live births and FD cases registered in official records between 1994 and 2019. The data were provided by the DEIS of the National Ministry of Health and were obtained from birth certificates and fetal death certificates. Given the temporal depth of the records, conversion tables were used to standardize the codes in the databases between the transition from ICD-9 to ICD-10 codification (from 1996 onwards). The causes of FD were grouped according to the categories proposed by Hoyert and Gregory (2016): P00) fetus affected by maternal conditions unrelated to the current pregnancy; P01) fetus affected by maternal complications of pregnancy; P02) fetus affected by complications of the placenta, umbilical cord, and membranes; P95) fetal death of unspecified

cause; Q00-Q99) congenital malformations, deformities, and chromosomal abnormalities; Other) all other possible causes combined.
In order to identify possible trends over time, the entire study period was divided into the following five-years intervals: 1994-1998, 1999-2003, 2004-2008, 2009-2013, 2014-2019.

Statistical analysis

On the basis of these records, the total number of FDs and the Fetal Death Rates (FDRs) were calculated (FDR = number of fetal deaths / total number of newborns *1000), irrespective of sex/gender, for the whole country, the five geographical regions, and for each province and its departments (see more details below, in Data Missing).

To identify significant changes in FDRs over the period under review, we used joint point regression analysis (computed in Joinpoint Regression Program 5.0.2., 2023). This method identifies the year(s) in which a significant trend change occurs, and calculates the annual percentage change (APC) in rates between the trend change points. Each p-value is determined using Monte Carlo methods. The overall asymptotic significance level is maintained using a Bonferroni correction.

Spatial analysis

Moran's I Autocorrelation Index is widely used in order to investigate whether the spatial pattern of a given attribute appears clustered, dispersed, or random (Moran, 1950). The analysis included the 525 country's departments and their FDR values, using the Rook or contiguity criterion. Based on the Global Moran Index, a map of the Local Indicators of Spatial Association (LISA) was calculated following the methodology outlined by Rey, Arribas-Bel, & Wolf (2021). This determines whether a given value in a department and the mean of its neighbors are more similar (high-high or HH, low-low or LL) or different (high-low or HL, low-high or LH) than would be expected by chance. Significance was determined at the 0.05 confidence level utilizing the Monte Carlo test (999 permutations) under the null hypothesis of random distribution.
The analysis was performed using the exploratory spatial data analysis library ESDA of the PySAL: Python Spatial Analysis Library Meta-Package (Rey & Anselin, 2007).

Missing data

Data were complete for the variables needed to study the spatial variation of FD, with the sole exception of the 1994-1999 interval. For that period, data are not disaggregated at the departmental level. Since trend analyses were only performed at regional level, we reported those characteristics without adjustments for missing data.

Ethics approval

This study follows the bioethical guidelines proposed by the Argentine National Ministry of Health (2011), which exempt epidemiological studies that use public or publicly available records or information from obtaining informed consent.

3 | RESULTS

The databases documented 18.405.630 live births from 1994 to 2019, with a total of 173.330 FDs. Table 1 presents the FDRs for the country and by region along the established time intervals. The general trend is for FDRs to decrease over time. The regions of Cuyo and NEA stand out for having considerably higher FDRs than the rest for the first two periods considered (1994-1998, 1999-2003) and for showing a significant decrease further on. Among regions, Patagonia had the lowest rates, whereas Cuyo exhibited the most remarkable reduction throughout the period. The NWA region was the sole area that showed a net increase in FDR at the end of the research time period. For the country as a whole, the FDR declined by 3.8 points.

Table 1. Fetal deaths rates (FDRs) by region and for the entire country, presented in five-year intervals.

Figure 2 displays the joinpoint regression analysis results, highlighting interregional discrepancies in the FDR trends over the years. Cuyo and Patagonia exhibit a consistent decrease in the FDR observed at 2 APC. NWA shows a stable decrease in mortality since 2002. NEA shows its peak in the year 1999, after which it begins to decline until 2010. Significance and statistical analysis of the annual percentage change (APC), along with the 95% confidence interval (CI), are presented in Table 2. There was a notable rise in the observed rates in all regions between 1999 and 2003, which took place amid a severe socio-economic crisis in Argentina. The same situation occurred between

2011 and 2012, with this increase being especially important in the NWA region (red dotted line in Figure 2).

Figure 2. Fetal deaths rates (FDRs) joinpoint regression analysis by region and for the entire country (1994–2019 interval). Dotted lines represent observed rates; continuous lines are fitted rates based on joinpoint analysis.

Table 2. Annual percentage change (APC) and its 95% confidence interval (CI) of FDRs according to joinpoint regression analysis (1994–2019).

Figure 3 shows the percentage of fetal deaths categorized by cause at both the national and regional levels over the five-years periods. For Argentina as a whole, P02 (complications of the placenta, umbilical cord, and membranes), P95 (unspecified cause) and "other causes" are the categories with the highest percentages of occurrence. Together, they explain more than 80% of the deaths (for all levels of analysis). On the other hand, Q00-Q99, P00 and P01 have the lowest percentages in general. The prevalence of FDs due to congenital causes (Q00-Q99) remains stable and relatively low both nationally and within each region, with the highest percentages found in the Central and Patagonia regions. FDs related to maternal conditions (P00 and P01) also present in general a relatively low and stable occurrence.

For Argentina and all its regions, there is a notable decline in fetal deaths caused by P02. Overall, and for all regions (except for Central) the percentage is reduced by more than half at the end of the period considered, especially in the Cuyo region.

On the other hand, the number of FDs classified as unspecified (P95) shows a general upward trend. This trend began for all regions around the second period (1999-2003), which as noted above coincided with a severe socio-economic crisis that affected the whole country. This upward trend is particularly pronounced in the NWA and NEA regions, while the Cuyo region managed to reverse the trend and halve it by the end of the period studied.

Figure 3. Stacked bar charts displaying the percentages of fetal deaths categorized by causes over five-year intervals at both a regional and national level. P00: Fetus affected by maternal conditions unrelated to the current pregnancy; P01: Fetus affected by maternal complications of pregnancy; P02: Fetus affected by complications of the placenta, umbilical cord, and membranes; P95: Fetal

death of unspecified cause; Q00-Q99: Congenital malformations, deformities, and chromosomal abnormalities; Other: All other possible causes combined.

The spatial analysis is displayed in Figure 4, showcasing the FDR departmental clusters and their statistical significance over five-year periods. According to Moran's local spatial analysis, FDR's hot spots, characterized by high rates surrounded by high rates, are more prevalent in some northern regions of the country. Certain areas of Patagonia and the central region have cold spots, which are defined as areas with low rates surrounded by other areas with low rates. Furthermore, all analyses indicated positive spatial autocorrelation, as demonstrated by significant global Moran's I values that were positively signed, with p-values below 0.001. The Moran's I values for the complete period from 1999 to 2019 was 0.57.

Figure 4. The maps display the geographic clustering of FDR departmental values. Departments with high FDR that are surrounded by high FDR neighbors are colored in red (HH). Departments with high FDR but surrounded by low FDR are colored in orange (HL), and departments with low FDR but surrounded by high FDR are highlighted in light blue (LH). Departments with low FDR surrounded by low FDR are colored in blue (LL). In gray, non-significant (NS).

4 | DISCUSSION

Fetal death rates

Significant rate decreases were observed in the Center, Cuyo, and Patagonia regions, which exhibited progressively lower values over the five-year periods. Cuyo has the highest FDR during the initial interval but records the second-lowest value at the end of the period analyzed. This is the largest net reduction in any region. Coincidentally, these three regions have the lowest maternal mortality rates in the country, below the national average (DEIS, 2020)

In the Northwest and Northeast regions, however, the rates of fetal deaths either remain high from the beginning or, worse still, increase instead of decreasing as in the case of NWA. This region experienced a significant rise in its FDR during the second five-year period (specifically between 2000 and 2003, see Figure 2), from which it did not entirely recover in subsequent years. This

highlights the need to analyze the phenomenon at different spatial and time scales, in order to avoid interpretive biases. Clearly, the persistence of high levels of FD in the northern provinces is associated with specific regional and/or local variables.

The infant mortality rate (IMR) is a reliable indicator of the level of development of a population. This is because the IMR results from the interaction between population characteristics, risk or pathogenic factors and the social, physical and biological environment (Bhem, 2011). In Argentina, the NWA has the highest IMR compared to the so-called Humid Pampa (a subregion of the Central Region) (Bolsi, Paolasso & Longhi, 2006), and in 2010 it was the second region with the highest IMR after the NEA (Mazzeo, 2015). The high IMR in the NWA region could be related to environmental factors, such as the altitude: this region reaches up to 5000 meters above sea level. A study by Chapur, Alfaro, Bronberg, & Dipierri (2017) found that postneonatal mortality (28-365 days after birth) was directly related to this geographic feature. The ecology of high-altitude areas is complex due to low oxygen levels, low temperatures, intense solar radiation, and low ambient humidity. These variables have an impact on the quality of life of the population in these areas, in addition to the economic and structural difficulties in accessing health services and/or health information. The NWA is a region with a low level of economic development and the highest rate of poverty in the country (INDEC, 2023), which puts it at a comparative disadvantage to other regions and would make it particularly vulnerable to high rates of FD.

Loiacono, Guevel & Rosa (2020) analyzed the possible relationships between social inequality and fetal mortality in Argentina during the years 2007-2016. The authors concluded that the fetal mortality rate (FMR) is higher in strata with a very unfavorable socioeconomic situation. The gap between the extremes widened towards the last biennium considered (2015-2016). It was also observed that the most disadvantaged and inequitable areas are located in the north and center of the country, while the most advantaged departments are located in the south, with some exceptions.

Similar conclusions can be drawn from studies conducted in other populations regarding how environmental factors, living conditions, and healthcare infrastructure influence health outcomes. Martins, Rezende, de Mattos Almeida, & Félix Lana (2013), using data on perinatal deaths between 2003 and 2007 in the state of Belo Horizonte, Brazil, found that mortality decreased as maternal education increased. In South Asia and particularly in sub-Saharan Africa, around 60% of fetal deaths still occur in rural areas. In these countries, rural families are often the poorest and have

limited access to antenatal care, family planning services, and emergency obstetric care (Blencowe et al., 2016). A systematic review on inequalities and stillbirths in the United Kingdom found that comprehensive research on social inequalities and stillbirths remains underdeveloped in this country, despite repeated evidence that the risk of FD is associated with poverty and ethnicity (Kingdon et al., 2019).

With regard to causes

Improvements in obstetric care, health system facilities, and early monitoring of pregnancy may explain the decline in FDs associated with maternal and/or labor factors (codes P00-P02, Figure 3) over the study period. These changes allow for more controlled deliveries knowing potential complications. At the same time, there has been an increase in causes not specified on the death certificate (code P95). This is probably due to the lack of information on many deaths, which would require a more detailed post-mortem examination. Death certificates registered under code P95 are significantly more common in the north-west and north-east of the country. As mentioned above, these are two of the regions of Argentina with the greatest socio-economic disparities. This situation is more evident in the NWA region, but especially in the last five years (2014-2019). Clearly, the need for adequate staff and hospital infrastructure to provide proper prenatal and postnatal care and to analyze causes of death has been underfunded. The same situation was reported by Martins, Rezende, de Mattos Almeida, & Félix Lana (2013), who concluded that the highest unspecified FMRs occurred in areas of very high socioeconomic risk.

In the Patagonia region, there was also a significant change after the five-year period 1994-1998, with a marked decrease in FD associated with maternal factors, complications of pregnancy or childbirth, and, as in the Central region, a remarkable homogeneity in the most frequent causes from 2009 to the present. The situation is different in Cuyo, where the causes that are included in the "other" category have become much more common in the last five years.

Concerning FD caused by congenital malformations (CM), the percentages remained similar in all periods, both within each region over time and between regions. This phenomenon requires further analysis in terms of the proportion of infant deaths due to CM in relation to all infant deaths, which has tended to increase in recent decades in different countries and also in Argentina, as an expression of the epidemiological transition of infant mortality. This transition can be defined as the

change from a phase dominated by infectious diseases to one dominated by chronic degenerative diseases (Omran, 1971; Rosano Rosano, Botto, Botting, & Mastroiacovo, 2000; Bronberg, Chapur, & Dipierri,, 2021). For 18 countries with reliable data, congenital anomalies account for a median of 7.4 percent of stillbirths (Blecowe et al., 2016).

As noted above, estimates of FD causality are limited by inadequate routine reporting and/or limitations in classification systems. Although information on predictor variables is scarce, some studies have provided evidence in this regard. A case-control study conducted in Ethiopia found significant associations between FD and multiple pregnancies, preterm birth, cesarean section, hypertension during pregnancy, and lack of antenatal care (Abebe, Shitu, Workye, & Mose, 2021). To address this issue, it is recommended to prioritize high-quality facility-based care and invest in community demand and birth planning. Even slight delays in accessing appropriate care can result in death or disability for newborns and women (Upadhyay, Krishnan, Chinnakali, & Odukoya, 2014). In 2016, the World Health Organization (WHO) established the three-delay model to help identify problems in maternal care strategies. Delays in maternal healthcare can arise from various factors, including failure to recognize the need for care (such as delayed seeking of care, lack of awareness of risk signs, or refusal of care), difficulty accessing care (such as inadequate prenatal care, transportation issues, or geographical barriers), and delays in receiving quality care at healthcare facilities (such as delayed diagnosis, communication issues between hospitals and regulatory centers, lack of trained personnel, delayed referral/transfer of cases, delayed treatment initiation, or inadequate patient management). A case-control study was conducted based on prospective surveillance of fetal deaths and live births in a specialized care hospital for high-risk pregnancies in Northeastern Brazil (Martins et al., 2019). The delay in early diagnosis of morbidities and detection of risks during prenatal appointments were strongly associated with fetal deaths, resulting in delayed timely treatment to avoid adverse outcomes. The unequal distribution of obstetric beds is reinforced by the concentration of health services in large urban centers, leaving minor cities and rural populations without qualified medical coverage. Prevention cannot be individualized, strategies to reduce fetal mortality require coordinated and continuous action at multiple levels of care.

5 | CONCLUSIONS

Thanks to collaborating with the Department of Statistics and Health Information (DEIS) of the National Ministry of Health, a comprehensive database was created, facilitating the first epidemiological study delving into the spatial and temporal distribution of FDRs. The investigation presents a preliminary analysis of trends across all administrative levels. Despite significant heterogeneity observed throughout the country's regions, we highlight the following: First, fetal death rates (FDRs) demonstrate a negative secular trend, with greater evidence and significance measurable in Argentina's most developed regions (Cuyo, Center, and Patagonia). Second, FDRs that result from placenta, umbilical cord, and membrane complications tend to decrease. Meanwhile, there has been an increase in FDRs due to nonspecific causes, particularly in the NWA and NEA regions.

Argentina has implemented nationwide policies that have proven to be highly beneficial for public health. One such measure is the fortification of flour with folic acid, aimed at preventing congenital neural tube defects. As a result, there has been a significant reduction in fetal deaths caused by these defects (Calvo and Biglieri, 2008; Bronberg et al., 2023). Given the continuing decline in fertility rates in recent years (DEIS, 2020), it is crucial to maintain and expand efforts to minimize fetal mortality throughout the country, specifically in the most affected regions and provinces.

Author Contributions

Anahí Ruderman: Data curation (equal); formal analysis (lead); writing – review and editing (lead). Arturo L. Morales: Data curation (equal); visualization (lead); formal analysis (equal); writing – review and editing (equal). José Edgardo Dipierri: Data collection (lead); supervision (equal); writing – review and editing (equal). Jorge Iván Martínez: Formal analysis (lead); Maria Soledad Silva: Review and editing (equal). Soledad de Azevedo: Writing – review and editing (equal). Damian L. Taire: Conceptualization (lead); writing – review and editing (equal). Virginia Ramallo: Supervision (equal); writing – review and editing (lead).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Ruderman, Anahí - ORCID 0000-0002-9610-2997

Morales, Arturo Leonardo - ORCID 0000-0002-3980-8862

Dipierri, José Edgardo - ORCID 0000-0002-1679-0727

Martinez, Jorge Ivan - ORCID 0000-0001-9051-5451

De Azevedo, Soledad - ORCID 0000-0003-4601-0717

Taire, Damián Leonardo - ORCID 0000-0001-6505-1560

Ramallo, Virginia - ORCID 0000-0002-7856-4856

REFERENCES

Abebe, H., Shitu, S., Workye, H., & Mose, A. (2021) Predictors of stillbirth among women who had given birth in Southern Ethiopia, 2020: A case control study. PLoS ONE 16(5), e0249865. https:// doi.org/10.1371/journal.pone.0249865

Aminu, M., Bar-Zeev, S., van den Broek, N. (2017) Cause of and factors associated with stillbirth: a systematic review of classification systems. Acta Obstetricia et Gynecologica Scandinavica, 96(5), 519-528. doi: 10.1111/aogs.13126.

Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical analysis, 27(2), 93-115.

Anselin, L. (2019). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In Spatial analytical perspectives on GIS (pp. 111-126). Routledge.

Bhem, H. (2011). Determinantes económicos y sociales de la mortalidad en América Latina. Salud Colectiva, 7(2), 231-51.

Blencowe, H., Bottecchia, M., Kwesiga, D., Akuze, J., Haider, M. M., Galiwango, E., …, Every Newborn-INDEPTH Study Collaborative Group. (2021) Stillbirth outcome capture and classifica-

tion in population-based surveys: EN-INDEPTH study. Population Health Metrics 19(1),13. doi: 10.1186/s12963-020-00239-8.

Blencowe, H., Cousens, S., Jassir, F. B., Say, L., Chou, D., Mathers, C., …, Lancet Stillbirth Epidemiology Investigator Group. (2016) National, regional, and worldwide estimates of stillbirth rates in 2015, with trends from 2000: a systematic analysis. Lancet Glob Health 4(2), e98-e108.

Bolsi, A., Paolasso, P., & Longhi, F. (2005). El Norte Grande argentino entre el progreso y la pobreza. Población & Sociedad, 12-13, 231-70.

Bronberg, R. A., Chapur, V. F., & Dipierri, J. E. (2021). Tendencia secular (1980-2018) de las muertes infantiles por malformaciones congénitas en Argentina. Revista de la Facultad de Ciencias Médicas de Cordoba, 78(3), 287-293. doi: 10.31053/1853.0605.v78.n3.32300.

Bronberg, R., Martinez, J., Morales, L., Ruderman, A., Taire, D., Ramallo, V., & Dipierri J. (2023). Prevalence and secular trend of neural tube defects in fetal deaths in Argentina, 1994-2019. Birth Defects Research, 115(18),1737-1745. doi: 10.1002/bdr2.2248.

Bronberg, R., Schuler-Faccini, L., Ramallo, V., Alfaro, E,. & Dipierri J. (2014). Spatial and temporal analysis of infant mortality from congenital malformations in Brazil (1996-2010). Journal of Community Genetics, 5(3), 269-82. doi: 10.1007/s12687-013-0170-0.

Calvo, E. & Biglieri, A. (2008). Impacto de la fortificación con ácido fólico sobre el estado nutricional en mujeres y la prevalencia de defectos del tubo neural. Archivos Argentinos de Pediatría, 106(6), 492-498.

Chapur, V. F., Alfaro E. L., Bronberg, R., & Dipierri, J. E. (2017). Relación de la mortalidad infantil con la altura geográfica en el Noroeste Argentino. Archivos Argentinos de Pediatría, 115(5), 462-469.

Cousens, S., Blencowe, H., Stanton, C., Chou, D., Ahmed, S., Steinhardt, L., … & Lawn, J. E. (2011) National, regional, and worldwide estimates of stillbirth rates in 2009 with trends since 1995: a systematic analysis. Lancet, 377(9774), 1319-30.

de Bernis, L., Kinney, M. V., Stones, W., Hoope-Bender, P. T., Vivio, D., Hopkins Leisher, S., ..., Preventable Stillbirths Series Advisory Group. (2016). Stillbirths: ending preventable deaths by 2030. Lancet, http://dx.doi. org/10.1016/S0140-6736(15)00954-X

DEIS (Dirección de Estadísticas e Información de Salud). (2017). Presentacion-anuario-2017-DEIS.-pdf. Available in: http://www.deis.msal.gov.ar/wp-content/uploads/2019/01/Presentacion-anuario-2017-DEIS.pdf

Flenady, V., Wojcieszek, A. M, Middleton, P., Ellwood, D., Erwich, J. J., Coory, M., …, The Lancet Stillbirths in High Income Countries Investigator Group. (2016) Stillbirths: recall to action in high-income countries. Lancet, 387, (10019), 691-702, http://dx.doi.org/10.1016/S0140-6736(15)01020-X.

Frøen, J. F., Friberg, I. K., Lawn, J. E., Bhutta, Z. A., Pattinson, R. C., Allanson, E. R., …, The Lancet Stillbirths In High-Income Countries Investigator Group (2016) Stillbirths: progress and unfinished business. Lancet 387, (10018), 574-586.

Heazell, A. E., Siassakos, D., Blencowe, H., Burden, C, Bhutta, Z. A., Cacciatore, J., …, Lancet Ending Preventable Stillbirths investigator group (2016). Stillbirths: economic and psychosocial consequences. Lancet, 387(10018), 604-616. doi: 10.1016/S0140-6736(15)00836-3.

Hoyert, D. L., & Gregory, E. C. (2016) Cause of Fetal Death: Data From the Fetal Death Report, 2014. National vital statistics reports life tables, 65(7):1-25.

Instituto Nacional de Estadística y Censos (INDEC) (2023). Incidencia de la pobreza y la indigencia en 31 aglomerados urbanos. Primer semestre de 2023. Reportes Técnicos, 7 (205). Consulted in: https://www.indec.gob.ar/indec/web/Nivel4-Tema-4-46-152

Joinpoint Regression Program, Version 5.0.2 - May 2023; Statistical Methodology and Applications Branch, Surveillance Research Program, National Cancer Institute.

Kiguli, J., Kirunda, R., Nabaliisa, J., Nalwadda, C. K. (2021). Inadequate evidence on stillbirths: rethinking public health. Lancet, 398, 727-729.

Kingdon, C., Roberts, D., Turner, M. A., Storey, C., Crossland, N., William Finlayson, K. & Downe, S. (2019). Inequalities and stillbirth in the UK: a metanarrative review. BMJ Open. doi:10.1136/bmjopen-2019-029672.

Loiacono, K. V., Guevel, C., & Rosa, E. A. (2020). Inequidad social posiblemente relacionada con mortalidad fetal en Argentina en 2007-2016. Revista Argentina de Salud Pública, 12, 14.

Martins, E. F., Rezende, E. M., Almeida, M. C. de M, & Lana, F. C. F. (2013). Mortalidad perinatal y desigualdades socio-espaciales. Revista Latino-Americana de Enfermagem, 21(5), 1062-1070. https://doi.org/10.1590/S0104-11692013000500008.

Martins, M. C. F., de Lucena Feitosa, F. E., Viana Júnior, A. B., Correia, L. L., Ibiapina, F. L. P., Pacagnella, R. C., & Carvalho, F. H. C. (2019). Pregnancies with an outcome of fetal death present higher risk of delays in obstetric care: A case-control study. *PLoS ONE 14*(4): e0216037. https://doi.org/10.1371/journal.pone.0216037

Mazzeo, V. (2015). La mortalidad infantil en Argentina. Análisis de sus cambios y de las diferencias regionales. Población Y Desarrollo - Argonautas Y Caminantes, 10, 9–20. https://doi.org/10.5377/pdac.v10i0.1734

Moran, P. A. (1950). Notes on continuous stochastic phenomena. Biometrika, 37(1/2), 17-23.

Omran, A. R. (1971). The epidemiologic transition. A theory of the epidemiology of population change. The Milbank Memorial Fund Quarterly, 49(4), 509-38.

Organización Panamericana de la Salud (OPS). (2017). Lineamientos básicos para el análisis de la mortalidad. Washington, D.C.

Rey, S. J., D. Arribas-Bel, & L. J. Wolf. (2021). Geographic Data Science with Python and the PyData Stack. Boca Raton, FL: CRC Press.

Rey, S.J. & L. Anselin (2007). Review of Regional Studies 37, 5-27.

Rosano, A., Botto, L. D., Botting, B., & Mastroiacovo, P. (2000). Infant mortality and congenital anomalies from 1950 to 1994: an international perspective. Journal of Epidemiology and Community Health, 54(9), 660-6. doi:10.1136/jech.54.9.660.

Scalise, C., Cordasco, F., Sacco, M. A., Ricci, P., & Aquila, I. (2022). The Importance of Post-Mortem Investigations in Stillbirths: Case Studies and a Review of the Literature. International journal of environmental research and public health, 19(14), 8817. https://doi.org/10.3390/ijerph19148817

Souza, J. P., & Bahl, R. (2022). PURPOSe study: understanding the burden of stillbirths in South Asia. Lancet Glob Health, 10(7), e930-e931. doi: 10.1016/S2214-109X(22)00218-2.

UN (2015a) Transforming our world: the 2030 agenda for sustainable development. Resolution adopted by the General Assembly on September 25, 2015. New York: United Nations.

UN (2015b) Every Woman Every Child's Global Strategy for Women's, Children's and Adolescents' Health (2016-2030).

Upadhyay, R. P., Krishnan, A., Rai, S. K., Chinnakali, P., & Odukoya, O. (2013). Need to focus beyond the medical causes: a systematic review of the social factors affecting neonatal deaths. Paediatric and Perinatal Epidemiology, 28(2), 127-37. doi: 10.1111/ppe.12098.

World Health Organization - WHO (2016a). The WHO application of ICD-10 to deaths during the perinatal period: ICD-PM. Geneva: World Health Organization;.URL: https://apps.who.int/iris/bitstream/handle/10665/249515/9789241549752-eng.

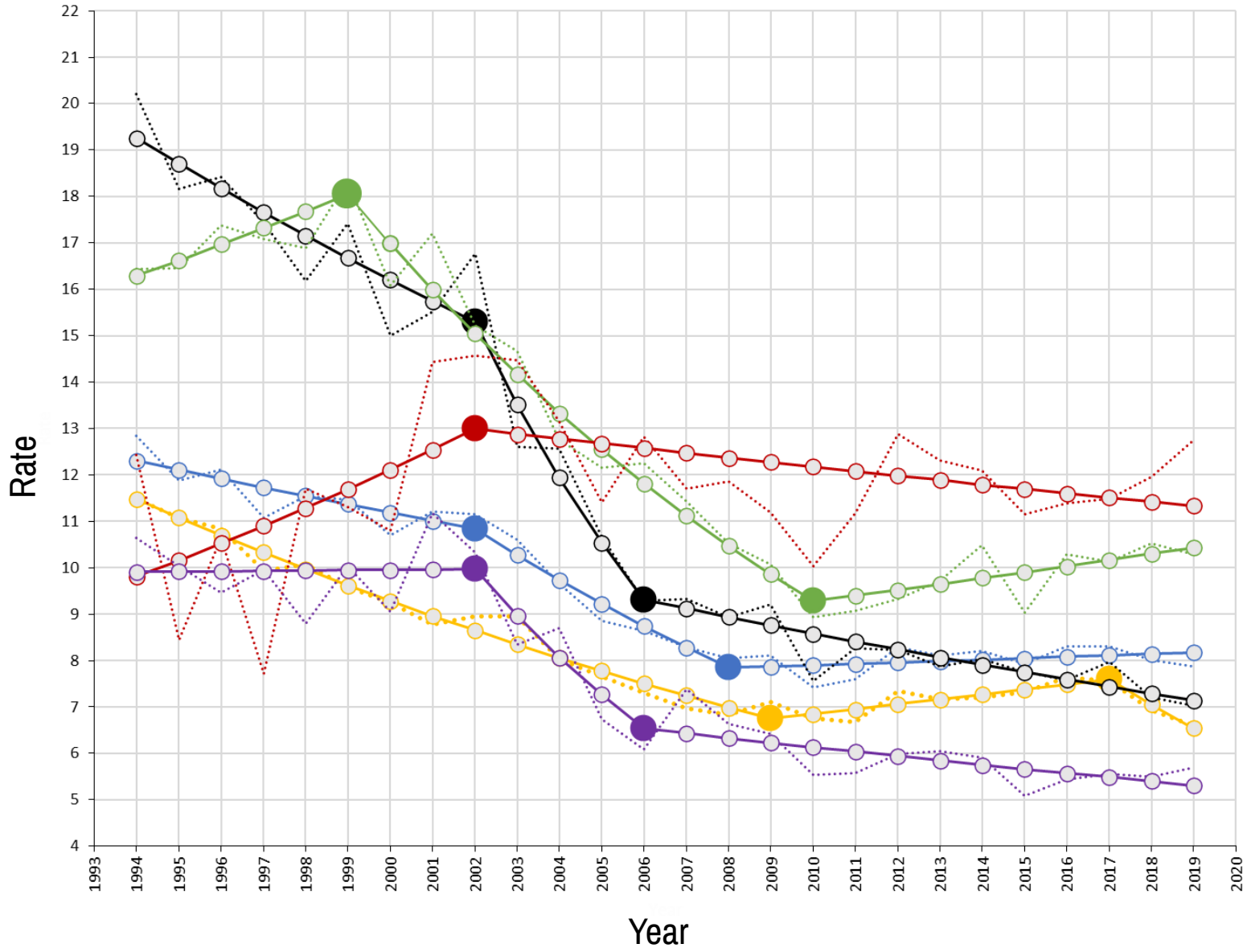
World Health Organization - WHO (2016b) Making every baby count: audit and review of stillbirths and neonatal deaths. Geneva: World Health Organization; Available from: http://

www.who.int/maternal_child_adolescent/documents/stillbirth-neonatal-death-review/en/

TABLE 1. Fetal deaths rates by region and for the entire country, presented in five-year intervals.

Region	1994-1998	1999-2003	2004-2008	2009-2013	2014-2019
Center	11,3	9,4	7,5	7,2	7,7
NEA	17,2	16,4	12	9,4	10,1
NWA	10,4	13,4	12,6	11,3	11,4
Cuyo	18,3	15,7	10,2	8,5	7,6
Patagonia	10	10	7	5,9	5,4
Argentina	11,9	11	8,7	7,9	8,1

	1st Period trend				*2nd Period trend*				*3rd Period trend*			
Region	**Period**	**APC**	**CI (95%)**	**T test**	**Period**	**APC**	**CI (95%)**	**T test**	**Period**	**APC**	**CI (95%)**	**T test**
Argentina	1994-2002	-1.6*	(-2.6--0.6)	-3.3	2002-2008	-5.2*	(-7.4--3.0)	-4.9	2008-2019	0.4	(-0.4-1.1)	1.0
Center	1994-2009	-3.5*	(-3.8--3.1)	-21.4	2009-2017	1.5*	(0.3-2.7)	2.7	2017-2019	-7.2	(-15.8-2.2)	-1.6
Cuyo	1994-2002	-2.8*	(-4.0--1.6)	-4.8	2002-2006	-11.7*	(-17.6--5.3)	-3.8	2006-2019	-2.0*	(-2.9--1.2)	-5.0
NEA	1994-1999	2.1	(-0.4-4.5)	1.8	1999-2010	-5.9*	(-6.8--5.0)	-13.2	2010-2019	1.3	(-0.0-2.7)	2.1
NWA	1994-2002	3.6	(-0.4-7.6)	1.9	2002-2019	-0.8	(-2.0-0.4)	-1.4				
Patagonia	1994-2002	0.1	(-2.2-2.4)	0.1	2002-2006	-10.0	(-20.4-1.8)	-1.8	2006-2019	-1.6*	(-2.9--0.2)	-2.4


Rate
Year
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

Received: 18 January 2023 | Revised: 1 June 2023 | Accepted: 3 June 2023

DOI: 10.1002/ajhb.23938

ORIGINAL ARTICLE

American Journal of Human Biology WILEY

Mapping spatial morbidity patterns for bronchiolitis related to socioeconomic estimators: A spatial epidemiology approach to identify health disparities in Puerto Madryn, Argentina

Bruno A. Pazos[1,2,3] | **Arturo L. Morales**[1,2,3] | **Virginia Ramallo**[1] | **Rolando González-José**[1,4] | **Soledad de Azevedo**[1] | **Damián L. Taire**[1,5]

[1]Instituto Patagónico de Ciencias Sociales y Humanas, Centro Nacional Patagónico (IPCSH), Consejo Nacional de Investigaciones Científicas y Técnicas, Puerto Madryn, Argentina

[2]Laboratorio de Ciencias de las Imágenes, Departamento de Ingeniería Eléctrica y Computadoras, Universidad Nacional del Sur, Bahía Blanca, Argentina

[3]Departamento de Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco, Trelew, Argentina

[4]Programa de Referencia y Biobanco Genómico de la Población Argentina (PoblAr), Secretaría de Planeamiento y Políticas en Ciencia, Tecnología e Innovación, Ministerio de Ciencia, Tecnología e Innovación, CABA, Argentina

[5]Departamento de Neumonología Pediátrica, Hospital Zonal "Dr. Andrés R. Isola", Puerto Madryn, Argentina

Correspondence
Soledad de Azevedo, Instituto Patagónico de Ciencias Sociales y Humanas, Centro Nacional Patagónico (IPCSH), Consejo Nacional de Investigaciones Científicas y Técnicas, Puerto Madryn, Argentina.
Email: deazevedo@cenpat-conicet.gob.ar

Funding information
CONICET, Proyecto Unidad Ejecutora IPCSH, Grant/Award Number: 22920160100081CO

Abstract

Objectives: To describe the frequency of hospitalizations of infants under 1 year of age with bronchiolitis in Puerto Madryn, Argentina, and to study the spatial distribution of cases throughout the city in relation to socioeconomic indicators. To visualize and better understand the underlying processes behind the local manifestation of the disease by creating a vulnerability map of the city.

Methods: We performed a cross-sectional study of all patients discharged for bronchiolitis from the local public Hospital in 2017, considering length of hospital stay, readmission rate, patient age, home address and socioeconomic indicators (household overcrowding). To understand the local spatial distribution of the disease and its relationship to overcrowding, we used GIS and Moran's global and local spatial autocorrelation indices.

Results: The spatial distribution of bronchiolitis cases was not random, but significantly aggregated. Of the 120 hospitalized children, 100 infants (83.33%) live in areas identified as having at least one unsatisfied basic need (UBN). We found a positive and statistically significant relationship between frequency of cases and percentage of overcrowded housing by census radius.

Conclusions: A clear association was found between bronchiolitis and neighborhoods with UBNs, and overcrowding is likely to be a particularly important explanatory factor in this association. By combining GIS tools, spatial statistics, geo-referenced epidemiological data, and population-level information, vulnerability maps can be created to facilitate visualization of priority areas for development and implementation of more effective health interventions. Incorporating the spatial and syndemic perspective into health studies makes important contributions to the understanding of local health-disease processes.

1 | INTRODUCTION

Acute Lower Respiratory Tract Infections (ALRTI) are one of the main causes of death in the world, with more than 4 million deaths per year (FSRI, 2017). In particular, bronchiolitis and pneumonia (with or without complications) are the most relevant because of their impact on morbidity and mortality on child patients (Arch. argent. pediatr., 2015). In Argentina, ALRTI still represents an important cause of morbidity and mortality (MINSAL, 2010). Around 500–600 children under 5 years of age die annually from respiratory infections, representing the third leading cause of death after perinatal deaths and congenital and chromosomal abnormalities. In 2005, ALRTI were responsible for 68 605 hospital discharges of patients under 5 years old, representing 21.5% of the total registered discharges for patients in that age range in the same year ($N = 318\ 347$). Additionally, ALRTI were the second most common cause of hospitalization in this group of patients, only preceded by complications during perinatal period (22.06%) and over other infectious diseases as a whole (11.2% of cases) (MINSAL, 2007).

The term bronchiolitis is generally applied to the first episode of wheezing in infants under 12 months of age (Meissner, 2016). Respiratory syncytial virus or RSV, a worldwide pathogen with global seasonal patterns largely driven by climate, is the leading cause of these infections. In most countries in the southern hemisphere, the RSV epidemic starts between March and June and declines from August to October (Obando-Pacheco et al., 2018). Several hypotheses have been proposed to explain the effect of climatic factors in the seasonality of RSV, including changes in host behavior (e.g., more time spent indoors, in closed environments, during cold or rainy weather), changes in host defense mechanisms (e.g., impairment of mucociliary clearance with inhalation of cold, dry air), and changes in viral infectivity and stability under different climatic conditions (Barros et al., 2014; Mahishale, 2014; Patz et al., 2014; Paynter et al., 2010; Pica & Bouvier, 2012; Sloan et al., 2011). Besides its morbidity and mortality, this pathology demands an important amount of health resources, in particular when hospitalization is required. In addition, the emergence of the COVID-19 pandemic has disrupted epidemiology, particularly for acute respiratory infections in children. For example, there have been global reports of smaller outbreaks than in previous years or delays in RSV seasons ranging from nearly 10 to more than 20 weeks. It is quite possible that these differences are related to how each country dealt with COVID-19 and, in particular, how they managed non-pharmacological interventions (e.g., widespread use of masks and social distancing) to mitigate the pandemic. Nevertheless, it appears that acute respiratory infections in children are returning to their usual pre-pandemic epidemiology (Ferrero et al., 2022).

In Argentina, the most recent national report concludes that although the mortality rate due to respiratory diseases in children under 5 years of age continues to show a downward trend (INER, 2018), there remains a remarkable disparity in its spatial and demographic distribution. Argentina is the second largest country in South America after Brazil (2 780 400 km^2 or 1 073 500 sq mi), with a population of 46 044 703 (INDEC, 2023). Its main administrative divisions are called provinces, which are subdivided into departments. The provinces are often grouped into geographic regions, such as Patagonia (which includes Neuquén, Río Negro, Chubut, Santa Cruz and Tierra del Fuego). Due to its extension, climatic conditions vary widely. For the past decade, the historical record of extreme coldest temperatures in the city of Puerto Madryn (Chubut, North Patagonia) has occurred during the months of June to July. At this latitude, the environmental conditions that lead to a significantly higher risk of community-transmitted respiratory viruses occur between fall and winter. They begin in May, peak between June and July, and can last until September (Cannizzaro & Nuñez de la Rosa, 2020).

In Argentina, in 2018, the majority of child deaths (children under five) were caused by ALRTI, and more than half (52.8%) were recorded between June and September. The efforts of the public health system are focused on the prevention and treatment of diseases during this critical period of life. Beyond the influence of climate, other factors must be considered to understand epidemiological data and their spatial distribution (INER, 2018). For decades, syndemic studies have pointed to the biosocial nature of disease, that is, environmental factors that cause or exacerbate disease or increase vulnerability or disease interactions. Social inequalities lead to adverse living and working conditions that promote the aggregation of both infectious and non-infectious diseases (Singer, 2013). With respect to respiratory health, these interactions may be immediate or delayed, because physiological and immunological mechanisms are concomitant and closely associated. RSV infection has been shown to act as a predisposing factor for bacterial co-infection and often correlates, in susceptible individuals or populations, with more severe disease than simple RSV infection (Oliva & Terrier, 2021). Moreover, according to the economic geography literature, although agglomeration, urbanization and migration are often key features associated with population growth and economic development, there is a risk that growth may be spatially unequal, at least over the short to medium term (McKay & Perge, 2015). Spatial inequality refers to

the uneven distribution of resources and services across different areas or locations, including health care, welfare, public services, household income and infrastructure. The spatial distribution of these characteristics can be examined according to proximity, distance, clustering and concentration. Understanding and measuring the nature of spatial differences and their trends can inform the development of policies, strategies and interventions to reduce morbidity and mortality and improve the access to health resources.

For example, the incidence of hospitalizations (which are often caused by RSV) can vary widely among communities with different levels of socioeconomic status. Zheng et al. (2022) recently analyzed comprehensive hospitalization databases from three U.S. states and found that children from families living in low socioeconomic status (low-SES) areas had the highest incidence of RSV-associated respiratory hospitalizations. Another comprehensive study of communicable respiratory diseases and associated socioeconomic variables was conducted in the poorest region of Chile (Araucanía), which also has the largest ethnic Mapuche population. Spatial statistics and geographic information systems (GIS) were combined with individual income indices and hospital discharge records aggregated by neighborhood. A statistically significant relationship was found between poverty and respiratory infections (Rojas, 2007). Several authors have already highlighted the importance of conducting multilevel analyses of health outcomes related to socioeconomic conditions, even at smaller territorial scales such as neighborhoods (Diez Roux & Mair, 2010; Kawachi & Berkman, 2003; Navazo et al., 2018; Pickett & Pearl, 2001). This type of research represents a further development of applied biological anthropology, and a conceptual and methodological improvement that allows health disparities and social determinants of health to be framed, measured, and analyzed along with the social inequalities that may shape them (Thayer et al., 2022).

Moreover, advances in geographic information systems, statistical methodology, and the availability of geographically referenced data on health and socio-environmental quality have created new opportunities to study and explain local geographic variations in disease (Elliott & Wartenberg, 2004). In the case of small areas, people and communities tend to cluster in space in systematic ways that can be highly predictive of disease risk (e.g., people with high socioeconomic status tend to live near other people with high incomes and in areas with better housing and schooling than those in lower-income areas). Spatial epidemiology is useful to describe and analyze health data in relation to demographic, environmental, behavioral, socioeconomic, genetic, and infection risk factors to explain local geographic variations in disease.

Besides the complexities observed in other contexts and populations, there is a lack of information incorporating these levels of analysis regarding the study of risk factors for bronchiolitis in Argentina in general, and in Patagonia in particular. Spatial epidemiology combined with evaluation of social determinants of health inequalities is still rare in the study of child (and mother–child) health problems. In this regard, household overcrowding (a condition in which the number of occupants exceeds the capacity of the available housing space) leads to adverse physical and mental health outcomes (WHO, 2018), and has been considered as an indicator of material deprivation and treated as a proxy for individual socioeconomic status (Cable & Sacker, 2019), as well as a risk factor for respiratory diseases.

Taking all this into account, here we aimed to describe the frequency of hospitalization of children with bronchiolitis in the public health system of the city of Puerto Madryn and to analyze the possible influence of household overcrowding, in order to identify possible local patterns or inequalities that facilitate the visualization of priority areas for the development and implementation of more effective health interventions. Specifically, we used GIS and spatial statistics to test the null hypothesis of a random spatial distribution of cases in the city and to identify possible hotspots and social determinants of bronchiolitis. It is expected to contribute to the generation of evidence-based interventions and the improvement of child welfare and care.

2 | METHODS

We designed a cross-sectional study analyzing patients under 1 year old hospitalized between January 1st, 2017 and December 31st, 2017 with a diagnosis of acute bronchiolitis, requiring oxygen therapy. Patients who did not require oxygen therapy and those with other concomitant respiratory or cardiologic diseases were not included. This exclusion criterion is based on the fact that less than 3% of infants require hospitalization in these circumstances, and those with comorbidities would require prolonged hospitalization, oxygen therapy, and/or mechanical ventilation (Arch. Argent. Pediatr., 2021). The protocol used and all ethical considerations were evaluated and approved by the Interdisciplinary Teaching and Research Committee of the Public Hospital Zonal "Dr. Andrés R. Isola" (15/12/2017).

The following data were collected from the hospital archive for each case: date of birth, age in months, admission and discharge date, home address, primary discharge diagnosis with its corresponding International Statistical Classification of Diseases and Related Health

Problems alphanumeric code J21 (ICD-10: 2016), and oxygen requirement. From the collected hospital dataset, we calculated the hospitalization or length of stay (LOS, expressed in days) and the readmission rate. This is the percentage of admitted patients who return to the hospital within a month of discharge.

To visualize our data, we used Geographic Information System (GIS). This tool integrates and relates different types of data, allowing the organization, visualization, storage, analysis and modeling of information linked to a spatial reference. It facilitates the incorporation of structural, socioeconomic and/or environmental aspects.

We first created a polygon map using the census radii boundaries information for the city of Puerto Madryn, obtained from the National Institute of Statistics and Census. Then, each patient's home address was translated into geo-referenced data (latitude and longitude). A layer of points was generated by georeferencing the addresses of diagnosed patients using Google Maps Geocoding API (Google, 2022; Van Rossum & Drake, 2009) and a custom Python algorithm developed by the authors. By combining these layers of points and polygons we accurately located bronchiolitis cases on the city polygon map, and determined the number of cases per unit (census radius).

In addition, we used data obtained from the National Institute of Statistics and Census to geographically code households according to the presence or absence of at least one unsatisfied basic needs (UBN). We used overcrowding, which measures the ratio of the number of bedrooms to the total number of inhabitants in each house. Operationally, critical overcrowding in a home is considered to exist when there are more than three people per room (Feres & Mancero, 2001; INDEC, 2010). The percentage of overcrowded households per census radius was computed and added as a new attribute layer to the polygon map.

2.1 | Spatial statistical analysis

To examine the spatial distribution of cases and to detect possible clustering, we used Moran's index of autocorrelation (Moran's *I*). Moran's *I* measures the correlation coefficient for the relationship between a variable and its surrounding values. It evaluates whether the expressed spatial pattern is clustered, dispersed, or random, given a set of features and an associated attribute (Moran, 1950). The first step is to define a neighborhood for a unit (polygon) of our map to construct a weight matrix that represents the spatial relationships that exist between the features in our dataset. Neighbor selection strategies can be based on contiguity/adjacency relationships, or on weights that are derived from distance-based relationships. In this case, a neighbor was defined according to Rook's contiguity criteria (two regions are considered neighboring if they share a common boundary). The Rook method is the most commonly used because of its simplicity, and here it allowed us to obtain the best Moran's *I* with the highest statistical significance. We then calculated the average number of neighboring cases (spatial-lag) and plotted cases-spatial-lag against cases for each census radius (the Moran scatterplot). The least-squares regression line that best fits the relationship between cases-spatial-lag and cases after normalizing the variables is the Global Moran's *I* coefficient. The index varies between −1 and 1, where positive values indicate positive spatial association, where neighboring units have close values, indicating a tendency to cluster (high or low values cluster near other high or low values, respectively); negative values indicate negative spatial association, where neighboring units have a tendency to disperse (high values repel other high values and tend to be close to low values); and values close to or equal to 0 indicate no spatial autocorrelation or random distribution.

Although the Global Moran's Index provides a general idea of the spatial behavior of the data, it can also be decomposed into its components, providing a localized measure of autocorrelation to obtain a map of "hotspots" and "coldspots" (Rey et al., 2021). The central idea of Local Indicators of Spatial Association (LISA) is to determine when a given value and the average of its neighbors are more similar (high–high or HH, low–low or LL) or different (high–low or HL, low–high or LH) than would be expected by chance. In addition to assessing the autocorrelation of a variable in space, it is possible to examine the relationship between two variables and their location in space using the bivariate extension of Moran's *I*, a measure of spatial autocorrelation that is used to simultaneously assess the degree of spatial dependence between two variables (Anselin et al., 2002). While LISAs measure the spatial cluster (similar neighbors) and spatial dispersion (scattered or different neighbors) of features of a variable and the lag of the same variable across neighbors, Bivariate Local Indicators of Spatial Association (BILISAs) focus on the spatial cluster and spatial dispersion between features of one variable and another different variable across neighbors (Anselin & Rey, 2014).

We computed a univariate global Moran's *I* on the cases of bronchiolitis to test the null hypothesis of a random spatial distribution of cases across the city. We then used a local bivariate Moran's *I* (BILISA) to measure the strength and direction of the spatial relationship between bronchiolitis cases and overcrowding and to visualize possible clustering across the city. Significance was estimated at the 0.05 confidence level using the Monte Carlo

test (999 permutations) under the (null) hypothesis of random distribution. The analysis was performed using the ESDA exploratory spatial data analysis library from the PySAL package: Python Spatial Analysis Library Meta-Package (Rey & Anselin, 2010).

3 | RESULTS

A total of 120 infants younger than 12 months discharged from the hospital met the inclusion criteria and were selected for this study. The median age was 4.45 months (IQR 3.9–5). The distribution of the group by age was 12 (10%) <1 month, 76 (63.33%) between 1 and 5 months, and the remaining 32 (26.66%) between 6 and 11 months. The mean length of stay (LOS) was 7.30 days. Fourteen patients (11.66%) had to be readmitted to the hospital, 7 of them within a period of even less than 1 month after their discharge. Overall, there was no significant association between patient age, length of stay, and readmission status (results not shown). It is important to note that of the 120 hospitalized cases, 100 (83.33%) lived in neighborhoods that were previously registered as having overcrowded households in both the 2001 and 2010 Argentine Permanent Household Surveys. Moreover, most of the cases requiring readmission (12 out of 14 or 85.7%) lived in these neighborhoods.

Figure 1 shows a map of the city with census radii colored according to the percentage of households with overcrowding (divided into four classes), overlaid with geo-referenced cases of bronchiolitis (both admitted and readmitted).

As can be seen in this figure, both the spatial distribution of the cases and the presence of overcrowded households along the census radii are not random (i.e., many cases seem to cluster from the center to the north and west, while the south and especially the east of the city have the lowest percentage of overcrowding and almost no cases of bronchiolitis). This emerging spatial pattern was statistically significant. We obtained a positive univariate global Moran's I of 0.31 ($p = .001$). The null hypothesis of spatially random distribution was rejected; the positive coefficient indicates that the spatial distribution of high values and/or low values in the dataset is more spatially clustered than would be expected if the underlying spatial processes were random.

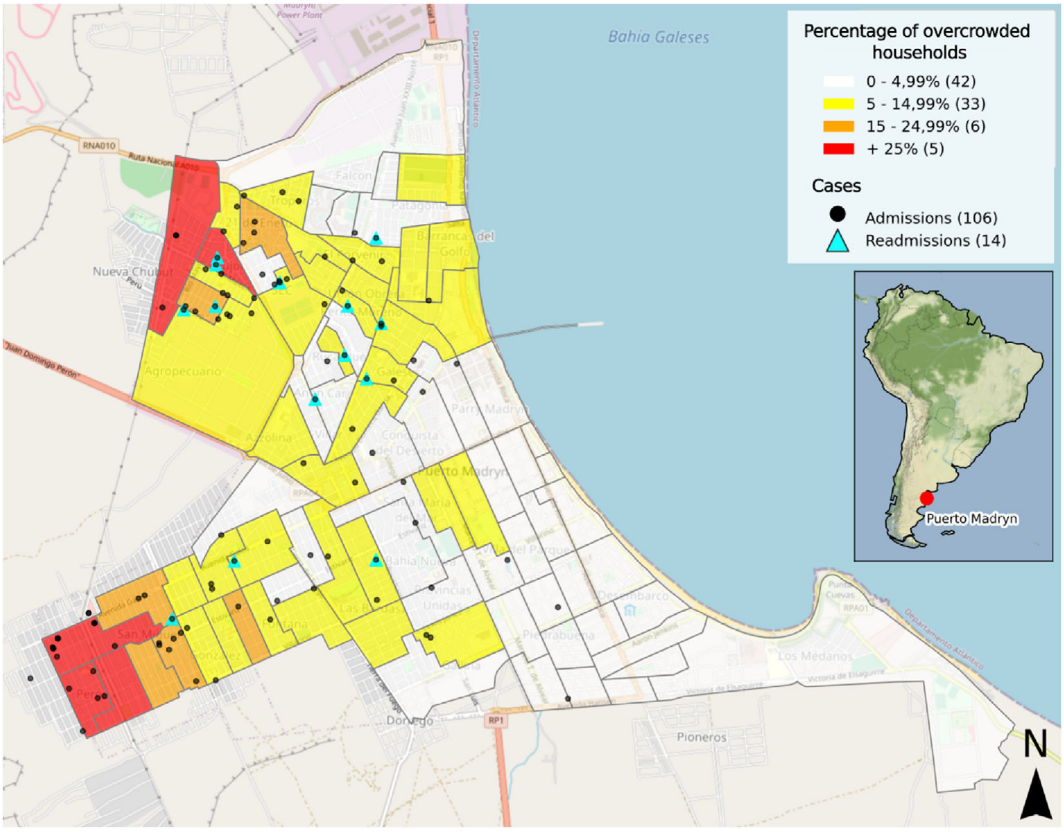


FIGURE 1 Map of Puerto Madryn with census radii colored by percentage of overcrowded households (in parentheses, the number of census radii falling into each category). Black dots and light blue triangles represent geo-referenced bronchiolitis cases (admissions and readmissions, respectively) in 2017.

Figure 2 shows the results of Moran's statistics for the bivariate spatial relationship between cases and overcrowding.

We obtained a Global Bivariate Moran's *I* coefficient of 0.45, which was statistically significant according to Monte Carlo permutations (*p*-value = .001), confirming an overall spatial pattern different from that expected by chance. The high and positive coefficient indicates that similar values of the two variables tend to cluster together in space. The Moran scatterplot (Figure 2A) provides a visual representation of the spatial associations in the neighborhood around each observation, allowing an assessment of how similar an observed value is to its neighboring observations. On the *x*-axis are the standardized values of bronchiolitis cases for each spatial unit (polygon) and on the *y*-axis the spatial lag of

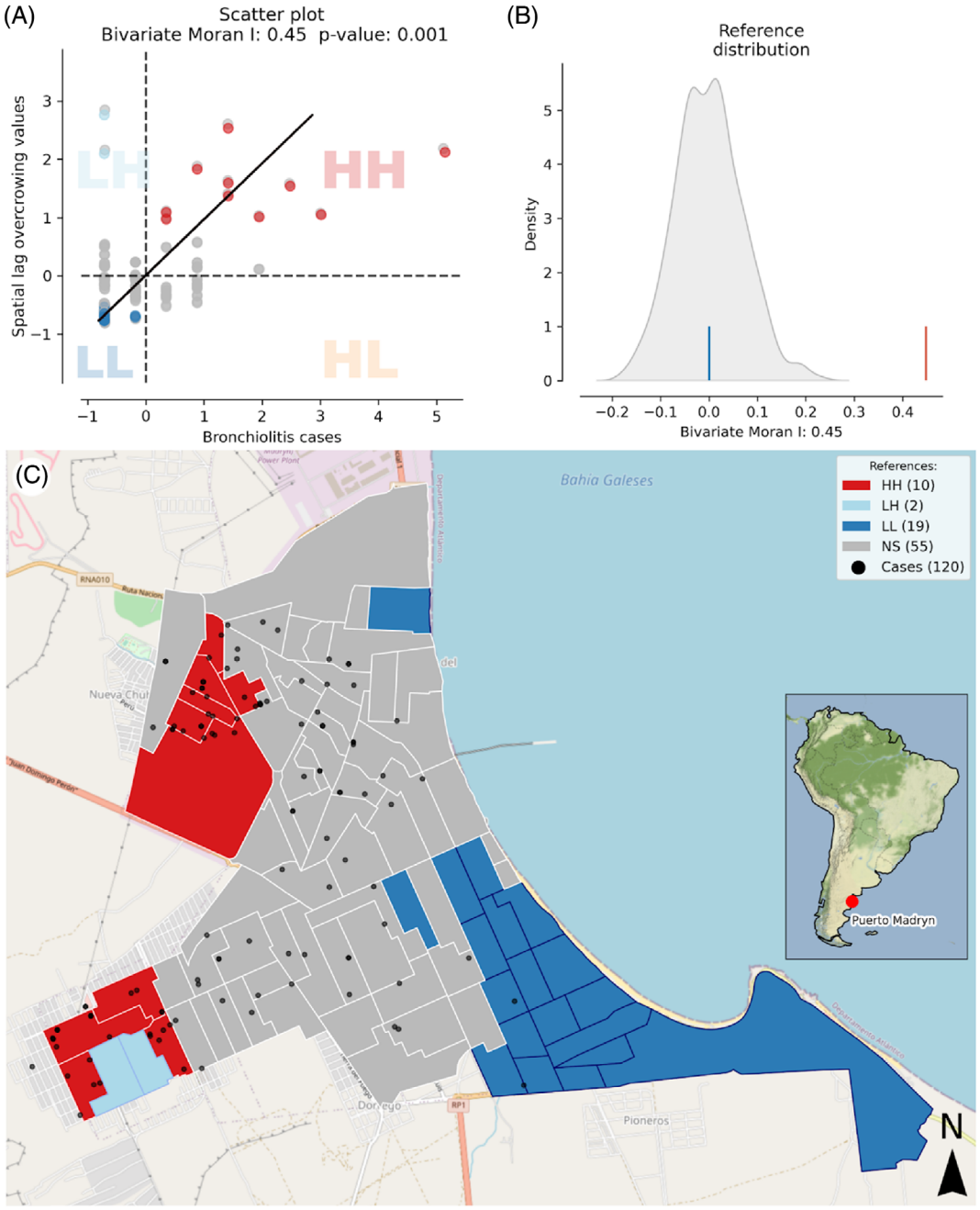


FIGURE 2 (A) Bivariate Moran scatterplot for bronchiolitis cases and overcrowding, showing four possible states: in red, high values of spatial lag surrounded by other high values (HH), in blue, low values surrounded by low values (LL) and the remaining combinations, HL (orange), LH (light blue), and non-significant or NS (gray). (B) Reference distribution for Bivariate Moran's *I* with its mean value in blue and the observed Moran's *I* value in red (0.45). (C) Bivariate Local Indicators for Spatial Association (BILISAs) map. Black dots represent bronchiolitis cases (total number in parentheses). Colored rectangles represent polygons corresponding to the states and color keys mentioned in (A) (in parentheses, the number of polygons or census radii falling into each of these categories).

overcrowding values. Points in the upper right (or high–high = HH) and lower left (or low–low = LL) quadrants indicate positive spatial association, with values higher and lower, respectively, than the sample mean. The lower right (or high–low = HL) and upper left (or low–high = LH) quadrants contain observations with negative spatial association (these observations have little similarity to their neighbors). Figure 2B shows the distribution of values generated by performing random permutations on the data location in space and calculating its Moran's *I* value. After 999 permutations, a reference distribution is obtained (mean value in blue) and compared against the observed Moran's *I* value (in red), which falls at the positive extreme of this distribution. Figure 2C shows each polygon on the map (census radii) colored according to its corresponding quadrant location in the Moran scatterplot, indicating which units have statistically significant autocorrelation between these variables. Two forms of positive significant local spatial autocorrelation are possible: significant clustering of HH values, or so-called "hot spots" (red polygons in Figure 2C), and significant clustering of LL values, or "cold spots" (blue polygons in Figure 2C). On the other hand, locations with significant but negative local autocorrelation occur when a low value is surrounded by locations with a high number of cases (LH, sky-blue polygons in Figure 2C). Again, the emerging spatial pattern seen in Figure 1 is quantified and supported here by BILISAs statistics. Confirming the spatial dependence between bronchiolitis cases and overcrowding, significant local spatial hot and cold spots were identified in the northwest and southeast of the city, respectively.

4 | DISCUSSION

Despite its viral etiology, the epidemiology of acute lower respiratory tract infection (ALRTI) is known to include socioeconomic, geographic, cultural, and genetic factors, increasing the need for localized registries and assessments (Nair et al., 2013). For example, demographic variables such as population density, paternal smoking at home, presence of a sibling, and history of hospitalization (Hyder et al., 2022; Pitzer et al., 2015) are known to influence respiratory virus transmission. Their nature and effects may be very specific and require local analyses. In this context, household overcrowding is a globally recognized risk factor for respiratory disease. However, there is a lack of research on its impact on hospital admissions due to ALRTI among children under five in Argentina.

The objective of this study was to register the spatial distribution of bronchiolitis in children under 1 year of age in the city of Puerto Madryn, North Patagonia, and to investigate its association with household overcrowding. Although no significant relationship was found between the age of the patients included here, their length of stay in the hospital and their readmission status, we found that a high proportion (83.33%) of infants hospitalized for bronchiolitis in 2017 lived in neighborhoods corresponding to homes with overcrowding, a measure of unsatisfied basic needs (UBN). The percentage of readmissions is an important measure of the quality of the health care system and, in bronchiolitis, may be a sensitive indicator of other problems, such as external/environmental factors. The concept of UBN is based on the establishment of minimum thresholds of well-being, according to universally accepted levels, which must be reached by satisfying a set of basic material needs. According to the census methodology adopted by the National Register, an important and key criterion used to diagnose a household with UBN is the existence of overcrowding conditions, given its wide range of negative health consequences. Precarious housing conditions also include sanitation problems (no access to piped water inside the house, no flush toilet, and/or no access to the sewerage system), no access to natural gas or electricity, minimal thermal insulation, inefficient heating systems, and minimal ventilation per person (DPE, 2010). But overall, of all the factors that constitute a deficit situation, overcrowding is considered by several authors to be the most serious indicator of a critical situation due to its multiple negative consequences (Dockery et al., 2021; Solari & Mare, 2012). Starting with the most obvious reason, overcrowding increases the number of cohabitants who spread respiratory infections (Mage et al., 2016), with infants being particularly vulnerable (Caballero et al., 2018; MINSAL, 2018; Tuñón, 2019). The spatial pattern shown here by this indicator (Figures 1 and 2) illustrates the problem of overcrowded housing in relation to acute lower respiratory infections in Puerto Madryn. Of course, this does not necessarily mean that overcrowding is the only (or even the most important, since it is the only explanatory variable included in this analysis) influence on bronchiolitis risk. However, overcrowding could be a proxy for other aspects of the families' environment that may covary with it, such as nutritional status, environmental pollution (air, water), among others.

Our findings are consistent with several studies around the world. For example, overcrowding was the most significant underlying factor responsible for the pattern observed in Nigeria in 2020. This study examined the spatial distribution of ALRTI cases in children under five across the country and found strong spatial variation and hotspots (Osayomi et al., 2020). Household overcrowding also appears to explain why children under five in Bangladesh have one of the highest ALRTI rates in the world (Islam et al., 2021).

In Argentina, Velázquez and Linares (2008) developed a GIS map of well-being indicators covering all

departments of the country in 1991 and 2001, based on the socio-economic dimensions of education, health, housing, and environmental conditions. For the Northern Patagonia region, they found that more than half of the households in the census were overcrowded, given a strong migratory process that has not been accompanied at the same rhythm by the construction and expansion of healthy housing. Migration, as a complex biological and social phenomenon involving individuals (migrants and their families) and societies, is particularly relevant in this local context. Between 1970 and 2010, Puerto Madryn multiplied its population by 13, from 6100 to almost 80 000 inhabitants, due to the development of fishing, construction, tourism and the expansion of the only aluminum factory in Argentina (Sassone et al., 2011). This dramatic population growth has changed the demographic structure and generated several collateral problems, such as urban poverty concentration and uneven restructuring of public services, among other socioeconomic and environmental aspects (Kaminker, 2015, 2020). This process has many dimensions and implications, including the increase in overcrowding. Previous demographic analyses (Kaminker, 2016, 2020) have shown that Puerto Madryn's residential expansion practice, called accelerated urbanization, has structurally segregated low-income households. According to the National Institute of Statistics and Censuses, based on information collected in 2010, 9.42% of the total population of Puerto Madryn lived in households with one or more UBN (INDEC, 2015). The urban planning that has been carried out in the city has not taken into account how the management of resources has contributed to the production and reproduction of inequalities. Severe weather events have highlighted these structural deficiencies, such as an extreme rainfall in 2016. The socio-environmental analysis of the extent and distribution of the resulting flooding (Bilmes et al., 2016) showed that the most affected areas corresponded to the southwestern and northwestern suburbs, exactly the same sectors of the city where, according to the present study, bronchiolitis cases were spatially clustered.

Regarding the characteristics of the local health system, it is important to note that Puerto Madryn has an average of 1200–1400 live births per year. It is estimated that half of them are born in the local public hospital (Hospital "Dr. Andrés R. Isola") and the rest in private medical centers in the city. The public system comprises one hospital and eight primary health centers where people can receive free care. The existence of both public and private health care could bias our results, leading to over- or under-representation of those served by one or the other. However, this is not the case in this study. The local hospital is the only place in the city equipped with a pediatric intensive care unit (ICU) that can meet the demand for severe cases of bronchiolitis, the focus of this study. Therefore, there is no difference between patients who have health insurance and those who do not, in Puerto Madryn they are all inevitably referred to the same hospital.

As mentioned above, the focus of our analysis was on the influence of overcrowding as a proxy for household socioeconomic status. Other characteristics of the family environment may also have an impact on the manifestation of the disease. These may be reflected in other UBNs not used here, or in variables not included in the government UBNs evaluation. Future studies will benefit from using spatial regression analysis to model the relationship between bronchiolitis and different variables (biological, socioeconomic and environmental), considering spatial dependence, to explain and predict local risk factors for this disease.

The production of district-level vulnerability maps using available national registers based on population and housing surveys is a useful source of information that can be cross-referenced with geo-referenced epidemiological data to facilitate visualization, identification and analysis of priority areas. The social gradient in health and disease that exists within and between countries is a global challenge (Marmot, 2003, 2005). This social gradient involves justice, moral and ethical dimensions and is largely shaped by health inequalities due to systematic, unnecessary, avoidable, unfair and unjust differences in health experiences and outcomes (Braveman, 2019; Thayer et al., 2022). These inequalities are caused by various factors such as poverty, discrimination and their consequences such as powerlessness and lack of access to fairly paid jobs, quality education and housing, safe environments and health care. Academic medicine and applied bioanthropology should play an important role in studying and measuring health inequalities, taking advantage of interdisciplinary (e.g., spatial epidemiology, human biology, and sociology) and multisectoral (e.g., academic researchers, health care providers, and government) approaches and collaborative efforts.

5 | CONCLUSION

Our strategy was a retrospective analysis that combined information at the population level with data at the individual level. To the best of our knowledge, this is the first study in Argentine Patagonia that combines hospital records, patients' home location, household well-being and access to basic needs. Although other authors have postulated that such approaches require internal validation to avoid bias (Singer et al., 2021), we believe this is

the most appropriate given the characteristics of local health facilities, known birth rates, cold season climate characteristics in the Patagonian region, and RSV seasonality. Studying this set of variables (whose annual behavior is expected and known) combined with GIS tools are useful to describe clusters of cases and concomitant conditions within our city. A clear association was found between ALRTI and neighborhoods with UBNs in Puerto Madryn. Overcrowding is likely to be a particularly important explanatory factor in this association. In the bibliography cited in this work, overcrowding has also been referred to as a relevant factor in respiratory diseases, but analyses cannot simply assume that the factors present in one community are the same for others. It is important to conduct local studies because communities are always unique in their health and disease processes. The current findings may have important implications for the development and implementation of more effective health interventions. Our interest in future work is to focus on the neighborhoods included in the clusters already identified and to analyze the progression of cases during and after the COVID-19 pandemic. The spatial mapping of diseases and their risk factors is a promising tool for improving our ability to understand the complex relationship between human health and socio-economic and environmental factors. This will allow us to determine whether there are other structural deficits that affect the most disadvantaged population groups (who are also the most dependent on the public health care system).

AUTHOR CONTRIBUTIONS

Bruno A. Pazos: Data curation (equal); software (equal); formal analysis (equal); visualization (equal); writing – review and editing (equal). **Arturo L. Morales**: Data curation (equal); software (equal); formal analysis (equal); visualization (equal); writing – review and editing (equal). **Virginia Ramallo**: Funding acquisition (equal); writing – review and editing (lead); supervision (equal). **Rolando González-José**: Funding acquisition (equal); writing – review and editing. **Soledad de Azevedo**: Data curation (equal); formal analysis (lead); writing – review and editing (lead). **Damián L. Taire**: Conceptualization (lead); data curation (equal); writing – original draft (lead); writing – review and editing (equal).

ACKNOWLEDGMENTS

We thank Mauro Novara for the elaboration of the cartography with a geographic information system, and the administrative staff of the Hospital Zonal "Dr. Andrés R. Isola," for their collaboration in the data registration necessary for this research during 2017. We would also like to thank two anonymous reviewers for their very useful comments to the present study.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Bruno A. Pazos https://orcid.org/0000-0002-0965-070X
Arturo L. Morales https://orcid.org/0000-0002-3980-8862
Virginia Ramallo https://orcid.org/0000-0002-7856-4856
Rolando González-José https://orcid.org/0000-0002-8128-9381
Soledad de Azevedo https://orcid.org/0000-0003-4601-0717
Damián L. Taire https://orcid.org/0000-0001-6505-1560

REFERENCES

Anselin, L., & Rey, S. J. (2014). *Modern spatial econometrics in practice: A guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press LLC.

Anselin, L., Syabri, I., & Smirnov, O. (2002). Visualizing multivariate spatial correlation with dynamically linked windows. In L. Anselin & S. Rey (Eds.), *New tools for spatial data analysis: Proceedings of the specialist meeting*. Center for Spatially Integrated Social Science (CSISS), University of California, Santa Barbara.

Archivos Argentinos de Pediatría. (2015). Recomendaciones para el manejo de las infecciones respiratorias agudas bajas en menores de 2 años. *Sociedad Argentina de Pediatría, Consensos*, *113*(4), 373–374. http://www.sap.org.ar/consensos.php

Archivos Argentinos de Pediatría. (2021). Recomendaciones para el manejo de las infecciones respiratorias agudas bajas en menores de 2 años. *Sociedad Argentina de Pediatría, Subcomisiones, Comités y Grupos de Trabajo*, *119*(4), S171–S197. https://www.sap.org.ar/docs/publicaciones/archivosarg/2021/v119n4a38s.pdf

Barros, V. R., Boninsegna, J. A., Camilloni, I. A., Chidiak, M., Magrín, G. O., & Rusticucci, M. (2014). Climate change in Argentina: Trends, projections, impacts and adaptation. *Wiley Interdisciplinary Reviews: Climate Change*, *6*(2), 151–169. https://doi.org/10.1002/wcc.316

Bilmes, A., Pessacg, N., Álvarez, M. P., Brandizi, L., Cuitiño, J. I., Kaminker, S., Bouza, P. J., Rostagno, C. M., Núñez de la Rosa, D., & Canizzaro, A. (2016). *Inundaciones en Puerto Madryn: Relevamiento y diagnóstico del evento del 21 de Enero de 2016*. Informe Técnico CCT CONICET-CENPAT. https://doi.org/10.13140/RG.2.2.15254.34886

Braveman, P. A. (2019). Swimming against the tide. *Academic Medicine*, *94*(2), 170–171. https://doi.org/10.1097/acm.0000000000002529

Caballero, M. T., Bianchi, A. M., Nuño, A., Ferretti, A. J., Polack, L. M., Remondino, I., Rodríguez, M. G., Orizzonte, L., Vallone, F., Bergel, E., & Polack, F. P. (2018). Mortality associated with acute respiratory infections among children at home. *The Journal of Infectious Diseases*, *219*(3), 358–364. https://doi.org/10.1093/infdis/jiy517

Cable, N., & Sacker, A. (2019). Validating overcrowding measures using the UK household longitudinal study. *SSM – Population Health*, *8*, 100439. https://doi.org/10.1016/j.ssmph.2019.100439

Cannizzaro, A., & Nuñez de la Rosa, D. (2020). *El COVID-19 según el clima*. https://cenpat.conicet.gov.ar/el-covid-19-segun-el-clima/

Diez Roux, A. V., & Mair, C. (2010). Neighborhoods and health. *Annals of the new York Academy of Sciences*, *1186*, 125–145. https://doi.org/10.1111/j.1749-6632.2009.05333.x

Dockery, A. M., Moskos, M., Isherwood, L., & Harris, M. (2021). *How many in a crowd? Assessing overcrowding measures in Australian housing, AHURI final report no. 382*. Australian Housing and Urban Research Institute Limited. https://doi.org/10.18408/ahuri8123401, https://www.ahuri.edu.au/research/final-reports/382

DPE – Dirección Provincial de Estadística de la provincia de Buenos Aires. (2010). *Métodos de Medición de la Pobreza. Conceptos y Aplicaciones en América Latina. Equipo de Trabajo de la Encuesta de Hogares y Empleo* (pp. 1–11). Entrelíneas de la Política Económica. http://sedici.unlp.edu.ar/bitstream/handle/10915/15399/Documento_completo.pdf?sequence=1

Elliott, P., & Wartenberg, D. (2004). Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives*, *112*(9), 998–1006. https://doi.org/10.1289/ehp.6735

Feres, J. C., & Mancero, X. (2001). *El método de las necesidades básicas insatisfechas (NBI)* (pp. 1–54). División de Estadísticas y Proyecciones Económicas, Naciones Unidas, CEPAL, Santiago de Chile. https://repositorio.cepal.org/bitstream/handle/11362/4784/S0102117_es.pdf?sequence=1

Ferrero, F., Ossorio, M. F., & Rial, M. J. (2022). The return of RSV during the COVID-19 pandemic. *Pediatric Pulmonology*, *57*(3), 770–771. https://doi.org/10.1002/ppul.25802

FSRI – Foro de las Sociedades Respiratorias Internacionales. (2017). *El impacto global de la Enfermedad Respiratoria. Segunda edición* (pp. 1–48). México, Asociación Latinoamericana de Tórax. https://gard-breathefreely.org/wp-content/uploads/2017/11/Firs2017_esp_web.pdf

Google. (2022). *Google maps geocoding API*. https://developers.google.com/maps/documentation/geocoding/overview

Hyder, R. T., Ahmed, S., & Ahmed, A. T. M. F. (2022). Evaluation of the clinical features of bronchiolitis in paediatric patients and exploration of the risk factors. *Journal of Monno Medical College*, *8*(2), 48–52.

INDEC – Instituto Nacional de Estadística y Censos. (2023). *Censo nacional de población, hogares y viviendas 2022: Resultados provisionales* (1a ed.). Ciudad Autónoma de Buenos Aires, Instituto Nacional de Estadística y Censos – INDEC. https://www.indec.gob.ar/ftp/cuadros/poblacion/cnphv2022_resultados_provisionales.pdf

INDEC – Instituto Nacional de Estadísticas y Censos. (2010). *Necesidades Básicas Insatisfechas en la República Argentina*. https://www.indec.gob.ar/indec/web/Nivel4-Tema-4-47-156

INDEC – Instituto Nacional de Estadísticas y Censos. (2015). *Unidades Geoestadísticas*. Cartografía y códigos geográficos del Sistema Estadístico Nacional. http://www.indec.gov.ar/codgeo.asp

INER – Instituto Nacional de Enfermedades Respiratorias "Emilio Coni". Administración Nacional de Laboratorios e Institutos de Salud (ANLIS) Ministerio de Salud de la Nación. (2018). *Actualización de la situación de la mortalidad por Enfermedades Respiratorias en Menores de 5 años en Argentina. Buenos Aires*. http://www.anlis.gov.ar/iner/wp-content/uploads/2020/02/Boletin-Actualizacion-2018-Enfermedades-Respiratorias-en-Menores-de-5-A%C3%B1os-1.pdf

Islam, M., Sultana, Z. Z., Iqbal, A., Ali, M., & Hossain, A. (2021). Effect of in-house crowding on childhood hospital admissions for acute respiratory infection: A matched case-control study in Bangladesh. *International Journal of Infectious Diseases*, *105*, 639–645. https://doi.org/10.1016/j.ijid.2021.03.002

Kaminker, S. A. (2015). Segregación residencial y proyectos de ciudad: Puerto Madryn como espacio en disputa. In H. Vessuri & G. Bocco (Eds.), *Conocimiento, paisaje, territorio: Procesos de cambio individual y colectivo* (1st ed.). Ediciones Universidad Nacional de la Patagonia Austral.

Kaminker, S. A. (2016). *Segregación Residencial en Puerto Madryn, Chubut (1991–2010). Formas y efectos de una urbanización acelerada en una ciudad intermedia de la Patagonia Central*. Instituto de Altos Estudios Sociales, Universidad Nacional de San Martín (IDAES, UNSAM). https://ri.unsam.edu.ar/handle/123456789/924

Kaminker, S. A. (2020). *Desigualdad, pobreza y construcción de la Ciudad en la Patagonia Central: Puerto Madryn, Chubut (1991–2010). Documento N° 6/2020. Secretaría de Investigación*. Instituto de Altos Estudios Sociales, Universidad Nacional de San Martín (IDAES, UNSAM).

Kawachi, I., & Berkman, L. F. (2003). Introduction. In I. Kawachi & L. F. Berkman (Eds.), *Neighborhoods and health*. Oxford University Press.

Mage, D. T., Latorre, M. L., Jenik, A. G., & Donner, E. M. (2016). An acute respiratory infection of a physiologically anemic infant is a more likely cause of SIDS than neurological prematurity. *Frontiers in Neurology*, *7*, 1–11. https://doi.org/10.3389/fneur.2016.00129

Mahishale, V. (2014). Climate change and respiratory health: Time to act!! *Journal of the Scientific Society*, *41*(3), 149. https://doi.org/10.4103/0974-5009.141196

Marmot, M. (2005). Social determinants of health inequalities. *The Lancet*, *365*(9464), 1099–1104. https://doi.org/10.1016/s0140-6736(05)71146-6

Marmot, M. G. (2003). Understanding social inequalities in health. *Perspectives in Biology and Medicine*, *46*(3 Suppl), S9–S23.

McKay, A., & Perge, E. (2015). Spatial inequality and its implications for growth–poverty-reduction relations. In A. McKay & E. Thorbecke (Eds.), *Economic growth and poverty reduction in sub-Saharan Africa: Current and emerging issues* (pp. 197–226). Oxford Academic. https://doi.org/10.1093/ACPROF:OSO/9780198728450.003.0007

Meissner, H. C. (2016). Viral bronchiolitis in children. *The New England Journal of Medicine*, *374*(1), 62–72. https://doi.org/10.1056/NEJMra1413456

MINSAL – Ministerio de Salud de la Nación. (2010). *Estadísticas Vitales. Información básica – año 2009. Serie 5 – Número 53*. Buenos Aires, Dirección de Estadísticas e Información de Salud. https://www.argentina.gob.ar/sites/default/files/serie5nro53.pdf

MINSAL – Ministerio de Salud de la Nación. (2018). *Dirección Nacional de Epidemiología. Boletín Integrado de Vigilancia N°*

395 – SE 03 (pp. 1–93). http://www.msal.gob.ar/images/stories/boletines/BIV_395_SE03.pdf

MINSAL – Ministerio de Salud de la Nación. Sistema Estadístico de Salud. (2007). *Egresos de Establecimientos Oficiales por Diagnóstico. 2005. Serie 11 – Número 1*. Buenos Aires, Dirección de Estadísticas e Información de Salud. https://www.argentina.gob.ar/sites/default/files/serie11nro1.pdf

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17–23.

Nair, H., Simões, E. A., Rudan, I., Gessner, B. D., Azziz-Baumgartner, E., Zhang, J. S. F., Feikin, D. R., Mackenzie, G. A., Moïisi, J. C., Roca, A., Baggett, H. C., Zaman, S. M., Singleton, R. J., Lucero, M. G., Chandran, A., Gentile, A., Cohen, C., Krishnan, A., Bhutta, Z. A., … Severe Acute Lower Respiratory Infections Working Group. (2013). Global and regional burden of hospital admissions for severe acute lower respiratory infections in young children in 2010: A systematic analysis. *Lancet*, *381*(9875), 1380–1390. https://doi.org/10.1016/S0140-6736(12)61901-1

Navazo, B., Dahinten, S. L., & Oyhenart, E. E. (2018). Malnutrición y pobreza estructural. Comparación de dos cohortes de escolares de Puerto Madryn, Argentina. *Revista de Salud Pública*, *20*(1), 60–66.

Obando-Pacheco, P., Justicia-Grande, A. J., Rivero-Calle, I., Rodríguez-Tenreiro, C., Sly, P., Ramilo, O., Mejías, A., Baraldi, E., Papadopoulos, N. G., Nair, H., Nunes, M. C., Kragten-Tabatabaie, L., Heikkinen, T., Greenough, A., Stein, R. T., Manzoni, P., Bont, L., & Martinón-Torres, F. (2018). Respiratory syncytial virus seasonality: A global overview. *The Journal of Infectious Diseases*, *217*(9), 1356–1364. https://doi.org/10.1093/infdis/jiy056

Oliva, J., & Terrier, O. (2021). Viral and bacterial co-infections in the lungs: Dangerous liaisons. *Viruses*, *13*(9), 1725. https://doi.org/10.3390/v13091725

Osayomi, T., Ogbonnaiye, O. B., & Iyanda, A. E. (2020). Hotspots and drivers of acute respiratory infection among children in Nigeria. *South African Journal of Child Health*, *14*(4), 224. https://doi.org/10.7196/sajch.2020.v14i4.01734

Patz, J. A., Frumkin, H., Holloway, T., Vimont, D. J., & Haines, A. (2014). Climate change. *JAMA*, *312*(15), 1565–1580. https://doi.org/10.1001/jama.2014.13186

Paynter, S., Ware, R. S., Weinstein, P., Williams, G., & Sly, P. D. (2010). Childhood pneumonia: A neglected, climate-sensitive disease? *The Lancet*, *376*(9755), 1804–1805. https://doi.org/10.1016/s0140-6736(10)62141-1

Pica, N., & Bouvier, N. M. (2012). Environmental factors affecting the transmission of respiratory viruses. *Current Opinion in Virology*, *2*(1), 90–95. https://doi.org/10.1016/j.coviro.2011.12.003

Pickett, K. E., & Pearl, M. (2001). Multilevel analyses of neighborhood socioeconomic context and health outcomes: A critical review. *Journal of Epidemiology and Community Health*, *55*(2), 111–122. https://doi.org/10.1136/jech.55.2.111

Pitzer, V. E., Viboud, C., Alonso, W. J., Wilcox, T., Metcalf, C. J., Steiner, C. A., Haynes, A. K., & Grenfell, B. T. (2015). Environmental drivers of the spatiotemporal dynamics of respiratory syncytial virus in the United States. *PLoS Pathogens*, *11*(1), e1004591. https://doi.org/10.1371/journal.ppat.1004591

Rey, S. J., & Anselin, L. (2010). PySAL: A python library of spatial analytical methods. In *Handbook of applied spatial analysis* (pp. 175–193). Springer.

Rey, S. J., Arribas-Bel, D., & Wolf, L. J. (2021). *Geographic data science with Python and the PyData stack*. CRC Press.

Rojas, F. (2007). Poverty determinants of acute respiratory infections among Mapuche indigenous peoples in Chile's ninth region of Araucania, using GIS and spatial statistics to identify health disparities. *International Journal of Health Geographics*, *6*, 26. https://doi.org/10.1186/1476-072X-6-26

Sassone, S., González, M., & Matossian, B. (2011). Ciudades patagónicas de la Argentina: Atracción, crecimiento y diversidad migratoria. In *Olimpíada de Geografía de la República Argentin*. (pp. 251–264). Universidad Nacional del Litoral.

Singer, M. (2013). Respiratory health and ecosyndemics in a time of global warming. *Health Sociology Review*, *22*(1), 98–111. https://doi.org/10.5172/hesr.2013.22.1.98

Singer, M., Bulled, N., Ostrach, B., & Lerman Ginzburg, S. (2021). Syndemics: A cross-disciplinary approach to complex epidemic events like COVID-19. *Annual Review of Anthropology*, *50*, 41–58. https://doi.org/10.1146/annurev-anthro-100919-121009

Sloan, C., Moore, M. L., & Hartert, T. (2011). Impact of pollution, climate, and sociodemographic factors on spatiotemporal dynamics of seasonal respiratory viruses. *Clinical and Translational Science*, *4*(1), 48–54. https://doi.org/10.1111/j.1752-8062.2010.00257.x

Solari, C. D., & Mare, R. D. (2012). Housing crowding effects on children's wellbeing. *Social Science Research*, *41*(2), 464–476. https://doi.org/10.1016/j.ssresearch.2011.09.012

Thayer, Z., Uwizeye, G., & McKerracher, L. (2022). Toolkit article: Approaches to measuring social inequities in health in human biology research. *American Journal of Human Biology*, *34*(12), e23804. https://doi.org/10.1002/ajhb.23804

Tuñón, I. (2019). Infancia(s). Progresos y retrocesos en clave de desigualdad. In *Documento estadístico. Barómetro de la Deuda Social Argentina. Serie Agenda para la Equidad (2017-2025)* (pp. 1–111). Edición Para Fundación Universidad Católica Argentina.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Velázquez, G., & Linares, S. (2008). Análisis de Autocorrelación espacial en variables de bienestar en la Argentina (1991-2001). *Vegueta*, *10*, 131–144.

World Health Organization. (2018). *WHO housing and health guidelines*. World Health Organization. https://www.ncbi.nlm.nih.gov/books/NBK535289/

Zheng, Z., Warren, J. L., Shapiro, E. D., Pitzer, V. E., & Weinberger, D. M. (2022). Estimated incidence of respiratory hospitalizations attributable to RSV infections across age and socioeconomic groups. *Pneumonia*, *14*, 6. https://doi.org/10.1186/s41479-022-00098-x

How to cite this article: Pazos, B. A., Morales, A. L., Ramallo, V., González-José, R., de Azevedo, S., & Taire, D. L. (2023). Mapping spatial morbidity patterns for bronchiolitis related to socioeconomic estimators: A spatial epidemiology approach to identify health disparities in Puerto Madryn, Argentina. *American Journal of Human Biology*, e23938. https://doi.org/10.1002/ajhb.23938

Received: 15 February 2023 | Revised: 11 August 2023 | Accepted: 23 August 2023

DOI: 10.1002/bdr2.2248

RESEARCH ARTICLE

Birth Defects Research | Society for Birth Defects Research & Prevention | WILEY

"Prevalence and secular trend of neural tube defects in fetal deaths in Argentina, 1994–2019"

Ruben Bronberg[1,2] | **Martinez Jorge**[3] | **Morales Leonardo**[4] | **Ruderman Anahi**[5] | **Taire Damian**[6] | **Ramallo Virginia**[5] | **Dipierri Jose**[3]

[1]Medical and Population Genetics Area, Ramos Mejía Hospital, Buenos Aires, Argentina

[2]Buenos Aires Government Research Committee, Buenos Aires, Argentina

[3]Institute of Ecoregions of the Andes, National Scientific and Technical Research Council (CONICET), Jujuy, Argentina

[4]Patagonian Institute of Social and Human Sciences (IPCSH CCT CONICET-CENPAT) and Department of Informatics, Faculty of Engineering National University of Patagonia San Juan Bosco (UNPSJB), Argentina

[5]Patagonian Institute of Social and Human Sciences (IPCSH CCT CONICET-CENPAT), Argentina

[6]Andrés Ísola Zonal Hospital, Puerto Madryn-Patagonian Institute of Social and Human Sciences (IPCSH CCT CONICET-CENPAT), Argentina

Correspondence
Ruben Bronberg, Medical and Population Genetics Area, Ramos Mejía Hospital, Buenos Aires, Argentina.
Email: rabronberg@intramed.net

Abstract

Background: Fetal deaths are a major source of information on the epidemiology of neural tube defects (NTDs; anencephaly and myelomeningocele). We analyzed NTDs prevalence and secular trend using fetal death records between 1994 and 2019 in Argentina.

Materials and Methods: Data were obtained from the Department of Statistics and Information of the Ministry of Health (DEIS). Using the number of fetal deaths due to anencephaly and myelomeningocele, we estimated the proportion of all fetal deaths due to anencephaly, myelomeningocele, and NTDs (anencephaly + myelomeningocele) during pre- and post-fortification period in Argentina. We also estimated the ratio of fetal deaths due to anencephaly, myelomeningocele, and NTDs (anencephaly + myelomeningocele) to 10,000 live births. Secular trend in the outcomes was analyzed using a Poisson model and Joinpoint regression analysis.

Results: In the entire period analyzed, the NTD proportion on fetal deaths was 1.32. In 1994, NTDs accounted for 34.7% of congenital malformations fetal deaths (CM) and 1.7% of all fetal deaths, whereas in 2019, these percentages were 9.4% and 0.5%, respectively. NTDs present a negative secular trend ($p < .05$). The risk of fetal death due to anencephaly and myelomeningocele decreases between 2005 and 2019 by 67% and 51% respectively ($p < .05$) in comparison to the period between 1994 and 2004 before the effective fortification of wheat flour used in the food industry destined for the domestic market.

Discussion and Conclusion: We found a significant decrease in the risk of all fetal deaths due to NTDs, particularly anencephaly, in Argentina over the study period, with most reduction observed during the mandatory flour fortification era (introduced in Argentina in 2002). The inclusion of fetal deaths in NTD surveillance, coupled or uncoupled with other pregnancy outcomes, is essential for monitoring preventive supplementation measures.

KEYWORDS

anencephaly, Argentina, fetal deaths, myelomeningocele, neural tube defects

1 | INTRODUCTION

Neural tube defects (NTDs) are a group of severe malformations associated with significant perinatal mortality and morbidity, and long-term disability (Blencowe et al., 2018). The two most common NTDs are anencephaly and myelomeningocele.

The way in which pregnancy outcomes, especially live births and birth defects are monitored in Argentina differs from that of other countries that have reported data on the frequency of NTDs. Specifically, in Argentina, fetal and infant deaths are mandatory and uniformly reported. The occurrence of congenital malformations is included in the cause of death information. For live births, congenital malformations are not reported on birth certificates and their notification to the National Registry of Congenital Anomalies (RENAC, 2021), which covers less than half of private sector births and about 59% of public sector births, is not mandatory. In death certificates for children under 1 year of age, the most frequent is anencephaly, characterized by the absence of the skull and both cerebral hemispheres produced by failure of the rostral neural tube to close between 23 and 25 days of gestation (Bronberg et al., 2011). Less severe and more frequent in newborns is myelomeningocele (RENAC, 2021). Pregnancies with NTD may end in spontaneous abortion, fetal death, elective termination of pregnancy, or an affected live newborn.

Many, but not all, NTDs can be prevented by periconceptional administration of folic acid, which reduces their prevalence by 50%–70% depending on the population (Blencowe et al., 2010). In Argentina, Law No. 25630—enacted in 2002 and regulated in 2003—requires mandatory fortification of locally commercialized wheat flour with iron, folic acid, thiamine, riboflavin, and niacin. According to the Global Fortification Data Exchange the proportion of industrially processed wheat flour was 100%, the population covered is 56.7% and the daily food availability was 45 g/c/d; https://fortificationdata.org/country-fortification-dashboard/?alpha3_code=ARG&lang=en).

There are national records on the prevalence of anencephaly and the effect of fortification on infant deaths (<1 year of age) and newborns. These results come from different studies and registries and use different designs. Based on the death certificates of children less than 1 year of age for the period 1998–2007 in Argentina, Bronberg et al. (2011) estimate a 53% reduction in mortality risk due to anencephaly at the national level after fortification, with regional variations. Based on data from the ECLAMC (Latin American Collaborative Study of Congenital Malformations) concerning to Argentina and the National Registry of Congenital Anomalies (RENAC), Bidondo et al. (2015) compared the prevalence observed in live births during the post-fortification period with that reported in the pre-fortification period (López-Camelo et al., 2010) and observed a significant decrease of 66% for anencephaly and encephalocele and 47% for spina bifida.

The specific prevalence of NTDs in fetal deaths and the effect of fortification with folic acid on their occurrence are unknown in Argentina. The prevalence of NTDs in fetal deaths is usually calculated by adding these to the prevalence in live newborns (Calvo & Biglieri, 2008).

Following the international recommendations of the Pan American Health Organization (PAHO) and the United Nations (UN), the Directorate of Health Statistics and Information (DEIS), Ministry of Health of the Nation defines fetal death as "death prior to the expulsion or extraction from its mother of a product of conception, irrespective of the duration of the pregnancy; the death is indicated by the fact that after such separation, the fetus does not breathe or show any other evidence of life, such as a beating of the heart, pulsation of the umbilical cord, or definite movement of voluntary muscles. Therefore, it is a vital event different from death, given that it occurs to those who have not reached live birth" (DEIS, 2019; PAHO, 2008; UN, 2014). This is the main characteristic of the input data on which the estimates of the present work are based, meeting the guidelines for accurate and transparent health estimates reporting (Stevens et al., 2016).

The objective is to analyze the proportion and prevalence of NTDs and their secular trend from the records of fetal deaths that occurred in Argentina between 1994 and 2019.

2 | MATERIALS AND METHODS

The data for this retrospective descriptive epidemiological study was provided by the DEIS and came from the certificates of live births and fetal deaths for the period between 1994 and 2019.

The variables used were: the absolute number of fetal deaths (18 weeks onwards), the total number of live newborns, and fetal deaths due to anencephaly and myelomeningocele coded according to the International Classification of Diseases, tenth revision ICD-10, and ICD-9. For the conversion between ICD-10 and ICD-9, the ICD Converter (http://www.conversorcie.com/index.php) was used.

Based on these data, the following quantities were calculated for Argentina as a whole, regardless of sex: (a) proportions of fetal deaths due to anencephaly, myelomeningocele, and NTD (anencephaly and myelomeningocele combined) over total fetal deaths; (b) prevalence

due to anencephaly, myelomeningocele, and NTD among fetal deaths per 10,000 live newborns.

The secular trend of NTD proportions and rates was analyzed using a Poisson model. To identify changes in mortality rate trends, join point regression was used for fetal deaths with anencephaly, myelomeningocele, and NTD using the Join Point Regression Program, Version 4.5.0.1 (Statistical Research and Applications Branch, National Cancer Institute). This method identifies the year(s) when a trend change is produced, calculates the annual percentage change (APC) in rates between trend-change points, and estimate the significance of the change in trend.

3 | RESULTS

In the period considered, 2368 fetal deaths due to NTDs were detected, the most frequent being anencephaly (90%), followed by myelomeningocele (7%). Overall, NTDs accounted for 1.3% of the 178,227 fetal deaths registered between 1994 and 2019 (Table 1).

At the beginning of the period, NTDs accounted for 34.7% of all malformations among fetal deaths and 1.7% of all fetal deaths. At the end of the period, these percentages were 9.4% for all malformations and 0.5% for fetal deaths.

Considering the entire study period (1994–2019) and according to the applied mathematical model of Poisson regression, a statistically significant decrease is observed for anencephaly (7.2% per year), for myelomeningocele (3.2% per year), for the total of congenital malformations (2.1% per year), and for the total of fetal deaths from all causes (2.2% per year). When the post-folic acid fortification period (2005–2019) is compared to the pre-fortification period (1994–2004), the risk of fetal death shows a statistically significant decrease of 67% due to anencephaly and a statistically significant decrease of 51% due to myelomeningocele, while all congenital malformations and fetal deaths decrease by 31% and 29% respectively (Table 2).

The Joinpoint analysis shows (Figures 1, 2), in the case of the proportions, for anencephaly, four APCs, of which only two are significant in the 1994–2003 segment (increase) and in the 2009–2019 segment (decrease). NTD presents two significant APCs in the 2003–2006 and 2009–2019 segments, both decreasing. Myelomeningocele exhibits one significant APC in the 1994–2003 segment (increase). Regarding prevalences*10,000 NB anencephaly exhibits two statistically significant declining APCs in the 2003–2006 and 2009–2019 segments; NTD exhibits three statistically significant APCs 1997–2003 (increase), 2003–2006 and 2009–2019 both declining, myelomeningocele 1 APC in the 1994–2003 segment (increase).

4 | DISCUSSION

Epidemiological information on NTD fetal deaths in Argentina covering the period from 1994 to 2019 is provided. Usually, the prevalence of NTD fetal deaths is included and calculated within infant deaths, so there is no clear assessment of the magnitude and impact of NTDs on fetal and infant deaths, even in Argentina (Atta et al., 2015; Calvo & Biglieri, 2008; Lo et al., 2014; Zaganjor et al., 2016). Therefore, this analysis marks the only epidemiological precedent in the country on fetal deaths due to NTDs.

According to the guidelines provided by the GATHER Group (Guidelines for Accurate and Transparent Health Estimates Reporting; Stevens et al., 2016), epidemiological studies must provide information on the data sources included and their main characteristics (name/institution of contact, the population represented, data collection method, year(s) of collection, etc.). In Argentina, data sources on the prevalence of NTDs in fetal deaths, infant deaths, and newborns are clearly differentiated. Fetal and infant deaths are compulsorily and systematically collected through the corresponding certificates that are processed by the DEIS (Dirección de Estadísticas e Información de la Salud, Ministerio de Salud; https://www.argentina.gob.ar/salud/deis). The DEIS depends on the Sistema Estadístico Nacional created by Law 17622/68, therefore the coverage of this information is that of a census (100% of the population) and it has a temporal depth that extends to the 1970s. In this work, only the digitized information available as of 1994 was used. On the other hand, congenital malformations at birth are not recorded on the birth certificate, but by the RENAC (Registro Nacional de las Anomalías Congénitas), which is not mandatory, it started in 2009 and has a coverage of 38.58% in the private sector and one of 59.15% in the public sector (RENAC, 2019). Therefore, there may be differences and uncertainties regarding the prevalence of NTDs due to these characteristics of the sources.

In relation to folic acid fortification, we observe, as in infant deaths (Bronberg et al., 2011; Calvo & Biglieri, 2008) and in newborns (Bidondo et al., 2015; López-Camelo et al., 2010), a statistically significant negative secular trend after fortification, but of greater magnitude with respect to infant deaths (53% vs. 66%). However, comparisons between sources should be made with caution considering the period covered by each source and, as will be discussed below, the numerator to which they are referred. The observed periods are different from previous precedents on NTD epidemiology and generally shorter (Bidondo et al., 2015; Bronberg et al., 2011; Calvo & Biglieri, 2008; López-Camelo

TABLE 1 Fetal deaths from anencephaly, myelomeningocele, neural tube defects, congenital malformations, and live newborns in Argentina during the period 1994–2019.

Year	Fetal deaths from anencephaly	Fetal deaths from myelomeningocele	Fetal deaths from NTDs (anencephaly +myelomeningocele)	Fetal deaths from congenital defects	Total fetal deaths	Livebirths	Proportion of NTDs fetal deaths among all congenital defects fetal deaths (IC95%)	Proportion of NTD fetal deaths among all fetal deaths/(IC95%)	Prevalence of fetal deaths with NTDs (per 10,000 livebirths (IC95%)
1994	146	9	155	447	8893	673,787	34.7 (29.4–40.6)	1.7 (1.5–2.0)	2.3 (2.0–2.7)
1995	119	8	127	330	8201	658,735	38.5 (32.1–45.8)	1.5 (1.3–1.8)	1.9 (1.6–2.3)
1996	124	6	130	405	8472	675,437	32.1 (26.8–38.1)	1.5 (1.3–1.8)	1.9 (1.6–2.3)
1997	122	7	129	441	7994	692,357	29.3 (24.4–34.8)	1.6 (1.4–1.9)	1.9 (1.6–2.2)
1998	127	3	130	420	8030	683,301	31 (25.9–36.8)	1.6 (1.4–1.9)	1.9 (1.6–2.3)
1999	125	9	134	430	7955	686,748	31.2 (26.1–36.9)	1.7 (1.4–2.0)	2 (1.6–2.3)
2000	133	9	142	448	7622	701,878	31.7 (26.7–37.4)	1.9 (1.6–2.2)	2 (1.7–2.4)
2001	160	13	173	525	7912	683,495	33 (28.2–38.2)	2.2 (1.9–2.5)	2.5 (2.2–2.9)
2002	169	13	182	467	7881	694,684	39 (33.5–45.1)	2.3 (2.0–2.7)	2.6 (2.3–3.0)
2003	159	12	171	444	7459	697,952	38.5 (33.0–44.7)	2.3 (2.0–2.7)	2.5 (2.1–2.9)
2004	105	9	114	436	7190	736,261	26.1 (21.6–31.4)	1.6 (1.3–1.9)	1.5 (1.3–1.9)
2005	70	5	75	350	6529	712,220	21.4 (16.9–26.9)	1.1 (0.9–1.4)	1.1 (0.8–1.3)
2006	56	3	59	313	6084	696,451	18.8 (14.3–24.3)	1 (0.7–1.3)	0.8 (0.7–1.1)

TABLE 1 (Continued)

Year	Fetal deaths from anencephaly	Fetal deaths from myelomeningocele	Fetal deaths from NTDs (anencephaly +myelomeningocele)	Fetal deaths from congenital defects	Total fetal deaths	Livebirths	Proportion of NTDs fetal deaths among all congenital defects fetal deaths (IC95%)	Proportion of NTD fetal deaths among all fetal deaths/(IC95%)	Prevalence of fetal deaths with NTDs (per 10,000 livebirths (IC95%)
2007	50	5	55	277	5878	700,792	19.9 (15.0–25.8)	0.9 (0.7–1.2)	0.8 (0.6–1.0)
2008	48	4	52	287	6091	746,460	18.1 (13.5–23.8)	0.9 (0.6–1.1)	0.7 (0.5–0.9)
2009	75	4	79	352	6137	756,176	22.4 (17.8–28.0)	1.3 (1.0–1.6)	1 (0.8–1.3)
2010	54	5	59	346	5735	758,042	17.1 (13.0–22.0)	1 (0.8–1.3)	0.8 (0.6–1.0)
2011	58	2	60	327	5864	758,042	18.3 (14.0–23.6)	1 (0.8–1.3)	0.8 (0.6–1.0)
2012	60	4	64	347	6220	783,318	18.4 (14.2–23.6)	1 (0.8–1.3)	0.8 (0.6–1.0)
2013	43	4	47	325	6333	754,603	14.5 (10.6–19.2)	0.7 (0.6–1.0)	0.6 (0.5–0.8)
2014	32	6	39	327	6443	777,012	11.9 (8.5–16.3)	0.6 (0.4–0.8)	0.5 (0.4–0.7)
2015	42	7	49	351	6158	770,040	14 (10.3–18.5)	0.8 (0.6–1.1)	0.6 (0.5–0.8)
2016	33	7	40	329	6075	728,035	12.2 (87–16.6)	0.7 (0.5–0.9)	0.5 (0.4–0.8)
2017	37	5	42	312	5880	704,609	13.5 (9.7–18.2)	0.7 (0.5–1.0)	0.6 (0.4–08)
2018	32	4	36	286	5897	685,394	12.6 (8.8–17.4)	0.6 (0.4–0.9)	0.5 (0.4–0.7)
2019	20	5	25	267	5294	685,394	9.4 (6.1–13.8)	0.5 (0.3–0.7)	0.4 (0.2–05)

TABLE 2 Secular trend and risk of NTDs, malformations, and fetal deaths.

	Secular trend 1994–2019	Relative risk ^{POS}	Proportional change ^{POS}
Anencepahly-associated fetal deaths	−0.072 *	0.33 *	↓ 67%
Myelomeningocele-associated fetal deaths	−0.032 *	0.49 *	↓51%
NTD-associated fetal deaths	−0.069 *	0.34 *	↓66%
Congenital defects-associated fetal deaths	−0.021 *	0.69 *	↓31%
All fetal deaths	−0.022 *	0.71 *	↓29%

Note: Reference *statistical significance 0.005; POS risk and percentage change of the post-fortification period (2005–2019) in relation to the pre-fortification baseline period (1995–2004); NTDs only include anencephaly and myelomeningocele.

et al., 2010). After the sharp drop in fetal deaths during the immediate post-fortification period between 2003 and 2006, the prevalence of NTDs, particularly anencephaly, continues to decline monotonically, representing only 9.4% of fetal congenital malformations in 2009 in contrast with the 34.7% prevalence it presented in 1994. When considering the rate of anencephaly*10,000 live newborns, the difference between the beginning and the end of the period is 17%.

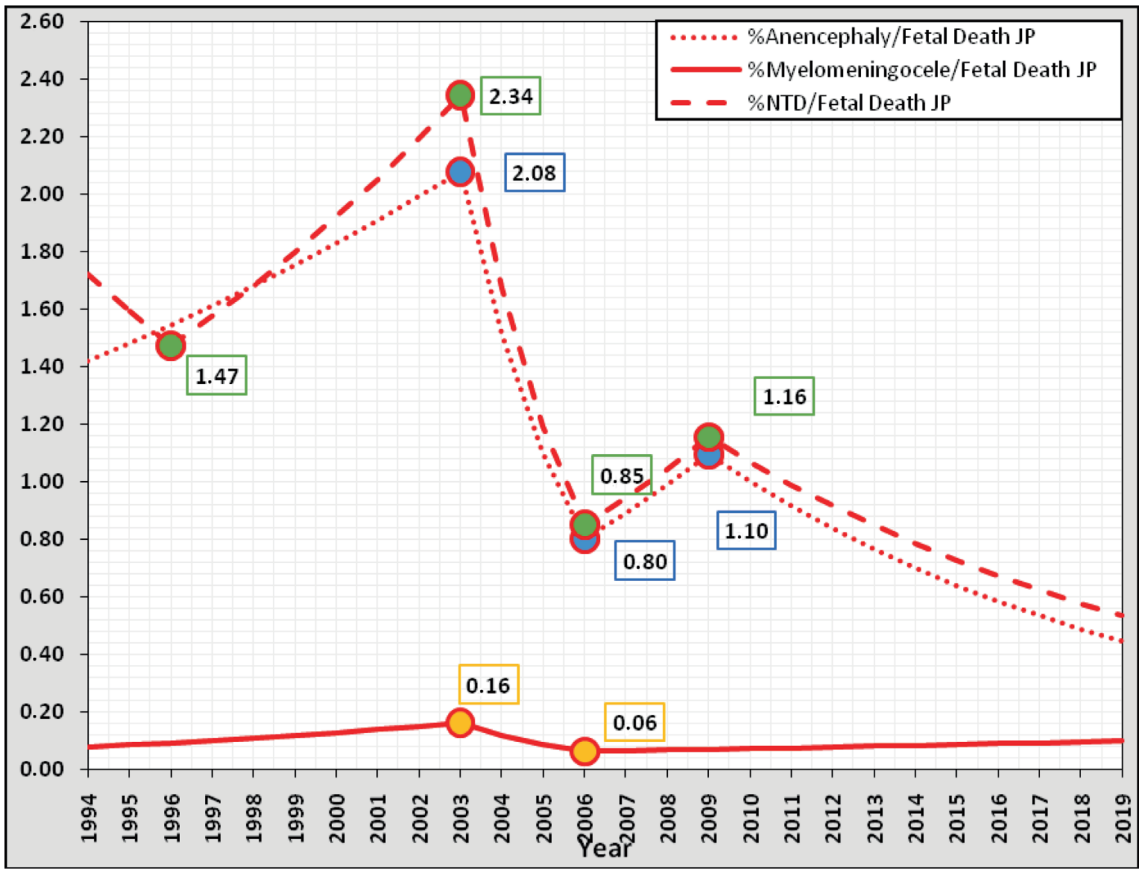
Between 1994 and 2019, a substantial decrease in fetal deaths was also observed (from 8893 in 1994 to 5294 in 2019) which can be interpreted as a statistically significant negative secular trend, with a 29% decrease in the risk of fetal death. It can be assumed that a large part of this decline corresponds to the decrease in the occurrence of NTDs.

Of all the reviews on the global prevalence of NTDs (Atta et al., 2015; Lo et al., 2014; Zaganjor et al., 2016), the only one that provides information on the estimation of NTDs in fetal deaths by the number of live newborns in the same year is that of Blencowe et al. (2018). For Latin America and the Caribbean, these authors provide an average prevalence of NTDs in fetal deaths of 0.7 (95% CI 0.2–1.4) *10,000 live newborns. In 2019, the prevalence of NTDs in fetal deaths in Argentina was lower, at 0.4 (95%CI 0.2–0.5), not differing substantially from the Latin American reference.

An important difference in the prevalence of NTDs in fetal versus infant deaths and live newborns is the representation of each malformation. While in fetal deaths anencephaly ostensibly predominates, in live newborns the opposite situation occurs, with myelomeningocele being more frequent. Since spina bifida cases are not diagnosed by ultrasound or radiographic examination that might provide details on the extent of vertebral defects, a misclassification bias cannot be ruled out, especially in cases of spina bifida occulta without skin manifestations (de Wals et al., 2008). Cases of myelomeningocele are scarce in fetal deaths and in some years nonexistent, which is probably why it presents a negative secular trend with a decrease of 51% in the period. These considerations indicate that anencephaly would be an excellent marker of malformation to analyze the epidemiology of NTDs in fetal deaths.

Considering the different sources of information, differences also arise in terms of the estimator used, percentages, and rates per 10,000 live newborns, in fetal and infant deaths. Taking Bronberg et al. (2011) as a reference, if we compare fetal deaths due to congenital malformations with infant deaths due to this cause during the same period (1998–2007), we observe that the percentage of anencephaly over the total of congenital malformations is 22% for fetal deaths and 7% for infant deaths, which corroborates its higher frequency in fetal deaths. However, when considering the rate of anencephaly per 10,000 live newborns, it is 1.5 in fetal and infant. These comparisons show the importance of the choice of the denominator and, in this particular case, using the number of congenital malformations would be more evident and discriminating than using the number of live newborns because the number of anencephaly cases is relatively small compared to that of live newborns (Blencowe et al., 2018).

The elective termination of pregnancy law was promulgated in 2022 in Argentina, there are no data on the elective termination of anencephaly and myelomeningocele. The strength of this research about the epidemiology of NTDs lies in the fact that fetal deaths in Argentina are systematically registered by the DEIS—this registry has been improved over time by including other variables not analyzed in this case (maternal age, weight, gestational age, sex, place of residence, etc.), and that for anencephaly, death certificates have a predictive value of 100% and a sensitivity of 86% (Lydon-Rochelle et al., 2005; Tairou et al., 2006). There were no relevant limitations for this analysis, except for the different versions of the ICD that were used throughout the 26 years of fetal death registration.

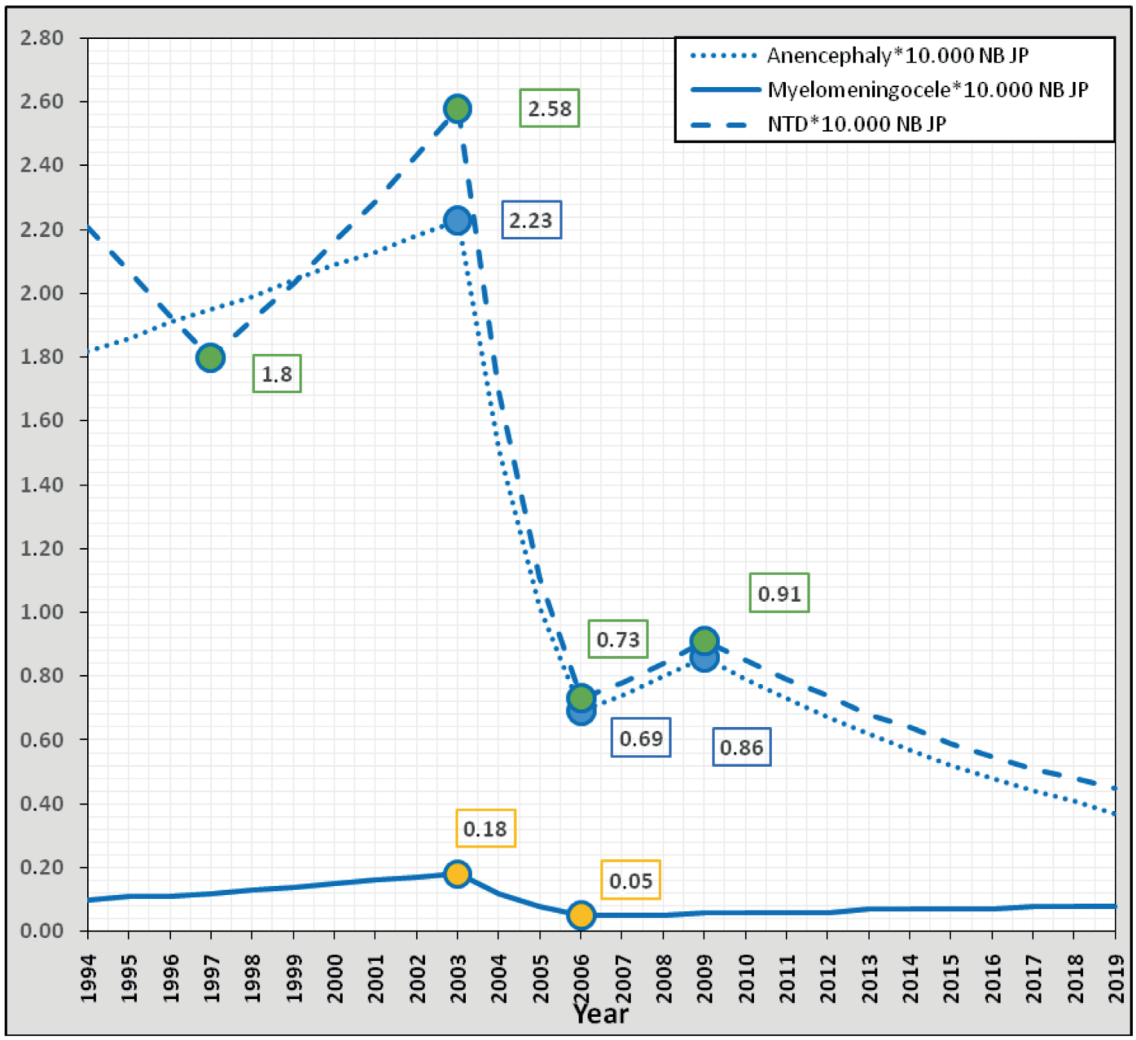


REFERENCES

Malformation	*APC***	*Lower CI*	*Upper CI*	*T test*
Anencephaly				
Segment 1994-2003	*4.3**	*1.9*	*6.8*	*3.8*
Segment 2003-2006	*-27.2*	*-48,1*	*2.3*	*-2*
Segment 2006-2009	*11*	*-28.9*	*73.2*	*0.5*
Segment 2009-2019	*-8.6**	*-12.1*	*-4.9*	*-4.9*
Neural tube defects				
Segment 1994-1997	*-7.5*	*-24.8*	*13.9*	*-0.8*
Segment 1997-2003	*6.8**	*3.3*	*10.5*	*4.3*
Segment 2003-2006	*-28.6**	*-44.9*	*-7.6*	*-2.8*
Segment 2006-2009	*10.7*	*-21*	*55.1*	*0.7*
Segment 2009-2019	*-7.4**	*-10.1*	*-4.7*	*-5.8*
Myelomeningocele				
Segment 1994-2003	*8.6**	*2.2*	*15.4*	*2.9*
Segment 2003-2006	*-27.6*	*-68.3*	*65.5*	*-0.8*
Segment 2006-2019	*3.7*	*-1.4*	*9.1*	*1.5*

*statistical significance 0.005; **APC: annual percentage change.

FIGURE 1 Joint point analysis of proportions of Anencephaly, Myelomeningocele and NTDs in Argentina during the period 1994–2019. *statistical significance 0.005; **APC, annual percentage change.



References

Malformation	*APC***	*Lower CI*	*Upper CI*	*T test*
Anencephaly				
Segment 1994-2003	*2.3*	*-0.4*	*5.1*	*1.8*
Segment 2003-2006	*-32.4**	*-54.3*	*-0.1*	*-2.1*
Segment 2006-2009	*7.8*	*-35.4*	*79.8*	*0.3*
Segment 2009-2019	*-8.1**	*-12.1*	*-3.8*	*-4*
Neural tube defects				
Segment 1994-1997	*-6.6*	*-16.2*	*4.2*	*-1.4*
Segment 1997-2003	*6.1**	*1.3*	*11.2*	*2.8*
Segment 2003-2006	*-34.4**	*-50.1*	*-13.9*	*-3.4*
Segment 2006-2009	*7.9*	*-24.3*	*53.7*	*0.5*
Segment 2009-2019	*-6.9**	*-9.7*	*-4*	*-5.1*
Myelomeningocele				
Segment 1994-2003	*6.6**	*0.1*	*13.5*	*2.1*
Segment 2003-2006	*-33.9*	*-71.9*	*55.5*	*-1*
Segment 2006-2019	*3.6*	*-1.7*	*9.2*	*1.4*

*statistical significance 0.005; **APC: annual percentage change.

FIGURE 2 Joint point analysis of prevalence of Anencephaly, Myelomeningocele, and NTDs in Argentina during the period 1994–2019. *statistical significance 0.005; **APC, annual percentage change.

5 | CONCLUSION

The analysis of fetal deaths in Argentina indicates that there is a significant decrease in the risk of NTD mortality, particularly anencephaly, as a result of the mandatory fortification of wheat flour with folic acid. The inclusion of fetal deaths, coupled or uncoupled with other pregnancy outcomes, in NTD surveillance is essential for the monitoring and investigation of these malformations. This is especially essential when malformations among live births are not comprehensively reported.

CONFLICT OF INTEREST STATEMENT

None of the authors have a conflict of interest to disclose.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Ruben Bronberg https://orcid.org/0000-0002-8859-3060

REFERENCES

Atta, C. A., Fiest, K. M., Frolkis, A. D., Jette, N., Pringsheim, T., St Germaine-Smith, C., Rajapakse, T., Kaplan, G. G., & Metcalfe, A. (2015). Global birth prevalence of spina bifida by folic acid fortification status: A systematic review and meta-analysis. *American Journal of Public Health*, *106*(1), e24–e34.

Bidondo, M. P., Liascovich, R., Barbero, P., & Groisman, B. (2015). Prevalence of neural tube defects and estimation of cases averted in the post-fortification period in Argentina. *Archivos Argentinos de Pediatría*, *113*(6), 498–501.

Blencowe, H., Cousens, S., Modell, B., & Lawn, J. (2010). Folic acid to reduce neonatal mortality from neural tube disorders. *International Journal of Epidemiology*, *39*(Suppl 1), 110–121.

Blencowe, H., Kancherla, V., Moorthie, S., Darlison, M. W., & Modell, B. (2018). Estimates of global and regional prevalence of neural tube defects for 2015: A systematic analysis. *Annals of the New York Academy of Sciences*, *1414*(1), 31–46.

Bronberg, R., Alfaro, E., Chaves, E., Andrade, A., Gili, J., López Camelo, J., & Dipierri, J. (2011). Anencephaly related infant mortality in Argentina: Spatial and temporal analysis (1998–2007). *Archivos Argentinos de Pediatría*, *109*(2), 117–123.

Calvo, E., & Biglieri, A. (2008). Impact of folic acid fortification on women's nutritional status and on the prevalence of neural tube defects. *Archivos Argentinos de Pediatría*, *106*(6), 492–498.

de Wals, P., Tairou, F., van Allen, M. I., Lowry, R. B., Evans, J. A., van den Hof, M. C., Crowley, M., Uh, S. H., Zimmer, P., Sibbald, B., Fernandez, B., Lee, N. S., & Niyonsenga, T. (2008). Spina bifida before and after folic acid fortification in Canada. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, *82*(9), 622–626.

Directorate of Health Statistics and Information (DEIS. (2019). Registration of causes of death and mortality statistics. Retrieved from http://www.deis.msal.gov.ar/wp-content/uploads/2019/09/resumen-sobre-certificacion-medica-de-causas-de-muerte.pdf

Lo, A., Polšek, D., & Sidhu, S. (2014). Estimating the burden of neural tube defects in low- and middle-income countries. *Journal of Global Health*, *4*(1), 010402.

López-Camelo, J. S., Castilla, E. E., & Orioli, I. M. (2010). Folic acid flour fortification: Impact on the frequencies of 52 congenital anomaly types in three south American countries. *American Journal of Medical Genetics. Part A*, *152A*(10), 2444–2458.

Lydon-Rochelle, M. T., Cárdenas, V., Nelson, J. L., Tomashek, K. M., Mueller, B. A., & Easterling, T. R. (2005). Validity of maternal and perinatal risk factors reported on fetal death certificates. *American Journal of Public Health*, *95*(11), 1948–1951.

National network of congenital anomalies of Argentina (RENAC). (2019). Epidemiological analysis on congenital anomalies in newborns, registered during 2018. Retrieved from http://www.anlis.gov.ar/cenagem/wp-content/uploads/2017/07/Reporte-RENAC-2019.pdf

Pan American Health Organization (PAHO). (2008). *International statistical classification of diseases and related health problems*, Tenth Revision (Vol. 2, p. 554).

RENAC. (2021). Análisis epidemiológico sobre las anomalías congénitas en recién nacidos, registradas durante 2020 en la República Argentina. Ministerio de Salud. https://www.ine.gov.ar/renac/Rep2021.pdf

Stevens, G. A., Alkema, L., Black, R. E., Boerma, J. T., Collins, G. S., Ezzati, M., Grove, J. T., Hogan, D. R., Hogan, M. C., Horton, R., Lawn, J. E., Marušić, A., Mathers, C. D., Murray, C. J., Rudan, I., Salomon, J. A., Simpson, P. J., Vos, T., Welch, V., & (The GATHER Working Group). (2016). Guidelines for accurate and transparent health estimates reporting: The GATHER statement. *Lancet*, *388*(10062), e19–e23.

Tairou, F., de Wals, P., & Bastide, A. (2006). Validity of death and stillbirth certificates and hospital discharge summaries for the identification of neural tube defects in Quebec City. *Chronic Diseases in Canada*, *27*(3), 120–124.

United nations (UN. (2014). Principles and recommendations for a vital statistics system. Statistical Reports. Serie M No.19/Rev.3. Retrieved from https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Principles_and_Recommendations/CRVS/M19Rev3-S.pdf

Zaganjor, I., Sekkarie, A., Tsang, B. L., Williams, J., Razzaghi, H., Mulinare, J., Sniezek, J. E., Cannon, M. J., & Rosenthal, J. (2016). Describing the prevalence of neural tube defects worldwide: A systematic literature review. *PLoS One*, *11*(4), e0151586.

How to cite this article: Bronberg, R., Jorge, M., Leonardo, M., Anahi, R., Damian, T., Virginia, R., & Jose, D. (2023). "Prevalence and secular trend of neural tube defects in fetal deaths in Argentina, 1994–2019". *Birth Defects Research*, 1–9. https://doi.org/10.1002/bdr2.2248

Received: 2 September 2022 | Revised: 16 December 2022 | Accepted: 17 January 2023

DOI: 10.1002/ajpa.24702

RESEARCH ARTICLE

AMERICAN JOURNAL OF BIOLOGICAL ANTHROPOLOGY
The Official Journal of the American Association of Biological Anthropologists

WILEY

Genetic and self-perceived ancestries in Argentina: Beyond the three-hybrid model

Anahí Ruderman[1,2] | Pierre Luisi[2,3] | Carolina Paschetta[1,2] | Tamara Teodoroff[1] | Luis Orlando Pérez[1,2] | Soledad de Azevedo[1,2] | Magda Alexandra Trujillo-Jiménez[1,2] | Pablo Navarro[1,2,4,5] | Leonardo Morales[1,2,4,5] | Bruno Pazos[1,2,4,5] | Rolando González-José[1,2] | Virginia Ramallo[1,2]

[1]Instituto Patagónico de Ciencias Sociales y Humanas, CCT CONICET CENPAT, Puerto Madryn, Chubut, Argentina

[2]Programa de Referencia y Biobanco Genómico de la Población Argentina. Secetaría de Planeamiento y Politicas, Ministerio de Ciencia, Tecnología e Innovación, Ciudad Autónoma de Buenos Aires, Argentina

[3]Departamento de Antropología, Facultad de Filosofía y Humanidades, Universidad Nacional de Córdoba, Córdoba, Córdoba, Argentina

[4]Laboratorio de Ciencias de las Imágenes, Departamento de Ingeniería Eléctrica y Computadoras, Universidad Nacional del Sur, Bahía Blanca, Buenos Aires, Argentina

[5]Departamento de Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco, Trelew, Chubut, Argentina

Correspondence
Anahí Ruderman and Rolando González-José, Instituto Patagónico de Ciencias Sociales y Humanas—CCT CONICET CENPAT, Puerto Madryn, Argentina.
Email: ruderman@cenpat-conicet.gob.ar; rolando@cenpat-conicet.gob.ar

Pierre Luisi, Departamento de Antropología, Facultad de Filosofía y Humanidades, Universidad Nacional de Córdoba, Córdoba, Argentina.
Email: pierre.luisi@unc.edu.ar

Present address
Pierre Luisi, Institut Pasteur, Université de Paris Cité, Paris, France.

Funding information
Consejo Nacional de Investigaciones Científicas y Técnicas, Grant/Award Number: PIP D.111/16; Fondo para la Investigación Científica y Tecnológica, Grant/Award Number: PICT 3206

Abstract

Objectives: The increased availability of genome-wide data allows capturing the fine genetic structure of present days populations. Here we analyze the genetic ancestry at a fine scale of an Argentinean Patagonia population to understand the origins beyond the three-hybrid model, and to compare these results with volunteers' self-perceived ancestry in a broad context encompassed by historical and familiar information.

Materials and Methods: We compare high-throughput genotyping data for 92 individuals that we generated to data sets from the literature by applying fully haplotype-based methods to examine patterns of human population substructure. The volunteers filled out a semi-structured questionnaire, including questions about their history, ancestors, and self-perceived ancestry. Finally, we used non-parametric tests in order to compare genomic ancestry against self-perception.

Results: Genetic ancestry from Iberian populations accounted for 0.176 (Spain and Basque origins), while the component associated with Italian populations accounted for 0.140. We observed a 0.169 Native American genetic ancestry. Participants significantly over- and under- self-perceived Native American and European origins, respectively. Components of origins from North Africa to Central South Asia accounted for 0.225 of the genetic ancestry in the sample, with significantly higher proportions for people that mentioned such origins in their genealogical history.

Discussion: We captured the fine-genetic architecture of a Puerto Madryn population sample in Chubut province, showing that self-perceived ancestry remains a poor proxy for genetic ancestry. The presence of North Africa to Central South Asia components and its correlate with self-perception of these origins justifies its inclusion in future miscegenation studies in Argentina.

KEYWORDS
admixture, ancestry, Argentinean Patagonia

Anahí Ruderman and Pierre Luisi contributed equally to this work

1 | INTRODUCTON

Concepts like race and genetic ancestry have distinct impacts on several aspects of modern life, including access to work, social policies, and public health key decisions such as drug administration, therapy design, and clinical trials, among others (Paschetta et al., 2021). The concept of race as a classificatory criterion for analyzing the biological diversity of the human species is more and more avoided in biological anthropology studies. Instead, geographic populations are more frequently used as units of analysis and comparison. Delimiting human populations under study represents a great challenge: various criteria are often applied, such as the current political division of the territory or the physical characteristics of the environment (e.g., populations from the Caucasus region, Sub-Saharan Africa, Amazonia, etc.). In other cases, populations are defined in terms of cultural aspects such as language or religion (e.g., Bantu-speaker groups or populations of Jewish origin, respectively). Whether in the context of research or state statistics, the choice of one criterion over another to define human populations is not innocuous but rather influences the entity or importance that population groups obtain in the social imaginary, which in turn has an impact on the collective process of identity construction.

Many studies aimed at determining the genetic ancestry of samples from different sub-populations across Latin America. Also widespread in Argentina, those studied leveraged a large range of molecular markers. Early approaches involved initial investigations on polymorphisms of the ABO and Rh systems in Buenos Aires (Etcheverry, 1947, 1949; Palazzo & Tenconi, 1939). These papers detected gene frequencies similar to those reported in subjects from Spain and Italy. Later, Palatnik (1966), revealed a gradient in the ABO*O allele distribution, with greater frequency in those regions with a large Indigenous presence. For the first time, the author used the labels of "Indigenous" and "European" components to highlight the parental populations.

In the following decades, several research groups have carried out genetic ancestry studies in urban and rural populations (Avena et al., 2003; Avena et al., 2010; Di Fabio Roca et al., 2011; Dipierri et al., 1999; López Camelo et al., 1996; Morales et al., 2000, among others). López Camelo et al. (1996) were the first to include the African genetic component in their analyses, thus increasing the complexity of the approaches under a three-hybrid model. Since then, a scheme of three-way admixture (i.e., with European, Native American, and African parental populations) was settled and remained identical in practically all miscegenation studies in Argentina (Carnese et al., 2011; García et al., 2015; Luisi et al., 2020; Parolin et al., 2012; Pauro et al., 2010; Salas et al., 2008, among others), not because all these works ignored the historical and current migratory processes from other origins, but because the limitations of the employed methods and/or availability of fine-scale data leveraged as reference made it difficult to prove.

Studies on genetic ancestry hardly ever include explicitly other origins, and if they do, they are not primary studies, using samples taken in Argentina to perform comparisons of trends at the country level (e.g., Ongaro et al., 2019). In general, the three-hybrid model has also been predominant in most countries in Latin America, and efforts to understand the recent demographic history that shaped genetic diversity have been scarce in the region (see Chacón-Duque et al., 2018). The recently increased availability of genome-wide data now offers the chance to capture the genetic complexity of present days Latin American populations at a finer scale.

As it is now well documented, since the conquest of the Americas and the establishment of colonies in the continent, a process of miscegenation began between the Native populations, the European conquerors (mostly Spanish in the case of Argentina), and the captured African populations brought to the region as slaves. Later, massive migratory movements contributed to modifying cultural, phenotypic, and genetic Latin American diversity. Between 1857 and 1960, 7,600,000 people arrived in Argentina from overseas (Devoto, 2007). At the beginning of the 20th century, only 54.5% of over 1,232,000 inhabitants of Buenos Aires city were born in Argentina. A large proportion of immigrants came from European countries such as Italy, Spain, Germany, Great Britain, Russia, Poland, France, and Slovenia, among others (Devoto, 2007; Margulis, 1977), but also from Levant (principally Lebanon, Syria), the Caucasus (Armenia), Asian, and North African countries (Egypt and Morocco) (Besteme, 1988). Indeed, the third largest immigration flow came from Syrian-Lebanese populations, following the flows from Italian and Spanish populations. Actually, Argentina was the South American country that received the highest percentage of Syrian-Lebanese migrants since the beginning of the 19th century.

Specifically, the province of Chubut, located in Argentinean Patagonia, was also characterized by receiving an important migration of Welsh origin during the second half of the 19th century (Margulis, 1977). From the second half of the 20th century onwards, the population that arrived in Chubut came from neighboring countries, as well as from other provinces of Argentina. During the 1960s–1980s there was a strong policy aimed to attract migrants from the North of the country to the southern provinces (Escudero & Rubilar, 2017). All the processes described above have influenced the genetic make-up of the current population, and their individual and collective identities.

Self-perceived and genetic ancestry are both complementary ways of approaching the history and identity of people. We define here self-perceived ancestry as the origins from particular human groups that an individual assumes he or she has through his or her ancestors. It is a dynamic and complex abstraction that can be influenced by several factors such as collective and family history, education, social and political context, physical appearance, and so forth. On the other hand, genetic ancestry estimates the origins relying on the inference of genetic relationships of an individual to representative members of particular human groups. A discrepancy between these two equally valid variables has already been described in Latin America (Paschetta et al., 2021; Ruiz-Linares et al., 2014). Comparing genetic and self-perceived ancestries in a population is relevant since classifications based on ancestry are now widespread in social and biomedical research, in some cases using genetic ancestry and

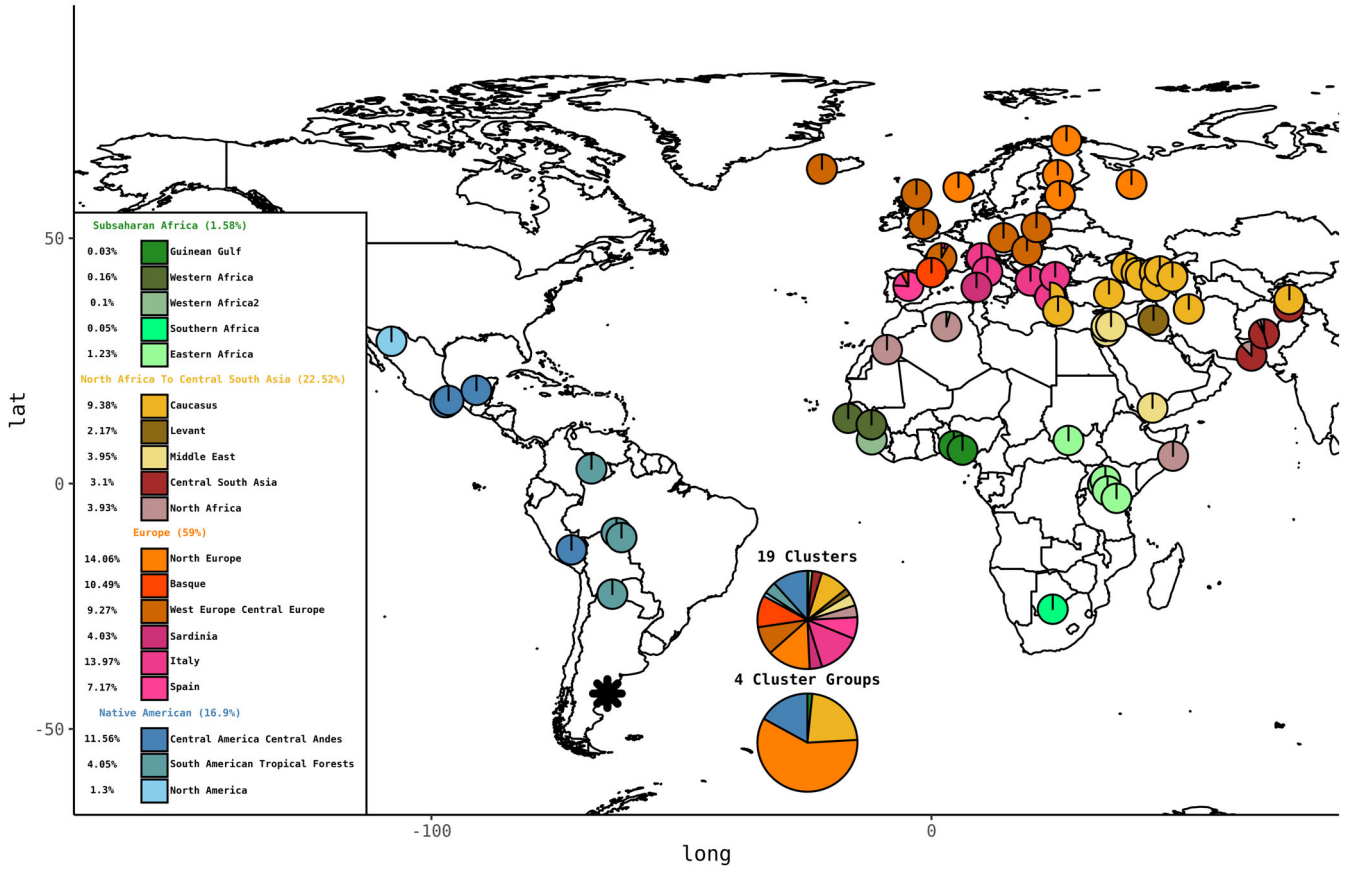


FIGURE 1 Haplotype-based genetic ancestry results. Small pie charts show the proportions of individuals from each reference population assigned to the different groups of donor clusters (see Supplementary Figure S3 and Supplementary Table S2 for the actual numbers). The big pie chart shows the estimates of genetic ancestry proportions for the Patagonian individuals from Puerto Madryn (location marked with an asterisk) when considering 19 donor clusters or by grouping those into four main groups (see also Supplementary Table S3).

self-reported ancestry as if they were synonyms. This is particularly important in populations with a long history of miscegenation, such as Argentina. Here, we aim at analyzing the genetic ancestry at a fine-scale of an Argentinean Patagonia population to understand the admixture complexity beyond the classical three-hybrid model, leveraging newly generated high-throughput genotyping data for 92 cosmopolitan individuals from Puerto Madryn city on the East Coast of the Chubut Province (see Figure 1). Also, we compare the volunteers inferred genetic ancestry estimates with their self-perceived ancestry to shed light on the complex process of identity as determined by a broad context of historical and familiar information.

2 | SUBJECTS AND METHODS

2.1 | Sample collection

During 2018, an open call for volunteers was made in Puerto Madryn city to participate in a study about complex diseases and the geographical origins of the population. The research project was publicized through social networks and local media inviting the community to participate. After being informed about the objectives and scope of the project, the volunteers who decided to participate signed a free and informed consent form previously approved by the local Bioethics Committee depending on the Ministry of Health of the Chubut Province. Before submitting the present manuscript, we also presented the results to the community, providing information about the methodology used in each step of the study (e.g., how we estimate genetic ancestry proportions) trying to use an accessible language. During this meeting, we also discussed about the reach and limitations of our results.

Using a comprehensive protocol including semi-structured questionnaires, we gathered socioeconomic data, in combination with individual and familiar health information, as well as a DNA sample from 143 individuals (44 males and 99 females) over 18 years of age. 96 samples were genotyped with Applied Biosystems Axiom Precision Medicine Research Array. We included all male individuals for genotyping in order to maintain a balanced number of female and male samples. Then we prioritized the female participants who were born in Puerto Madryn or have been living in the city the longest.

The research project, documents and protocol were approved by the Bioethics Committee of the Northern Programmatic Area (Comité de Bioética del Área Programática Norte; Resolution No. 19/17), Ministry of Health of the province of Chubut. All data were analyzed preserving the absolute anonymity and privacy of the volunteers.

2.2 | Identity and family history variables

For each volunteer, a semi-structured questionnaire was filled out, including questions about his or her history and ancestors. The full name, native language, and place of birth of father, mother, grandfathers, and grandmothers were queried (Supplementary Table S1). In addition to family history, self-perception of ancestry at continental level (Native American, African, European) was addressed, by using a five-point scale for each one: (1) 0–20% (none or very low), (2) 20%–40% (low), (3) 40%–60% (moderate), (4) 60%–80% (high), and (5) 80%–100% (very high or total), in order to compare with genetic ancestry data (Ruiz-Linares et al., 2014). Moreover, a field with open answers allowed participants to provide any complementary information about ancestries from other origins.

2.3 | Analyzed genotype data

To infer genetic ancestry in this sample, comparative genotype data was compiled for reference populations representing major biogeographic regions across the globe. These samples were obtained from the public databases 1000 Genomes (1KGP; Auton et al., 2015), Human Genome Diversity Project (HGDP; Bergström et al., 2020) and Simons Genome Diversity Project (SGDP; Mallick et al., 2016) (Supplementary Table S2). We filtered the variants to retain only bi-allelic single polymorphism positions (SNPs) and excluding variants with ambiguous A/T and G/C genotypes. Individuals with more than 5% missing genotypes for these positions were not included. We also excluded the individuals to avoid first and second-degree relatedness (estimated with *King*; Manichaikul et al., 2010). To avoid unbalanced sample sizes across populations, we randomly sampled 25 individuals for non-admixed populations that originally include more individuals. Genotyping data for Puerto Madryn individuals was added to this dataset using the same filters (one and three individuals were removed due to missingness and relatedness, respectively). SNPs with minor allele frequency (MAF) below 1% and a missing genotype rate greater than 2% were removed. As a result, the starting data set contained 1862 samples from the reference populations and 92 individuals from Puerto Madryn, and 330,555 SNPs.

2.4 | Genetic ancestry

We first explored the data using *Admixture* software (Alexander et al., 2009), on a set of 275,076 LD-pruned SNPs (–indep-pairwise flag in *plink1.9* [Chang et al., 2015] with default parameters) including all reference populations, using a number of putative ancestral population from 3 to 20, with one replicate per *K*. We observed that some populations did not exhibit any contribution to the genetic pool of the study sample (Supplementary Figure S1), and they were removed for downstream analyses using a purely quantitative criterion based on genetic ancestry estimates obtained at $K = 11$: we kept only populations that exhibits as a major genetic component any that segregates at more than 2% in at least one Puerto Madryn individual. The retained data set for estimating the reported genetic ancestry proportions leveraging two different techniques included 1332 (1240 + 92) samples (see Supplementary Table S2).

First, we obtained genetic ancestry proportion estimates using *Admixture* software (Alexander et al., 2009), on a set of 271,887 LD-pruned SNPs (pruning performed as before). For each number of putative ancestral populations *K* varying between 3 and 18, we performed 10 independent runs, and we retained the one exhibiting the highest likelihood. Results are presented in Supplementary Figure S2.

Second, we also applied fully haplotype-based methods, which provide higher resolution than allele-only approaches, to examine patterns of human population substructure (Lawson et al., 2012). This approach requires phased genotypes and thus, the haplotype phases were inferred using shapeIT2 (Delaneau et al., 2012) using the 1000 genomes recombination map that reflects averaged recombination rates in Sub-Saharan Africa, Europe, and East Asia (Auton et al., 2015).

The haplotype-based estimations were obtained through a six-stage procedure, for which we considered as "donor individual" any non-American individual and American individuals with >95% Native American ancestry estimated with Admixture, and "recipient individuals" the remaining ones (i.e., American individuals with <95% Native American ancestry).

At stage 1, we estimated the nuisance parameters Ne (i.e., 'recombination scaling constant') and M (i.e. per site mutation rate') with 10 iterations of the expectation–maximization algorithm in *CHROMOPAINTER v2* applied to chromosomes 3, 7, 10, 18, and 22 and considering all donor individuals. We obtained estimates of $Ne = 249.44$ and $M = 8.1883 \times 10^{-5}$, that we used as fixed values for subsequent runs with *CHROMOPAINTER*.

At stage 2, we used again *CHROMOPAINTER* algorithm to "paint" each donor individual's chromosomes as a combination of fragments received from all other donor individuals.

At stage 3, we clustered the "donor" individuals into homogeneous clusters according to their genetic ancestry, by applying 3 consecutive runs of *fineSTRUCTURE* to the outputs of *CHROMOPAINTER* (i.e., the painted "donor" chromosomes). Indeed, we inferred an approximative clustering with 200,000 MCMC iterations thinned every 10,000 and preceded by 100,000 burn in iterations, that was used as burn-in for by a second run with 1,000,000 more iterations. The MCMC file thus obtained was fed into a third run, to infer the tree structure with option –T 1 (Leslie et al., 2015). At the end of stage 3, the donors individuals clustered into 102 groups.

The stage 4 was aimed at increasing the intelligibility of subsequent analyses by reducing the number of donor clusters. While maintaining the consistency of the geographic clustering criteria, we iteratively regrouped in pairs from the outer branches to the innermost node. We thus obtained 19 clusters, named according to the majority geographic origin of the included individuals (Supplementary Figure S3): six European (Basque, Italy, Spain, Sardinia, West and Central Europe, and North Europe), five Sub-Saharan (Eastern Africa, Guinea Gulf, Southern African, and 2 for Western Africa), three Native

American (North America, Central America-Central Andes, and South American Tropical Forests), and five groups belonging to a broad geographic region encompassing from North Africa to Central Asia (Caucasus, Levant, North Africa, Middle East and Central South Asia).

At stage 5, we used *CHROMOPAINTER* to paint each recipient (admixed American) individual as a combination of genomic fragments inherited by "donor individuals" pooled using these 19 clusters. We used 10 Expectation–Maximization iterations (−ip flag).

For the final stage 6, we applied *SourceFind* (Chacón-Duque et al., 2018), a Bayesian method that allows to estimate the proportion of the genomes of recipient individuals inherited from each donor groups. Indeed, *SourceFind* models the copying vector for recipient (i.e., the *CHROMOPAINTER* outputs obtained at stage 5) as a weighted mixture of copying vectors from the donors performing 400,000 MCMC iterations thinned every 1000, and preceded by 100,000 burn-in iterations. We disabled "self-copying," and we set the number of surrogates that can be used to form the target group in each MCMC iteration to 8, while the expected number of such surrogates was set to 4. We also used 100 equally sized proportions to be assigned to the selected surrogates at each MCM iteration.

Finally, the proportions of four main genetic ancestry groups were assessed by adding the genetic contribution from the different donor groups described before: North Africa to Central South Asia, Native American, Sub-Saharan African, and European.

3 | RESULTS

Through a model-based approach with Admixture software using SGDP, HGDP, and 1KGP panel references, we first obtained genetic ancestry proportions estimates for Puerto Madryn individuals, which can be intended as representative of admixed populations from Northern Patagonia. Based on cross-validation scores, we retained the model with $K = 5$ ancestral populations (Supplementary Figure S2A). We observed one specific ancestry component for populations belonging to each three following population groups: Sub-Saharan African, European, and Native American. One other is observed in Caucasus, Middle East and Central South Asia while the last one encompasses populations from Caucasus, Middle East, North Africa and South Europe (Supplementary Figure S2B,C).

The haplotype-based methods proved to provide clearer boundaries among the main reference (donors) groups of individuals (small pie-charts in Figure 1), especially for the European component (Supplementary Figures S4C and S4D). We grouped them into main four main geographic regions: Europe, Sub-Saharan Africa, Native Americans, and North Africa/Levant/Caucasus/Middle East/Central South Asia (hereafter referred as from North Africa to Central South Asia). We acknowledge that the clusters belonging to the North Africa to Central South Asia group are rather heterogeneous genetically (Supplementary Figure S3), what reflects the extensively described genetic gradient in this region (Li et al., 2008; also observed in our *Admixture* analyses shown in Supplementary Figures S2 and S4C). However, we decided to group these five clusters to perform comparisons or measure the potential discrepancy between genetic and self-perceived ancestries, based on the questionnaires filled by the volunteers. Although the ancestry components identified through both methods slightly differ in their geographic boundaries, their proportions in Puerto Madryn individuals highly correlate (Supplementary Figure S4 A and B). For a more in-depth interpretation of the results, we focus the following analyses on genetic ancestry estimations obtained by haplotype-based methods.

3.1 | Comparisons between self-perceived and genetic ancestries

According to the questionnaires, there is a tendency for self-perceived Native American and African ancestry to be lower than European ancestry. Forty-eight percent of the volunteers considered that they had a very low percentage of Native American ancestry, even close to zero (Table 1). This tendency was even stronger for African origins, as 91% of the volunteers responsed in the range 0–20% (none or very low) for this ancestry. On the other hand, only 12% of the participants considered that they have little or no European ancestry at all.

We also asked about self-perception of other origins and 10% of the volunteers reported an ancestral component different from Native American, European, and African. In these cases, the answers were diverse and difficult to quantify: "I have an origin in Arab countries," "Yes, from Asia," "Oriental," "40% Arab," "Yes, Lebanese," "Arab," "Muslim," "Asian at some point," "from Turkey," "Moorish," among others.

Although self-perceived and genetic ancestry are significantly correlated for the European and Native American components ($p > 0.558$; $p < 2.00 \times 10^{-9}$; Figure 2a,b), individuals tend to overall overestimate their Native American ancestry and to underestimate the European one, as evidenced by significant paired Student's tests ($p < 2.26 \times 10^{-2}$; Figure 2c). Moreover, this general trend is inconsistent across classes of self-perceived ancestry, as we observed that volunteers who self-perceived as having very high or total Native American ancestry do not exhibit such almost exclusive genetic ancestry (right-hand category in Figure 2b), mostly because they self-perceived none or very low European ancestry (left-hand category in Figure 2a). As a whole, self-perceived ancestry remains a very poor proxy for genetic ancestry.

TABLE 1 Distribution of participants falling in different Native American, European, and African self-perceived ancestry categories

	Native American	European	African
0–20%	48.2	11.9	90.9
20%–40%	9.8	20.3	7.7
40%–60%	17.5	17.5	0.7
60%–80%	17.5	18.2	0.0
80%–100%	7	30	0.0
Do not know	0.0	2.1	0.7

Unfortunately, we were unable to perform the same analyses for the North Africa to Central South Asia ancestries (these origins were not included in the questionnaire in the form of the five-point scale for continental ancestry) nor for Sub-Saharan African component (its genetic representation in Puerto Madryn sample was too scarce). However, we observed that people that recognize a genealogical link with North Africa to Central South Asia origins tend to exhibit significantly higher genetic ancestries for these regions than individuals that do not (Wilcoxon test $p = 2.71 \times 10^{-2}$; Figure 2d).

The results concerning the comparisons between self-perceived and genetic Native American ancestry are consistent when leveraging genetic ancestry proportion estimates with *Admixture* instead of *SourceFind* (Supplementary Figure S5). However, this does not hold true for the European and the North Africa to Central South Asia genetic components, most likely because of the gradient of the genetic ancestries among reference populations in these regions.

4 | DISCUSSION

The European genetic ancestry component most represented in the Puerto Madryn sample is derived from Iberian populations (0.071 from Spain and 0.104 of Basque origin), followed by components associated with Italian populations and with Northern Europe populations (0.140 for both components). These genetic contributions are congruent with the historical gene flow from those origins to Latin America in general, and to Argentina and Patagonia in particular. In chronological order, the component from the Iberian Peninsula was

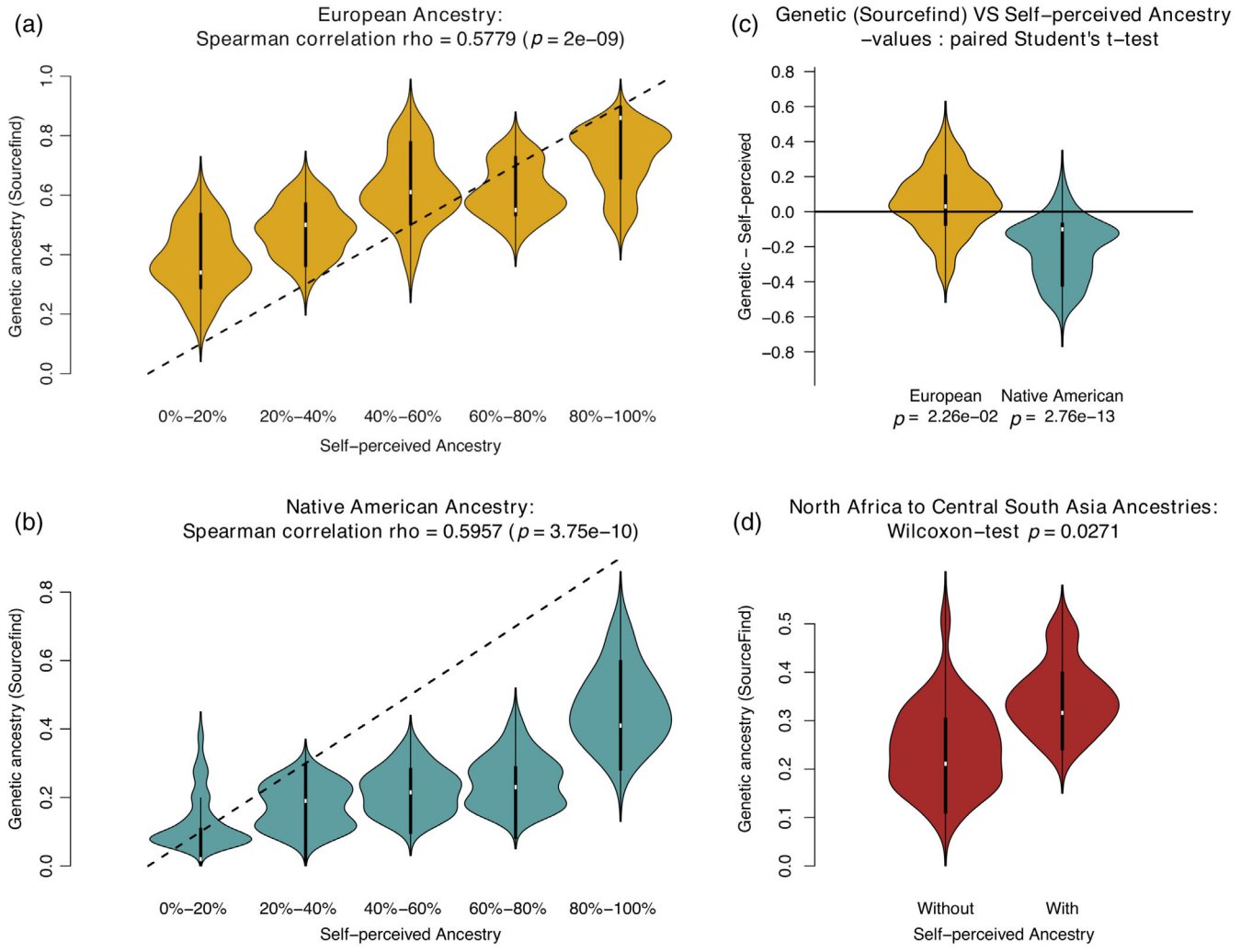


FIGURE 2 Comparisons of genetic and self-perceived ancestries in Puerto Madryn. (a–b) Violin plots of European and Native American genetic ancestry estimates in individuals classified into five groups according to their self-perceived ancestry. The dashed line (y = x) represents the ideal case of genetic and self-perceived ancestry being identical. We report the Spearman's correlation estimate ρ and the associated p-value. (c) Distribution of the differences in European and Native American genetic and self-perceived ancestries. We report for each ancestry component the p-value for the paired Student's t-test between genetic and self-perceived ancestries. (d) Comparison of the North Africa to Central South Asia genetic ancestries proportion estimates between individuals that self-identified a link with these origins in their genealogy and those who did not. We report the p-value for Wilcoxon test between genetic proportion estimates between both groups.

introduced from the Conquest onward, while the great migratory waves of the 19th and 20th centuries introduced new variants from different regions of Europe, mainly Italy and Spain but also largely from Russia (Margulis, 1977). Regarding the Basque origin, its arrival in America dates back to the colonial era, especially since the last third of the 18th century (Cruset, 2018). One of the preferred destinations was the area of the Río de la Plata. Then, during the great migratory wave between 1870 and 1930, a large number of persons arrived in Argentina from the sub-Pyrenean Basque Country (provinces of Alava, Guipuzcoa, Navarra and Vizcaya), and spread throughout a large part of Argentina (Torry, 2011).

In our analysis, we found 0.169 Native American genetic ancestry, comprised at the lower limit of the confidence interval reported with Ancestry Informative markers (AIMs) from a previous article (Parolin et al., 2019) studying samples from Puerto Madryn city and other four different localities in Patagonia. These differences in the percentages of Native ancestry may be due to different volunteer recruitment strategies or a possible sample size effect.

The contribution of genetic ancestry components from the regions encompassed within the North Africa to Central South Asia geographical range accounted to 0.225 of the genetic ancestry of Puerto Madryn. This result can be explained in the light of several historical facts. First, after the conquest of America in the 15th century, many Sephardic Jewish suffered religious persecution and were forced to converted to Catholicism by the Spanish crown. In this great diaspora movement, 100,000–300,000 Jews (estimates vary) left Spain. That migratory signal was detected in a previous genetic study by Chacón-Duque et al. (2018), who reported the presence of an ancestry component from the Southern/Eastern Mediterranean in Latin America.

However, other demographic events could explain why we observe components of North Africa to Central South Asia genetic ancestry in Argentina. Indeed, the third massive migration in Argentina after those from Spain and Italia was from Syria and Lebanon (Díaz-Jatuf, 2013), what is particularly well reflected by our genetic ancestry estimates. The main causes of expulsion were linked to the deep socioeconomic crisis that resulted from the decline of the archaic economic structures imposed on Syria and Lebanon by the Ottoman Empire since the second half of the 19th century. At present, it is estimated that 3.5 million people in Argentina come (or has a direct ancestor) from this region. The first immigrants from the Ottoman Empire (mostly Christian Arabs and Shiite Muslim Arabs) arrived in the 1860s, principally in the Northeast and Northwest of Argentina. However, from the beginning of the 20th century, Syrian-Lebanese immigrants also arrived and established in Patagonia, through the commercial networks (Chávez, 2019). Although among these immigrants there was a certain inbreeding tendency, marital unions between Syrian-Lebanese men and postcolonial admixed women ("Criollas") or native women have been documented (Ibarra & Hernández, 2016).

Moreover, during the 19th and 20th centuries, Argentina was the destination for multiple migratory waves of Jewish origin from Eastern Europe. The circumstances were diverse, from economic reasons to political and religious persecution. This process involved about 150,000 people, from its beginnings until the end of the 1920s. Then, between 1930 and 1945, more than 45,000 people settled in the country fleeing the Nazi regime (Avni, 1983). In addition to these historical references, contemporary studies point to the Middle East region as the geographical setting that was the initial focus of the genetic variability of today's Jewish groups (Atzmon et al., 2010; Behar et al., 2010). The European Ashkenazi and Sephardi communities have approximately 70% of their ancestry from the Middle East, also sharing a common origin with other groups from the Caucasus, approximately 3000 years ago. Regarding the Puerto Madryn sample, some participants carrying a high percentage of this genetic origin reported ancestors whose native language is Yiddish (belonging to Ashkenazi Jewish communities). Other studies also pointed out that the North African genetic component observed in the Iberian Peninsula probably reflects the impact of Arab expansion since the 7th century and the subsequent expansion of Christian kingdoms (Arauna et al., 2019).

The genetic ancestry component with the lowest proportion in the whole sample (0.016) is related to current populations from Sub-Saharan Africa. In previous studies in Argentina (Avena et al., 2012; Corach et al., 2010), this genetic finding has been associated mostly to the slave exploitation of individuals, mainly males, brought as labor force from Africa to America. Here, we bring more evidence on the particular importance of Eastern African populations in the Sub-Saharan African genetic origins in Argentina, as previously suggested (Luisi et al., 2020).

At different times, important migratory movements from different regions of the world have been integrated into the Argentinean population. Although the largest volume of migrants came from the European continent, we must avoid shrink the migration analysis by focusing only on the most statistically frequent group. Recognizing the diversity will help to build reliable ideas about our past and previous generations. This, in turn, has potential impacts, not only on better understanding the history of Argentina since colonialism, but also on the field of health. Indeed, medicine is increasingly looking at the origins of individuals and populations, since health and disease are influenced by the intersection of features that are associated with genetically inferred and socially constructed origins (Rebbeck et al., 2022). In parallel, Argentina and other Latin American nations discuss several approaches to better translating the benefits of research on genetic and non-genetic diversity of their population to public health policies, with the aim at incorporating local components into the precision medicine initiatives already ongoing (Dopazo et al., 2019).

4.1 | Difficulties to register self-perceived ancestry

Self-perceived ancestry and genomics may be related, but in our analysis, we observed discrepancies between these two variables. People may find it difficult to summarize their family and genealogical history in terms of percentages, as requested in the questionnaire. It involves

a certain level of abstraction and, fundamentally, it is an infrequent question for that aspect of their life. Individuals who self-reported Native ancestry tend to overestimate the genomic component of that origin (see Figure 2b). The same has been documented in previous works on Latin American populations, comparing individual reports in percentage values and associated genetic ancestry (Paschetta et al., 2021; Ruiz-Linares et al., 2014). In Brazil, Santos et al. (2009) explored the association between biological/genetic information and perceptions about skin color and race, and reported that the participants estimated their Native ancestry well above the levels revealed by genomic testing. Among the possible causes, it is worth mentioning the struggle, resistance, and processes of identity construction of Native communities, which have contributed to a more favorable context for the legitimization and acceptance of an identity linked to Native populations. These affirmative actions have materialized, in Latin American countries, in policies favorable to minorities of African and Native American descent carried out in recent decades (Paschetta et al., 2021). In Argentina, the 1994 reform of the National Constitution recognizes the "ethnic and cultural pre-existence of Argentina's Indigenous peoples" as well as "the legal status of their communities, and the community possession and ownership of the lands they traditionally occupy." These processes could have enabled a vindication of Indigenous identities by a broad sector of society.

We would like to emphasize that volunteers in our sample with a genomic component linked to current populations from North Africa to Central South Asia expressed self-perception of similar origins. This perception of a past linked to migratory movements from these regions justifies the inclusion of more categories in future population genomics studies in Argentina, particularly those that also ask about people's self-perception. By using another model of questioning, where these categories of self-recognition are reflected, we could contribute to open panoramas of greater depth and complexity with which to think about our origins. As has already been shown, research, and especially the way in which the results of scientific studies are communicated, can influence people's social constructions and notions about certain topics (García et al., 2016; Pin & Gutteling, 2009; Sturgis et al., 2004).

4.2 | Beyond the three-hybrid model

The three-hybrid model has become a canon in American settlement studies. However, it is limited. When we think of the first moment in which Indigenous populations and groups of other origins made contact, we go back to the end of the 15th century and the beginning of the 16th century, 500 years ago from the present. By embracing these limits, we are omitting earlier processes that shaped the genetic diversity in populations from each continent before the colonial contact. Movements that have clearly contributed to the current diversity of our territory (e.g., migrations from the Near and Middle East, and from countries on the border between Asia and Europe, such Russia or Armenia) are also generally omitted. These groups have been overlooked in population studies, even when there is a historical record or oral memory about this past.

By perceiving and transmitting a simplified model of population ancestry with only three parental populations, scientists are likely contributing to the construction of an imaginary group and society. It is necessary to consider the possibility of more complex and comprehensive schemes on the history of the human groups that arrived in the American continent. Those who migrated to Argentina had a history of miscegenation of their own, involving different regions of different continents, several generations before. They have yet to be considered in investigations aimed at reconstructing the population history of Argentina. The European population, as they are recognized today in their political boundaries, is the result of countless processes of gene flow, some very recent, with populations from different places of Asia and Africa (Arauna et al., 2019; Botigué et al., 2013; Plaza et al., 2003; Secher et al., 2014).

In this context, we should consider about the classifications used in genomic ancestry studies. It is possible that in Latin American populations, and particularly those in Argentina, we designate as European a large percentage of our ancestry that is not. This may contribute to reinforce the idea, historically rooted in colonial agendas and currently present in the imagination of a certain sector of society, that European populations are practically the founders of Latin Americans and, therefore, those with the greatest presence in biological terms. National narratives still have a long way to go toward a greater openness in the recognition of all populations who have contributed to what is today Latin America. However, challenging current analytical frameworks, such as the three hybrid model, offers the opportunity to perform rigorous analyses of population genomic histories that can complement social and cultural understandings of Latin America today.

4.3 | Conclusions

Our study shows the studied Argentinean Patagonia population in this article is demographically diverse. Within some confidence this can be extended to different regions of Argentina and Latin America in general, as the oral and written history has taught us. Given the current availability of multi-geographic genomic databases, it is recommended to incorporate a broader vision in the design of studies on the genetic variability of our populations, including varied geographical origins. On the other hand, self-perceived and genetic ancestry are both complementary ways of approaching the history and identity of people, that do not necessary overlap and should not be considered synonymous, particularly in medical practice.

AUTHOR CONTRIBUTIONS

Carolina Paschetta: Conceptualization (supporting); data curation (equal); funding acquisition (equal); writing – original draft (supporting). **Tamara Teodoroff:** Data curation (equal). **Luis Orlando Pérez:** Data curation (equal); funding acquisition (equal). **Soledad de Azevedo:** Data curation (equal); funding acquisition (equal). **Magda Alexandra Trujillo-Jiménez:** Data curation (equal). **Pablo Navarro:** Data curation (equal). **Leonardo Morales:** Data curation (equal). **Bruno Pazos:** Data curation (equal). **Virginia Rmallo:** Data curation (equal); funding

acquisition (equal); project administration (lead); supervision (lead); writing – original draft (supporting).

ACKNOWLEDGMENTS

The authors specially thank to all the volunteers who were part of this project. Without their commitment, our work would not have been possible. We are also deeply grateful to the health personnel of the Primary Care Centers "Favaloro" and Fontana" of Puerto Madryn. Their help, along months and at different stages of data collection, was invaluable. Our heartfelt thanks to the Hemotherapy Service and the Laboratory of the Hospital Zonal de Puerto Madryn "Dr. Andrés Ísola" for training us, with patience and warmth, in biological material extraction techniques. Sincere gratitude is extended to Lara Rubio Arauna for answering many questions concerning the implementation of the haplotype-based methods. Finally, we thank the Argentinean National Agency for the Promotion of Research, Technological Development and Innovation (PICT 3206) and the Argentinean National Council for Scientific and Technical Research (PIP D.111/16), for the funding that made this work possible.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data analyzed here comprises both newly generated and previously reported data sets. Access to publicly available datasets should be requested through the distribution channels indicated in each published study. Allele frequency data of the Puerto Madryn newly generated samples have been registered the public database https://ri.conicet.gov.ar/handle/11336/179519. Following this project's regulations, informed consent, the Argentinean National Law N° 25.326 of Personal Data Protection, and the Ministry of Health resolutions N° 2940/2020 and 1480/2011, genotypic information would be shared upon prior evaluation of the request and through a Material Transfer Agreement. Data cannot be used for commercial purposes or to identify the sample donors.

ORCID

Anahí Ruderman https://orcid.org/0000-0002-9610-2997
Pierre Luisi https://orcid.org/0000-0002-2542-3689
Carolina Paschetta https://orcid.org/0000-0002-5869-3570
Luis Orlando Pérez https://orcid.org/0000-0003-2939-7753
Soledad de Azevedo https://orcid.org/0000-0003-4601-0717
Magda Alexandra Trujillo-Jiménez https://orcid.org/0000-0001-5506-3496
Pablo Navarro https://orcid.org/0000-0003-2180-449X
Leonardo Morales https://orcid.org/0000-0002-3980-8862
Bruno Pazos https://orcid.org/0000-0002-0965-070X
Rolando González-José https://orcid.org/0000-0002-8128-9381
Virginia Ramallo https://orcid.org/0000-0002-7856-4856

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Arauna, L. R., Hellenthal, G., & Comas, D. (2019). Dissecting human north African gene-flow into its western coastal surroundings. Proceedings. *Biological Sciences*, *286*(1902), 20190471. https://doi.org/10.1098/rspb.2019.0471

Atzmon, G., Hao, L., Pe'er, I., Velez, C., Pearlman, A., Palamara, P. F., Morrow, B., Friedman, E., Oddoux, C., Burns, E., & Ostrer, H. (2010). Abraham's children in the genome era: Major Jewish diaspora populations comprise distinct genetic clusters with shared middle eastern ancestry. *American Journal of Human Genetics*, *86*(6), 850–859. https://doi.org/10.1016/j.ajhg.2010.04.015

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., et al. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Avena, S., Via, M., Ziv, E., Pérez-Stable, E. J., Gignoux, C. R., Dejean, C., Huntsman, S., Torres-Mejía, G., Dutil, J., Matta, J. L., Beckman, K., González Burchard, E., Parolin, M. L., Goicoechea, A., Acreche, N., Boquet, M., Ríos Part, M., Fernández, V., Rey, J., … Fejerman, L. (2012). Heterogeneity in genetic admixture across different regions of Argentina. *PLoS ONE*, *7*(4), e34695. https://doi.org/10.1371/journal.pone.0034695

Avena, S. A., Goicoechea, A. S., Clapsos, R., Dugoujon, J. M., Dejean, C. B., Perosino, C., & Carnese, F. R. (2003). Aporte aborigen y africano de diferentes regiones de la Argentina en Buenos Aires. En Actas Sextas Jornadas Nacionales de Antropología Biológica. Catamarca, Universidad Nacional de Catamarca.

Avena, S. A., Parolin, M. L., Boquet, M., Dejean, C. B., Postillone, M. B., Álvarez Trentini, Y., Di Fabio Rocca, F., Mansilla, F., Jones, L., Dugoujon, J. M., & Carnese, F. R. (2010). Mezcla génica y linajes uniparentales en Esquel (provincia del Chubut). Su comparación con otras muestras poblacionales argentinas. *Journal of Basic & Applied Genetics*, *21*(1), 1–14.

Avni, H. (1983). *Argentina y la historia de la inmigración judía, 1810–1950, Capítulo 2* (p. 1983). Magnes Press.

Behar, D. M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., Rootsi, S., Chaubey, G., Kutuev, I., Yudkovsky, G., Khusnutdinova, E. K., Balanovsky, O., Semino, O., Pereira, L., Comas, D., Gurwitz, D., Bonne-Tamir, B., Parfitt, T., Hammer, M. F., … Villems, R. (2010). The genome-wide structure of the Jewish people. *Nature*, *466*, 238–242. https://doi.org/10.1038/nature09103

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J. F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., … Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, *367*(6484), eaay5012. https://doi.org/10.1126/science.aay5012

Besteme, J. O. (1988). La inmigración sirio-libanesa en la Argentina. *Estudios Migratorios Latinoamericanos*, *9*, 239–268.

Botigué, L. R., Henn, B. M., Gravel, S., Maples, B. K., Gignoux, C. R., Corona, E., Atzmon, G., Burns, E., Ostrer, H., Flores, C., Bertranpetit, J., Comas, D., & Bustamante, C. D. (2013). Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(29), 11791–11796. https://doi.org/10.1073/pnas.1306223110

Carnese, F. R., Avena, S. A., Parolin, M. L., Postillone, M. B., & Dejean, C. B. (2011). Gene admixture analysis through genetic markers and geneological data in a sample from the Buenos Aires metropolitan area. In S. Gibbon, R. Y. Ventura Santos, & M. Sans (Eds.), *Racial identities, genetic ancestry, and health in South America* (pp. 177–194). Palgrave Macmillan.

Chacón-Duque, J. C., Adhikari, K., Fuentes-Guajardo, M., Mendoza-Revilla, J., Acuña-Alonzo, V., Barquera, R., Quinto-Sánchez, M., Gómez-Valdés, J., Everardo Martínez, P., Villamil-Ramírez, H., Hünemeier, T., Ramallo, V., Silva de Cerqueira, C. C., Hurtado, M., Villegas, V., Granja, V., Villena, M., Vásquez, R., Llop, E., … Ruiz-Linares, A. (2018). Latin Americans show wide-spread converso ancestry and imprint of local native

ancestry on physical appearance. *Nature Communications*, *9*(1), 5388. https://doi.org/10.1038/s41467-018-07748-z

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7.

Chávez, M. (2019). Los inmigrantes sirio-libaneses y su inserción territorial en el sudeste de Río Negro, Argentina (1912-1930). *Magallania*, *47*(2), 5–19.

Corach, D., Lao, O., Bobillo, C., van Der Gaag, K., Zuniga, S., Vermeulen, M., van Duijn, K., Goedbloed, M., Vallone, P. M., Parson, W., de Knijff, P., & Kayser, M. (2010). Inferring continental ancestry of Argentineans from autosomal, Y-chromosomal and mitochondrial DNA. *Annals of Human Genetics*, *74*(1), 65–76. https://doi.org/10.1111/j.1469-1809.2009.00556.x

Cruset, M. E. (2018). Migración transnacional: la diáspora vasca en Argentina como agente de para-diplomacia. *Relaciones Internacionales*, *20*(40), 124.

Delaneau, O., Marchini, J., & Zagury, J. F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, *9*(2), 179–181.

Devoto, F. (2007). La inmigración de ultramar. En: Torrado, S. (Comp.). Población y bienestar en la Argentina del primero al segundo centenario. Una historia social del siglo XX. Tomo I. Serie Estudios del Bicentenario. Buenos Aires: ADHASA.

Di Fabio Roca, F., Solís, E., Ares, A., Romaldini, M., Rocco, F., Vaccaro, M. S., Avena, S. A., Dejean, C. B., Carnese, F. R., & De La Vega, E. D. (2011). Análisis de marcadores autosómicos y mezcla génica en la población de Rosario, provincia de Santa Fe. *En Actas Décimas Jornadas Nacionales de Antropología Biológica*, 112.

Díaz-Jatuf, J. (2013). Los árabes en Argentina / Abdeluahed Akmir. *AlQáfila: Revista bilingüe Online Sobre el Mundo árabe*, *1*(1), 273.

Dipierri, J. E., Alfaro, E., & Bejarano, I. F. (1999). Surnames, ABO system and miscegenation in highland populations of province of Jujuy (Nortwest argentine). *Homo*, *50*(1), 14–20.

Dopazo, H., Llera, A. S., Berenstein, M., & González-José, R. (2019). Genomas, enfermedades y medicina de precisión: Un proyecto nacional. *Ciencia, Tecnología y Política*, *2*, 7.

Escudero, H. B., & Rubilar, R. A. (2017). *Miradas migrantes en la provincia del Chubut: la interculturalidad en las aulas*. Editorial Universitaria de la Patagonia.

Etcheverry, M. A. (1947). El factor Rhesus en personas de ascendencia ibérica e itálica residentes en Argentina. *La Semana Médica*, *2*(2082), 500.

Etcheverry, M. A. (1949). Frecuencia de los tipos sanguíneos Rh en la población de Buenos Aires. *Revista de la Sociedad Argentina de Hematología y Hemoterapia*, *1*, 166–168.

García, A., Demarchi, D. A., Tovo-Rodrigues, L., Pauro, M., Callegari-Jacques, S., Salzano, F. M., & Hutz, M. H. (2015). High interpopulation homogeneity in Central Argentina as assessed by ancestry informative markers (AIMs). *Genetics and Molecular Biology*, *38*(3), 324–331.

García, A., Oliveira Rufino, R., Bergese, A. B., Agüero, J. F., Cuevas, A., Díaz-Rousseau, G., Pauro, M., Nores, R., Garita-Onandía, Y., Tavella, M. P., & Demarchi, D. A. (2016). El cruce entre las antropologías. Una mirada interdisciplinaria en torno a la genética de poblaciones, las memorias familiares y la construcción identitaria. *Revista Del Museo De Antropología*, *9*(2), 105–112. https://doi.org/10.31048/1852.4826.v9.n2.13614

Ibarra, H., & Hernández, C. (2016). *Estado, economía y sociedad. Trelew y su Hinterland. 1889–1999*. Mandala Libros.

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, *8*, e1002453.

Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E. C., Cunliffe, B., Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, Lawson, D. J., Falush, D., Freeman, C., Pirinen, M., Myers, S., Robinson, M., Donnelly, P., & Bodmer, W. (2015). The fine-scale genetic structure of the British population. *Nature*, *519*(7543), 309–314. https://doi.org/10.1038/nature14230

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., & Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, *319*(5866), 1100–1104.

López Camelo, J. S., Cabello, P. H., & Dutra, M. G. (1996). A simple model for the estimation of congenital malformations frequency in racially mixed populations. *Brazilian Journal of Genetics*, *19*(4), 659–663.

Luisi, P., García, A., Berros, J. M., Motti, J. M. B., Demarchi, D. A., Alfaro, E., Aquilano, E., Argüelles, C., Avena, S., Bailliet, G., Beltramo, J., Bravi, C. M., Cuello, M., Dejean, C., Dipierri, J. E., Jurado Medina, L. S., Lanata, J. L., Muzzio, M., Parolin, M. L., … Dopazo, H. (2020). Fine-scale genomic analyses of admixed individuals reveal unrecognized genetic ancestry components in Argentina. *PLoS One*, *15*(7), e0233808. https://doi.org/10.1371/journal.pone.0233808

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206. https://doi.org/10.1038/nature18964

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*, 2867–2873.

Margulis, M. (1977). Inmigraión y Desarrollo Capitalista. La migración europea a la Argentina. *Demografía y Economía*, *11*, 273–306.

Morales, J. O., Dipierri, E. J., Alfaro, E., & Bejarano, I. F. (2000). Distribution of the ABO system in the argentine northwest: Miscegenation and genetic diversity. *Interciencia*, *25*(9), 432–435.

Ongaro, L., Scliar, M. O., Flores, R., Raveane, A., Marnetto, D., Sarno, S., Gnecchi-Ruscone, G. A., Alarcón-Riquelme, M. E., Patin, E., Wangkumhang, P., Hellenthal, G., Gonzalez-Santos, M., King, R. J., Kouvatsi, A., Balanovsky, O., Balanovska, E., Atramentova, L., Turdikulova, S., Mastana, S., … Montinaro, F. (2019). The genomic impact of European colonization of the Americas. *Current Biology*, *29*(23), 3974–3986.e4. https://doi.org/10.1016/j.cub.2019.09.076

Palatnik, M. (1966). Seroantropología Argentina. *Sangre*, *11*, 395–412.

Palazzo, R., & Tenconi, J. (1939). Estadística sobre 15.000 clasificaciones de grupos sanguíneos, realizadas en Buenos Aires. *Semana Médica*, *2*, 459–460.

Parolin, M. L., Avena, S. A., Dejean, C. B., Jaureguiberry, S. M., Sambuco, L. A., & Carnese, F. R. (2012). Y-chromosomal STR haplotype diversity in a sample from the metropolitan area of Buenos Aires (Argentina). *Revista del Museo de Antropología*, *5*, 53–64.

Parolin, M. L., Toscanini, U. F., Velázquez, I. F., Llull, C., Berardi, G. L., Holley, A., Tamburrini, C., Avena, S., Francisco, R. Carnese, Lanata, J. L., Carnero, N. S., Arce, L. F., Basso, N. G., Pereira, R., & Gusmão, L. (2019). Genetic admixture patterns in Argentinian Patagonia. *PLoS One*, *14*(6), e0214830.

Paschetta, C., de Azevedo, S., Ramallo, V., Cintas, C., Pérez, O., Navarro, P., Bandieri, L., Sánchez, M. Q., Adhikari, K., Bortolini, M. C., Ferrara, G. P., Gallo, C., Bedoya, G., Rothhammer, F., Alonzo, V. A., Ruiz-Linares, A., & González-José, R. (2021). The impact of socioeconomic and phenotypic traits on self-perception of ethnicity in Latin America. *Scientific Reports*, *11*, 12617. https://doi.org/10.1038/s41598-021-92061-x

Pauro, M., García, A., Bravi, C. M., & Demarchi, D. A. (2010). Distribución de haplogrupos mitocondriales alóctonos en poblaciones rurales de Córdoba y San Luis. *Revista Argentina de Antropología Biológica*, *12*(1), 47–55.

Pin, R., & Gutteling, J. M. (2009). The development of public perception research in the genomics field: An empirical analysis of the literature in the field. *Science Communication*, *31*(1), 57–83. https://doi.org/10.1177/1075547008327273

Plaza, S., Calafell, F., Helal, A., Bouzerna, N., Lefranc, G., Bertranpetit, J., & Comas, D. (2003). Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. *Annals of Human Genetics*, *67*, 312–328. https://doi.org/10.1046/j.1469-1809.2003.00039.x

Rebbeck, T. R., Mahal, B., Maxwell, K. N., Garraway, I. P., & Yamoah, K. (2022). The distinct impacts of race and genetic ancestry on health. *Nature Medicine*, *28*, 890–893. https://doi.org/10.1038/s41591-022-01796-1

Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., Quinto-Sanchez, M., Jaramillo, C., Arias, W., Fuentes, M., Pizarro, M., Everardo, P., de Avila, F., Gómez-Valdés, J., León-Mimila, P., Hunemeier, T., Ramallo, V., Silva de Cerqueira, C. C., Burley, M. W., Konca, E., de Oliveira, M. Z., Veronez, M. R., … Gonzalez-José, R. (2014). Admixture in Latin America: Geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genetics*, *10*(9), e1004572. https://doi.org/10.1371/journal.pgen.1004572

Salas, A., Jaime, J. C., Álvarez-Iglesias, V., & Carracedo, A. (2008). Gender bias in the multiethnic composition of Central Argentina. *Journal of Human Genetics*, *53*, 662–674.

Santos, R. V., Fry, P. H., Monteiro, S., Maio, M. C., Rodrigues, J. C., Bastos-Rodrigues, L., & Pena, S. D. (2009). Color, race, and genomic ancestry in Brazil: Dialogues between anthropology and genetics. *Current Anthropology*, *50*(6), 787–819. https://doi.org/10.1086/644532

Secher, B., Fregel, R., Larruga, J. M., Cabrera, V. M., Endicott, P., Pestano, J. J., & González, A. M. (2014). The history of the north African mitochondrial DNA haplogroup U6 gene flow into the African, Eurasian and American continents. *BMC Evolutionary Biology*, *14*, 109. https://doi.org/10.1186/1471-2148-14-109

Sturgis, P., Cooper, H., Fife-Schaw, C., & Shepherd, R. (2004). Genomic science: emerging public opinion. In A. Park, J. Curtice, K. Thomson, C. Bromley, & M. Phillips (Eds.), *British social attitudes: The 21st report* (pp. 119–140). SAGE Publications.

Torry, E. (2011). El asociacionismo vasco en Argentina. Notas sobre sus componentes identitarios. IX sociology conference. faculty of social sciences, University of Buenos Aires, Buenos Aires.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ruderman, A., Luisi, P., Paschetta, C., Teodoroff, T., Pérez, L. O., de Azevedo, S., Trujillo-Jiménez, M. A., Navarro, P., Morales, L., Pazos, B., González-José, R., & Ramallo, V. (2023). Genetic and self-perceived ancestries in Argentina: Beyond the three-hybrid model. *American Journal of Biological Anthropology*, 1–11. https://doi.org/10.1002/ajpa.24702

FEATURE ARTICLE

Bulsarapp: Interactive Visual Analysis for Surname Trend Exploration

Leonardo Morales and Pablo Navarro, *Instituto Patagónico de Ciencias Sociales y Humanas, Centro Nacional Patagónico (CENPAT), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), CP U9120, Puerto Madryn, Argentina, and with the Departamento de Informática (DIT), Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco, V9410, Trelew, Argentina*

Celia Cintas, *IBM Research Africa, CP 00200, Nairobi, Kenya*

Rolando González-José and Virginia Ramallo, *Instituto Patagónico de Ciencias Sociales y Humanas, Centro Nacional Patagónico (CENPAT), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), CP U9120, Puerto Madryn, Argentina*

Claudio Delrieux, *Laboratorio de Ciencias de las Imágenes, Departamento de Ingeniería Eléctrica y Computadoras, Universidad Nacional del Sur, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), CP B8000, Bahía Blanca, Argentina*

The study of surnames for a given population, together with their distribution and spatial patterns identification, has been a long-standing problem in the fields of human biology, public health, and social sciences. The ancestry inferred from surname information can be a useful means to understand the dynamics of human populations. This knowledge allows us to characterize geographically the ethnicity of populations, and to understand the complex relationships between identity, migration, and health issues in a demographic view. However, in most cases, a detailed geolocalization of this data can be a daunting task. We propose a visual analytic tool that summarizes the heterogeneous surname and geographic information collected from Argentinean electoral rolls. This tool allows a massive data analysis, and facilitates interdisciplinary studies about population dynamics related to ancestry, migration, and health. It also offers an easy-to-use interface that allows interactive exploration of isonymy and surname origins, their distribution, and spatial trends in a high population density context.

Human populations are not static in time or space. The variability observed in visible phenotypic characteristics correlates with the nonvisible genomic level. The biological plasticity of our species, added to our capacity to create culture and technology, has allowed us to occupy the entire planet, with only very few places with extreme climates. For centuries, and even today, humans migrated from one territory to another, primarily for economic reasons. Migration generates changes both in the origin and in the destination places, giving rise to different processes of miscegenation. Although widespread and global, this phenomenon does not occur in all directions. Sometimes, geographical boundaries (oceans and mountain ridges) or cultural barriers (linguistic and religious) limit miscegenation and offspring. Marriages are never random, and several underlying reasons lead to the preference of one type of union while rejecting others. Measuring this trend and discriminating the isolation degree among populations is relevant for the associated medical issues (for instance, the well-known undesirable effect of consanguinity on DNA variability). Today, we know that congenital anomalies and many diseases had great genetic components in their etiology. Therefore,

0272-1716 © 2021 IEEE
Digital Object Identifier 10.1109/MCG.2021.3115052
Date of publication 24 September 2021; date of current version 15 July 2022.

knowing population inheritance aspects over time is essential in health research, prevention, diagnosis, and treatment. Moreover, a key factor is to disseminate this knowledge through educational activities involving potentially affected communities. For instance, the Brazilian National Institute on Population Medical Genetics (INaGeMP) conducts several projects, including a national census of Brazilian populations with a high frequency of genetic diseases and exposure to high-risk genetic factors in isolated populations (inbreeding), or clusters with a high rate of rare diseases. Their goal is to identify population groups and individuals at risk of transmitting disorders, provide genetic advice, enforce neonatal screening, and develop open educational programs to the community. Accurate diagnosis enables treatment and provides pathways for further studies.

A massive and inexpensive way to achieve these goals is to study surnames. A worldwide characteristic of human societies is the economic and reproductive aspects of life organized through a particular *family* system. Most familial organizations present surname inheritance, a family name from parents to children to determine kinship between different persons. For this reason, from the Middle Ages to the present, surnames have been a powerful information source. By studying the frequency and distribution of the surnames, demographic aspects of a society can be known, such as marriage preferences or inbreeding tendencies. Early surname studies have been popular because, in some instances, like surnames passing through the male line, they were considered an inexpensive proxy for Y-chromosome profiling. But in addition to this particular use, surnames also carry social and cultural information that adds potential value in many interdisciplinary approaches,[8] as a proxy for ancestry origin.[6]

In this work, we introduce Bulsarapp,[a] a web-based solution for Argentinean surnames exploration and demographic approaches. The aim is to simplify isonymic research tasks, like surname classification by geolinguistic origin, or surname distribution analysis, facilitating these tasks, and enabling them to be performed in feasible times. The tool provides several juxtaposed linked views, each specifically designed to focus on a given analysis task. Preliminary reports by expert users provided a strong positive feedback, and scientists in some research groups in Argentina already adopted the tool as an aid for their activities.

[a]Our humble tribute to Freddy Mercury.

RELATED WORK

According to O'Connor,[14] surnames are special ethnic traits. Contemporary societies maintain continuous administrative records, and the family names transmission is mostly regular. Due to these characteristics, surname research has a long tradition in Europe. For this reason, some online projects involve surname analysis using varied information sources, mainly telephone directories, to create surname distribution maps. These maps are increasingly used in human geography, for example, to characterize cultural regions. In most cases, these representations use simple statistics to study data distribution, and also provide basic visualization functions. There are country-specific visual tools, such as the Italian Mappa Dei Cognomi,[b] and similar surname maps for Germany, Spain, France, the United Kingdom, the Netherlands, Belgium, or Romania. Also, the Longley, Singleton, and Mateos "Public Profiler Project"[c] allows us to create a global map of the distribution of a specific surname. Cheshire and Longley[5] used electoral rolls to analyze the spatial distribution of surnames. With an enhanced version that identifies the names and precise locations of 41.6 million people. The authors also clustered surnames within the Great Britain territory.

Furthermore, there are projects with an interactive visualization for more complex analyses, like the Name Profiler Toolkit site, designed to enable the interactive exploration of forenames and surnames in the United States.[20] The authors developed a methodology using probability distributions of categorical spatial data and different indices estimation to test and compare several hypotheses, for instance immigration distribution. In China, Meng *et al.*[12] worked with the distribution of 100 popular surnames and their spatial correlation. The objective was to determine the relationship between spatial distribution and regional characteristics (geographical, cultural, etc.).

When people migrate, they also carry their surnames, and it is possible to analyze the movements of family names throughout a territory. In combination with spatial information, surnames are a rich resource for a variety of socioeconomic, geographic, and cultural studies. The study of family names, also called isonymic studies, helps us improve the understanding of demographic dynamics and historical and contemporary migration. However, isonymy as a fundamental aspect of population studies, is not covered in any of the projects mentioned above.

[b]https://www.mappadeicognomi.it/
[c]http://worldnames.publicprofiler.org/

In 1965, the geneticists Crown and Mange established the principles of the isonymic method, which allows us to obtain a kinship measure even in the absence of genealogical information, particularly in large populations or entire countries.[15] Isonymic analyses can be developed from different sample sizes. Thus, we can characterize a territory at different levels of political organization, for example, the entire country, a region, a state or province, a department or county, etc. The population structure of several Latin American countries (Argentina, Uruguay, Honduras, Chile, Venezuela, Brazil, among others) was analyzed through isonymy studies using electoral rolls, working with data from millions of people in different administrative districts. Since voting is mandatory in these countries, electoral rolls constitute a significant population sample, reaching 70% or more of the total number of inhabitants, and are updated on an annual basis. The most notable examples based on electoral registers in Argentina are the works of Dipierri *et al.*[9] in country-wide analyses, or Bronberg *et al.*[4] focusing only on the capital city and most densely populated area. Other studies are aimed to identify genetic isolates and their spatial distribution, analyzing possible health consequences of inbreeding,[10] or relationships between demographic and economic indicators with geographical surname distribution.[11] These analyses were developed with general-purpose software, which in general is not entirely appropriate for such identification tasks. Also, there is no specific software tool available for this kind of data. In this context, information visualization may enable a different level of complexity, which requires a combination of perspectives. First, a *vertical* perspective referring to region ratios or levels (national, regional, provincial, and departmental). Second, a *horizontal* perspective that includes the set of isonymic indicators obtained for each of the different population samples within each region/subregion. Even in medium-size population countries, the heterogeneity and amount of data necessary to arrive at a feasible understanding requires developing novel tools.

A NOVEL APPROACH FOR STUDYING ISONYMY

The approaches presented above offer mainly two different types of surname analysis tools. In the first group, some apply visualization techniques to present simple information such as surname distribution, geographical references, and basic statistics. A second group includes more complex analyses for surnames using different indices, showing results in static charts or graphs that are neither interactive nor actionable. Bulsarapp, in turn, also incorporates several other analyses, including novel aspects like applying simultaneously the isonymic method to obtain multiple and related population indices, the estimation of surname geolinguistic origins and their spatial distribution. All these analyses are integrated into an interactive cartographic presentation.

Design Goals

Bulsarapp was developed along interacting with scholars and domain experts in surname studies. These interviews and interactions allowed us to identify the most common research tasks and problems and to define three main design goals.

G1. Isonymy Trends and Relationships Exploration: The system must allow the data exploration at diverse scales and the isonymic values calculation at different territorial units, their spatial distribution, the analysis of relationships between them, and identify possible extreme values. Even for studies regarding previous years' datasets, researchers could use the tool to quickly obtain a dynamic context and focus their efforts to hypothesize how a given demographic scenario may impact future years.

G2. Surname Trace: Surnames can be grouped according to several criteria: those that share a common etymological derivation or source (based on toponymy, occupations, nicknames, or physical characteristics), their traceability to a specific geographical origin, by their common occurrence in history, or even because they can be associated with different ancestry groups to a greater or lesser degree. The visualization of the distribution of surname groups is essential to characterize and complement isonymic studies, and therefore the tool should allow these queries to be carried out. Researchers can discover an interest group of surnames and then focus their efforts on conducting specific analyses for in-depth knowledge of certain aspects of the population that bears a surname from that set, especially those analyses related to inherited diseases.

G3. Surname Origin Trend: Like other Latin American countries and former colonies around the world, Argentina has an admixture population, entailing a high surname diversity. This tool then must allow researchers to explore the frequency of a specific geolinguistic origin in a region and contextualize it concerning the other origins. For this, we defined and designed the specific views, together with their interaction and coordination required to meet these goals. The study of surname spatial distribution may help broaden the understanding of migratory processes, one of the fundamental variables regarding population dynamics.

Surname Data

Our primary database is the 2015 Argentinean electoral register. Argentina is the second largest country in South America, with 40,117,096 inhabitants, according to the 2010 National Census. The electoral roll contains data for 30,530,194 people. Even though this register has not been compiled for geographic analysis purposes, its spatial information can be reconstructed through geo-referencing. Each record contains 13 fields, including family name, gender, and information about the specific department and province in which each person is registered to vote. According to Argentine electoral legislation, this must coincide with the place of residence.

For the experiments regarding surname geolinguistic origins, we used two lists[1,13] with names already classified. Collectively, we obtained 65,023 surnames associated with any of 28 geolinguistic origin labels (like German, Spanish, Polish, etc.), including native Latin American origins. As a final input, we used a dataset with the hierarchy of territorial units in Argentina. This political organization distinguishes five regions, 24 provinces, and 528 departments.[d]

FROM SURNAMES TO ISONYMIC INDICATORS

Unlike other Latin American countries, the use of a single surname predominates in Argentina, inherited through the paternal line. Since 2015, newborns can be registered with maternal and paternal surnames, though this is not mandatory and is still infrequent. For this reason, in this work, we only considered the first surname. With the electoral register information, we calculated isonymic indicators from which researchers infer the population structure: total number of voters and different surnames, random isonymy, Fisher's alpha index, consanguinity coefficient, A index, B index, and Lasker's distance.

Random Isonymy (I_{NS}) is interpreted as the probability that two surnames, randomly extracted, are identical because they have been inherited from both parents and can predict the inbreeding frequency in a given region. It was calculated according to the following equation:[17]

$$I_{\mathrm{NS}} = \sum_k \left(\frac{N_{ki}}{N_i}\right)^2 - \left(\frac{1}{N_i}\right) \tag{1}$$

where N_{ki} is the relative frequency of the surname k for the region i and N_i is the size of the population in the region i.

[d]https://www.indec.gob.ar/ftp/cuadros/territorio/n010301.xls

Fisher's alpha index, according to Barrai *et al.*'s work,[3] is the inverse of random isonymy

$$\alpha = \frac{1}{I_{\mathrm{NS}}} \tag{2}$$

This value estimates the number of surnames with an equal frequency. A small alpha value indicates large genetic drift, whereas a large value indicates migration.

The consanguinity coefficient (F) is also related to the random isonymy, and according to Crow and Mange's work[7] it can be expressed as

$$F = \frac{I_{\mathrm{NS}}}{4}. \tag{3}$$

In turn, A and B indices illustrate, respectively, the percentage of the population with only one representative per surname, and the percentage of the population with the seven most frequent surnames. These indices, established by Rodriguez–Larralde,[16] are a useful tool for analyzing the demographic trend of each population over time. High A index values can be found in situations of expansion and rapid population growth (as a result of economic opportunities, such as the development of industries with great labor demand). Likewise, low A index values can be associated with stagnated communities, e.g., small towns or those short with economic or cultural perspectives, in which the population of productive age tends to migrate to larger urban areas in search of better opportunities. In these cases, the residents that stay generally belong to older age groups who no longer have children and remain as the only representatives of a surname. In the case of A index, it is necessary to have a demographic context to make a correct interpretation. In turn, high B index values are observed for isolated or small communities with high emigration rates, in which only few surnames were found in the majority of the population. Migration rates have a direct impact on this indicator.

Finally, Lasker's distances between regions (departments and provinces), according to Rodríguez-Larralde *et al.*'s work[18] can be evaluated as:

$$\mathrm{Lasker}(i,j) = -\log\left(\sum p_{ik}p_{jk}\right) \tag{4}$$

where p_{ik} and p_{jk} are the frequencies of surname k in the $i\mathrm{th}$ and $j\mathrm{th}$ regions, respectively. In normal demographic settings, Lasker's distance is linearly and significantly correlated with the the log of geographic distance, and thus any alteration of this trend is representative of specific population dynamics.

SYSTEM OVERVIEW

The raw data, described in the "Surname Data" section, is processed through different transformations

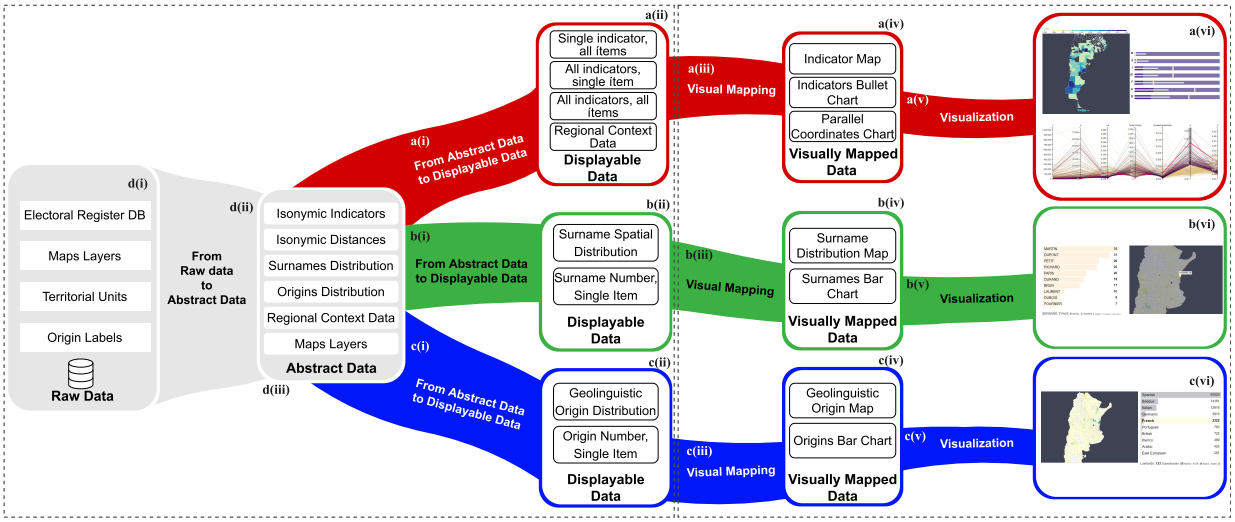


FIGURE 1. Pipeline description. The left dotted box defines the actual datasets to be displayed. The right defines how it will be displayed. The red, green, and blue branches represent transformations and the boxes represent the resulting data stages.

until reaching the final visualization presented to the user (see Figure 1). From the electoral roll, a data cleansing task extracts the relevant attributes: family name, department code, department description, province code, and province description. From the family name, composed of one or many surnames, we preserved the first one. We used the attributes corresponding to the polling place to geolocate each record. The Geographic Data Standardization Service provided by the Ministry of Modernization of Argentina was used for this. As a result, we assigned two codes for each territorial level (departmental and provincial) for all records. A population segment obtained from this coding is a "territorial unit," and within each one, we applied a sequence of calculation tasks and obtained all isonymic indicators [applying (1), (2), and (3)], Lasker's distance matrices for provinces and departments [see (4)], and the total number of bearers of each surname; Total surnames by geographical-linguistic origin required a previous step of assigning an origin label (if possible) to all electoral registry surnames (such as German, Spanish, Polish, Latin American native, etc.). Finally, extreme value identification and isonymic indicators required calculation for territorial unit groups. These groups depend on the administrative division. A department may belong to one of the major sets of departments inside a province. At the same time, the department may belong to one of the five main national regions. The same hierarchy applies to the provinces, in which the user may be able to select a group of those that make up one of the five regions, or the entire country. In any of these cases, the user selects the territorial units for the defined set, minimum, maximum, and average value of each of the seven isonymic indicators are computed.

All these tasks generate another body of data (abstract data) from which the different branches of our pipeline follow (see Figure 1) related to the three design goals. Branch "a" (in red) is related to the processing and visualization tasks of isonymic indicators, branch "b" (in green) is related to surname distribution tasks, and branch "c" (in blue) is associated with the origin distributions tasks. These three tasks for which the app provides support are described below.

Isonymic Indicators Task

Isonymic indicators and their context values make up three different sets of data to visualize, (branch "a (ii)" in Figure 1). Isonymic indicators for all territorial units are visualized in a parallel coordinates view. This allows users to overview first all indices and then to filter and see details on demand (visualization mantra) to explore relationships between them. The axes may be rearranged to facilitate the exploration of subjects of interest between two or more indicators. It is also possible to establish ranges on each axis to limit the isonymic values and determine how this operation impacts the behavior of the other indices. When ranges are indicated, a filter criterion is used. The items outside the criteria are no longer colored on the map (only their borders are preserved). In this way, the users can geographically visualize the specified set of data relationships. A more detailed graph view is

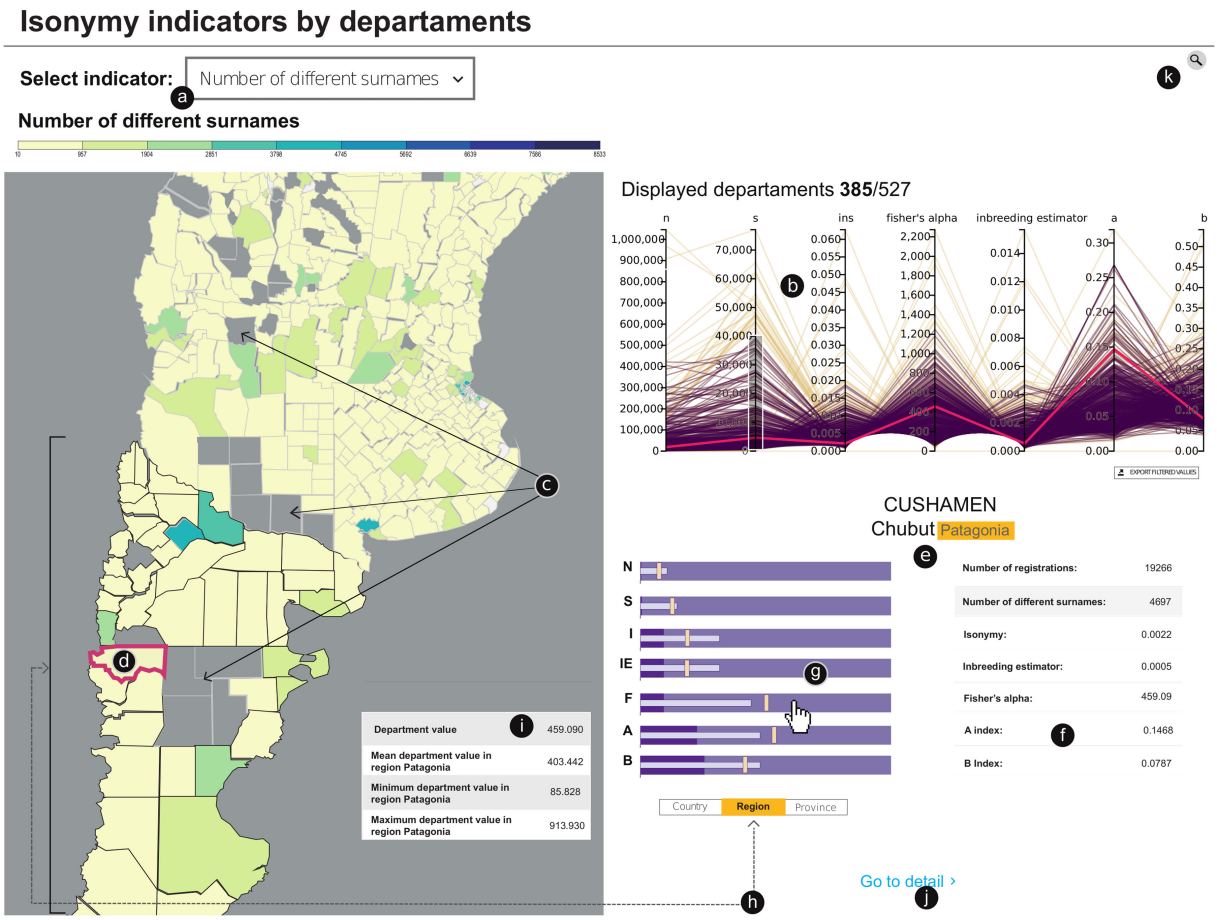


FIGURE 2. General view of isonymic information at the Argentinean departmental (county) level in the Bulsarapp web application. (a) The user can interact with the map by choosing any one of the seven isonymic indices from the selector and/or (b) setting ranges in the axes of a parallel coordinates chart. A range of values was established on the second axis. Lines within the range are darker. (c) Those territorial units that do not fall within the ranges of values specified by the user are left unpainted in the map, and only their contours are preserved. (d) When a particular region is selected, (e) a detail is displayed including the department name, the province, region to which it belongs, and (f) their respective values of isonymic indices. The line corresponding to the selected territorial unit on the map is highlighted in a different color in the parallel coordinates view. (g) A bullet chart allows the user to compare these values with the maximum, minimum, and mean values in the neighborhood. The longest bar represents the maximum value of the indicator in the selected neighborhood. The darkest bar with the same thickness represents the lowest value of the indicator in the neighborhood. The thinnest bar represents the mean value. The bullet represents the indicator value for the selected territorial unit. (h) The user can choose this neighborhood from three options: country, region, or province. The context area is highlighted with darker borders on the map. When the user hovers over any of the bars on this bullet chart, (i) a tool-tip is displayed with the bar's reference values. (j) A button is presented to move to the detail section. (k) A search button is displayed on the upper right corner to allow searching by department/province name.

shown in Figure 2. The map can be colored according to the places whose values fall within the criteria. This is extremely useful, for example, to detect and identify isolated populations (a case study will be presented in the "Usability Study" section).

Region-wise isonymic indicator values are represented over the territorial units using a choropleth map. This map allows users to detect spatial patterns in the specified parameters. When the user chooses another indicator in the selector, it is colored according to the new option. Finally, we placed all indicators values for a specific territorial unit on a bullet chart. Here, the context values depend on the selected geographical hierarchy (provincial, regional, or national).

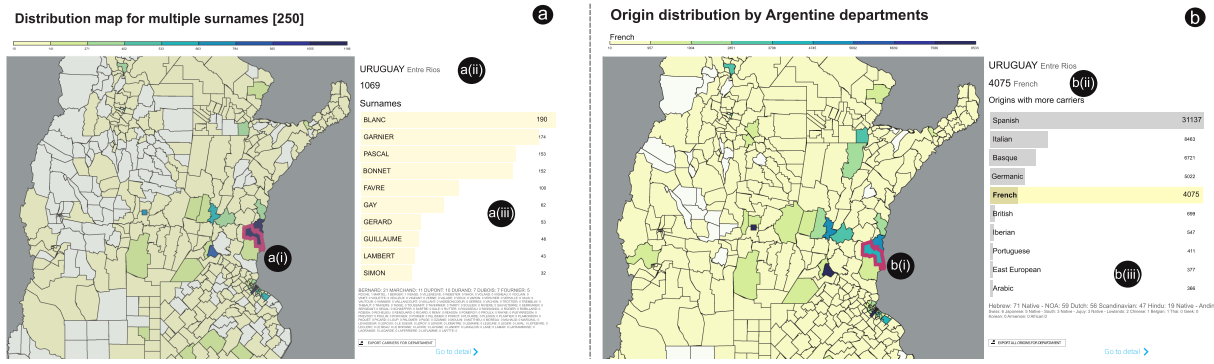


FIGURE 3. (a) Shows the surname distribution view. In this example, the view shows the result of a query for carriers of a set of 250 surnames of French origin collected from the web. a(i) When the user selects a unit in the map, the tool shows a detail. a(ii) This detail shows the name of the selected element, the region/province to which it belongs and the total number of carriers consulted. a(iii) Below is a list ordered by carriers of the consulted surnames. (b) On the right, shows the visualization of the distribution of surnames labeled with French origin. b(i) When the user selects a territorial unit, the tool displays a detail in a similar way to the distribution view. b(ii) The detail displays the number of surname bearers with the selected origin. b(iii) Shows a ranking of origins present in the unit (with horizontal bars). Both views are at a departmental level, and the spatial distribution pattern is similar. In both, the detail is shown for the "Uruguay" department, "Entre Ríos" province.

The user can characterize each item on the map with the corresponding isonymic maximum, minimum, and mean values present in its neighborhood. These context values are drawn in different lengths bars. The marker (the bullet) is located with the value for the item adequately selected on the map.

A selector with three options (provincial, regional or national) is located below the graph, with which the user can interact and change as desired. The items' edges in each neighborhood are highlighted on the map to provide geographic context. Finally, in this view, if the user clicks on a map item, the bullet chart is updated with new markers. Also, the corresponding line in the parallel coordinates graph is highlighted with an appropriate color. These three views are linked in a coordinated way.

Surname Distribution Task

The user first specifies a surname set (one or many) to explore their spatial distribution in this task. Then two plots constitute the obtained view, a choropleth map together with a bar graph. The map is painted with a color code according to the number of people carrying any of the consulted surnames. A summary table with the provinces and departments with the highest number of surname bearers is displayed next to the map. When the user selects an item on the map, the bearers number of each surname in the query is displayed. This list, sorted by prevalence, is presented in the horizontal bar graph. Figure 3(a) shows an example of this task.

Origins Distribution Task

In this task, the tool generates a gradient map colored by frequency. The data to visualize is the total number of surname bearers classified and labeled according to a geolinguistic specific origin, and the total number of different surnames for each origin. When a department or region is selected from the map, a summary with the most frequent surnames of the given origin in the selected region is displayed in a horizontal bar chart. A drop-down dialog allows the user to change the origin displayed, and the territorial unit analyzed. Figure 3(b) shows an example of this interaction. The highlighted bar in this graph corresponds to the origin that the user has selected to paint the map. As a final comment, the user can work with both geographic divisions (provinces or departments) in any of these three tasks. The user can access a detailed report of each element selected on the map in any of these divisions. This detail section takes information from each of the three main tasks described above, incorporates Lasker's distances calculated (from the abstract data) and presents a final report as shown below the bar chart in Figure 4(a).

USABILITY STUDY

Current isonymic studies reviewed to date mainly consider static and/or generalized reports about surname information. Therefore, if the objective is to characterize a population by its isonymy, and it is not included in the available report, researchers must have access to the data and perform the necessary calculations manually.

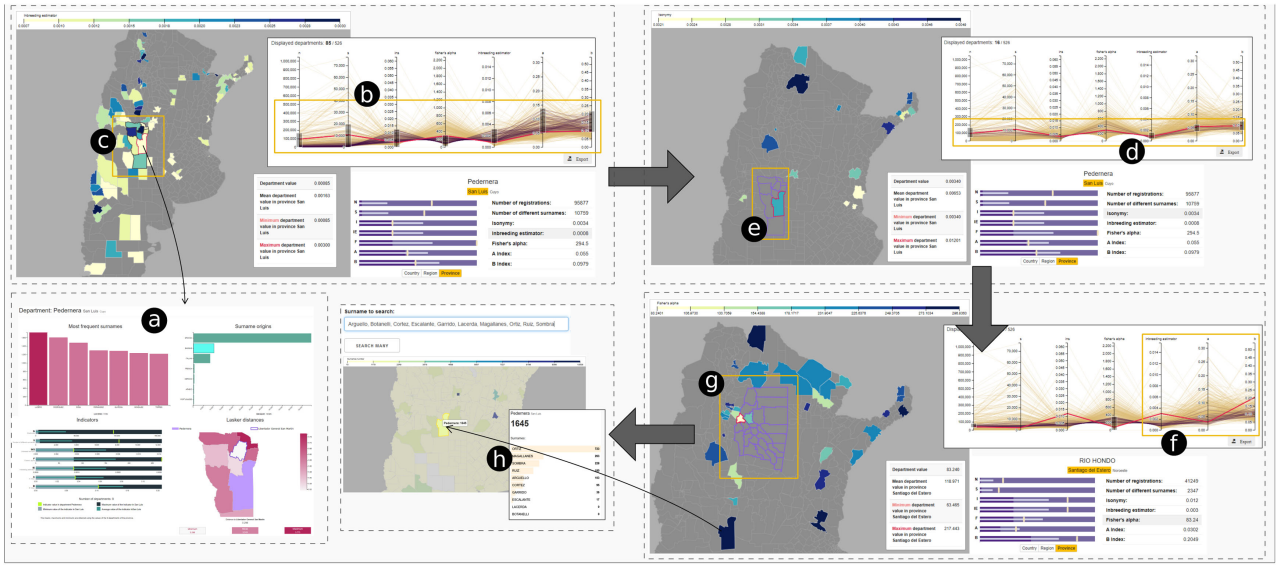


FIGURE 4. (a) Case study: Detection of population isolates. Details of department under study are inspected in the department view map. The maximum and minimum of each isonymic indicator are obtained for the province to which this department belongs. (b) Each axis on the parallel coordinate graph is bounded with those values. (c) Only the values of the filtered departments are represented in the map. Researchers may question themselves how much they know about the filtered populations: *e.g., "Is it possible to explain the resulting drawing on the map?"* (d) Then, on the parallel axes, each range is shortened with limits close to the indicators values for the department under study. (e) It is observed that there is no other department in the province with similar characteristics. In the map, other Argentinean departments with similar demographic features are drawn. (f) The restrictions on the axes "Inbreeding estimator," "A Indicator" and "B Indicator" are eliminated to see how many departments in the country are still rendered. The department with the lowest Fisher's Alpha indicates a decrease in variability, mainly due to genetic drift. It is a change between generations that happens due to random events. Although genetic drift happens in populations of all sizes, its effects tend to be stronger in small populations. (g) To contextualize this information, data from the other departments of the province are analyzed. (h) The study is furthered looking for the distribution of a set of surnames associated with a recessive disease present in this territorial unit.

To measure the general query time taken for our tool, we developed a series of experiments in order to navigate through the three sections of Bulsarapp. It comprises 20 tasks (T), each one with an associated multiple-choice question, whose correct answer was achieved using the tool (see Table 1). These questions aimed to characterize populations (provinces and departments) through their isonymic values, trace a group of surnames in a specific territorial unit and inquire about the surname origins in other areas. Domain experts completed the questionnaire as they conducted this experimentation of the Bulsarapp prototype. Finally, to obtain usability data, we ask the researchers to complete a survey with a classic system usability scale (SUS) and their set of Likert-scale questions.

In the experimentation step, the tasks were grouped according to the functionalities described in the "System Overview" section: isonymic indices (T1–T6), isonymy detail section (T7–T13), surname frequencies (T14–T18), and surname distribution by geolinguistic origin (T19, T20). The number of questions answered correctly can give us an indication of the effectiveness of the tool. To track the application efficiency, the time to complete all tasks was also measured. The experiment was carried out by nine participants (Female = 4, Male = 5, average age = 35). Seven of them answered the 20 questions correctly, while only two volunteers answered 19 correctly. The time it took to obtain specific data, for specific territorial units, was measured by users themselves, interacting directly with the application for the first time. The average time to complete the set of 20 tasks was 28 minutes and 30 seconds. In later sessions, knowing the mode of interaction, the required time decreased. The average SUS score obtained was 88.43, which makes us to consider that the application has good usability.[2]

After the usability study, and as a case study, we asked domain experts to identify any scenarios in which they could use Bulsarapp in the short term. One

TABLE 1. User tasks requested for usability study.

ID	Description	Options
T1, T2	Indicate which is the country region with highest values for a given index.	"To the south of..." or "to the north of..."
T3, T6	Indicate the value of a index for a given province/ department	Three numerical options.
T4, T5	Indicate if the set of values of all the indicators exceeds a specific threshold.	True or false.
T7	Indicate which are the most popular surnames in a department.	Three lists of surnames.
T8	Indicate how bearers have a specific surname.	Three numerical options
T9	Indicate the number of people who have a surname classified according to a specific origin	Three numerical options.
T10, T11	Compare department index values with the province average values.	True or false.
T12	Indicate if, for a specific department, the lowest Lasker's distance value relates it to a neighboring department.	True or false.
T13	Determinate if the Lasker's distance between two specific departments corresponds to the highest value.	True or false.
T14, T15	Given a surnames set, determine how many people bears them, in a particular department/province.	Three numerical options.
T16	Given a specific surname, determine the bearers percentage of total country population.	Three numerical options (percentages).
T17	Given a surnames set, determine which are the three departments/provinces where can be find most.	Three lists of regions.
T18, T19	Determine the bearers number for a given surname in a particular department.	Three numerical options.
T20	Determine if the surname origin sought corresponds to one of the most popular in terms of number of bearers.	True or false.

of the most prominent is the *population isolates detection*, in which using prior health databases, the researchers follow a plausible route for identifying a possible population isolate. In this scenario, the experts began their exploratory path by focusing on a well-known Argentine department. In it, they had already detected an isolated population. First they worked in the provincial order. They inspected the isonymic values of containing province of the department and used them to establish the ranges on the parallel axes. In this way, they sought to quickly determine how similar the country's departments are with those of the province under study. In a second step, small ranges were established, close to the isonymic values registered in the well-known department. Again new areas are highlighted on the map. In a third step they searched for a set of 11 surnames of families affected by an autosomic recessive disease, which had been identified in the health databases.

The complete exploratory path is illustrated in Figure 4. In the first two steps, the visualization through dynamic maps allowed to easily contextualize the department of interest and its neighbors, and to collect questions about their migratory past. For example, questions emerged such as: *How was the region historically occupied?*; *Where did the migrants come from?*; *What were the cities that most attracted settlers?*; *How has that trend changed over the last century?* In the third step, analyzing the spatial trend of a group of surnames, they found that this disease had a striking frequency in this specific region of the country, coinciding with the isonymic values that indicate a tendency to isolation and inbreeding, focusing on a possible correlation between variables.

DISCUSSION AND CONCLUSION

The proposal presented here ranges all phases of the framework proposed by Sedlmair *et al.*[19] After learning about the domain and processing the data, we deployed a web application prototype, with which we obtained a valuable exchange with domain experts. The use of interactive visualization tools like the one presented in this work appears to be a promising path to support multidisciplinary studies. Bulsarapp enables collaborative studies and broadens the end-users' research spectrum, enriching the subsequent investigation of human population dynamics (such as structure, migration, and social interactions), and provides visual representations of surnames through linked views to help isonymic researchers perform their tasks more efficiently. To the best of our knowledge, available tools or methodologies neither allow us to perform these tasks easily, nor offer the possibility of interactive visualization. In particular, the tool allows quick querying of surname information from three linked perspectives, isonymic structure, frequency of

surnames from a specified variable set as required by the user, and number of bearers of surnames from a particular geolinguistic origin. Furthermore, spatial exploration is possible by analyzing territorial units at the departmental and provincial levels. For the isonymic view, Bulsarapp allows us to explore isonymic values, outlier detection, and spatial trend descriptions. At the same time, it depicts the relationships between isonymic values for a territorial unit within a given ratio. Moreover, the tool allows the user to indicate restrictions in the isonymic indicators' values and quickly obtain the distribution in the territory of those units that comply with them. The surname frequencies view allows users to discover spatial patterns. Users can then easily compare the bearers' number of each queried surname, their geolinguistic origin classification, and territorial distribution.

We performed a usability study to evaluate the effectiveness and feasibility of our tool. For this, we measured the completion times of a series of tasks relevant to its specific use by scientists. The evaluation sessions yielded a satisfactory usability score according to the standard metrics. Domain experts also stated that Bulsarapp is a useful tool for detecting easily human populations that tend to move away from equilibrium and/or ideal conditions and explore their context variables. In contrast to their current work style, Bulsarapp allowed researchers to quickly carry out exploratory queries triggered by their curiosity, obtaining isonymic data, geographically characterized, and visually contextualized. Without this tool, an exploratory-confirmatory cycle like the one carried out in the presented experimentation becomes very lengthy and sometimes cumbersome. Among other advantages, the tool significantly accelerates the access to relevant information regarding isonymic structures, the distribution of surnames, and their respective origins. The users reported that the presentation provided through a web application also facilitates open access to this information, which enables researchers in other groups and new stakeholders to interact in new and constructive ways. Their final remarks are that for researchers in human population dynamics and related issues in the region, it is essential to develop novel and flexible research workflows.

The scientists participating in these experiments quickly adopted this visualization tool. They pointed out that the linked views significantly reduced the time and cognitive effort priorly required in some of their frequent research tasks. For instance, the coordinated view that represented together the choropleth map and the parallel coordinates facilitated a repetitive verification task that is necessary to constantly corroborate the geographical disposition of the regions that meet a given criterion of isonymic indicators. In our views, territorial units can be either hidden or presented, and thus this functionality enables to drill-down the understanding of the isonymic indicators. The researchers highlighted that the usefulness of these specific coordinated views increases as much as the researchers' knowledge of the country's population deepens. They also discovered a wider range of possibilities in using the application (isonymic indicators, distribution of surnames, and distribution of origins). As such, the use of linked views in different locations were important features for exploratory analysis tasks performed by the domain experts, which expanded the possibilities of reaching novel and useful results. This initial approach paves the way to a more thorough design study to visualize demographic structure information extracted from surnames.

Future work is aimed to facilitate interactive exploration of other demographic and population dynamics metrics, such as isolation, developing more extensive measures, and distance definitions. Furthermore, we will incorporate isonymic information for new spatial distribution and population division levels, for example, census circuits or town-wise. We are extending the underlying information processing platform to facilitate incorporating new data types and sources, and other isonymic distances available in the bibliography. The addition of machine learning features is also being considered as an aid in several contexts, for instance the use of recommendation systems may guide novel researchers in organizing their data gathering and presentation tasks. Additionally, we are performing further tests regarding the visualization pipeline, such as new linked views, and studies in the effectiveness of the choropleth maps.

ACKNOWLEDGMENTS

The authors would like to thank the researchers participating in the experimentation of the web tool and the Cámara Nacional Electoral Argentina authorities for providing the 2015 electoral register copy used in this work. They would also like to thank the developers of the open source technologies used in this work, which allowed them to significantly accelerate their development. This work was supported by grants and scholarships from the Argentine National Scientific and Technical Research Council (CONICET).

REFERENCES

1. M. E. Albeck, E. L. Alfaro, J. E. Dipierri, and E. R. Chaves, "Los apellidos de salta en el siglo xxi: Origen geo-lingüístico, diversidad y frecuencia," *Andes*, vol. 2, no. 28, pp. 1–20, 2017.

2. A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean: Adding an adjective rating scale," *J. Usability Stud.*, vol. 4, no. 3, pp. 114–123, 2009.
3. I. Barrai, C. Scapoli, M. Beretta, C. Nesti, E. Mamolini, and Á. Rodríguez-Larralde, "Isonymy and the genetic structure of Switzerland i. The distributions of surnames," *Ann. Hum. Biol.*, vol. 23, no. 6, pp. 431–455, 1996.
4. R. A. Bronberg *et al.*, "Isonymy structure of Buenos Aires city," *Hum. Biol.*, vol. 81, no. 4, pp. 447–461, 2009.
5. J. A. Cheshire and P. A. Longley, "Identifying spatial concentrations of surnames," *Int. J. Geographical Inf. Sci.*, vol. 26, no. 2, pp. 309–325, 2012.
6. S. E. Colantonio, G. W. Lasker, B. A. Kaplan, and V. Fuster, "Use of surname models in human population biology: A review of recent developments," *Hum. Biol.*, vol. 75, no. 6, pp. 785–807, 2003.
7. J. F. Crow and A. P. Mange, "Measurement of inbreeding from the frequency of marriages between persons of the same surname," *Eugenics Quart.*, vol. 12, no. 4, pp. 199–203, 1965.
8. P. Darlu *et al.*, "The family name as socio-cultural feature and genetic metaphor: From concepts to methods," *Hum. Biol.*, vol. 84, no. 2, pp. 169–214, 2012.
9. J. Dipierri, E. Alfaro, C. Scapoli, E. Mamolini, A. Rodriguez-Larralde, and I. Barrai, "Surnames in Argentina: A population study through isonymy," *Amer. J. Phys. Anthropol.*, vol. 128, no. 1, pp. 199–209, 2005.
10. J. Dipierri *et al.*, "Random inbreeding, isonymy, and population isolates in Argentina," *J. Community Genet.*, vol. 5, no. 3, pp. 241–248, 2014.
11. J. E. Dipierri, A. Rodríguez-Larralde, I. Barrai, E. G. Redomero, C. Alonso-Rodríguez, and E. L. Alfaro, "Consanguinity by random isonymy and socioeconomic development in Argentina: A population study," *J. Biosocial Sci.*, vol. 49, no. 3, pp. 322–333, 2017.
12. J. Meng, H. Chen, X. Liang, and J. Yan, "The empirical study of the spatial distribution of chinese surnames," in *Proc. IEEE Int. Conf. Cloud Comput. Big Data Anal.*, 2016, pp. 398–403.
13. L. Monasterio, "Surnames and ancestry in Brazil," *PLoS One*, vol. 12, no. 5, 2017, Art. no. e0176890.
14. M. I. O'connor, *Descendants of Totoliguoqui: Ethnicity and Economics in the Mayo Valley*, Univ. California Press, vol. 19, 1989.
15. M. Prost, G. Boëtsch, and E. Rabino-Massa, "The limitations of the isonymic method. From the model to an actual application on a computerized population register (vallouise 1350-1899)," *Int. J. Anthropol.*, vol. 20, no. 3/4, pp. 207–224, 2005.
16. A. Rodriguez-Larralde and I. Barrai, "Estudio genetico demografico del estado zulia, venezuela, a traves de isonimia," *Acta cientifica Venezolana*, vol. 49, no. 3, pp. 134–143, 1986.
17. A. Rodríguez-Larralde, G. Formica, C. Scapoli, M. Beretta, E. Mamolini, and I. Barrai, "Microevolution in perugia: Isonymy 1890-1990," *Ann. Hum. Biol.*, vol. 20, no. 3, pp. 261–274, 1993.
18. A. Rodríguez-Larralde, J. Morales, and I. Barrai, "Surname frequency and the isonymy structure of Venezuela," *Amer. J. Hum. Biol., Official J. Hum. Biol. Assoc.*, vol. 12, no. 3, pp. 352–362, 2000.
19. M. Sedlmair, M. Meyer, and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2431–2440, Dec. 2012.
20. F. Wang, B. Hansen, R. Simmons, and R. Maciejewski, "Name profiler toolkit," *IEEE Comput. Graphics Appl.*, vol. 37, no. 5, pp. 61–71, Sep. 2017.

LEONARDO MORALES is a doctoral fellow of the National Scientific and Technical Research Council, Buenos Aires, Argentina. His research interests include the visualization of information and machine learning techniques applied in the study of surnames, to improve the knowledge of the Argentine demographic structure. Morales received the bachelor's degree in information systems. He is the corresponding author of this article. Contact him at lmorales@cenpat-conicet.gob.ar

PABLO NAVARRO is a postdoctoral fellow of the National Scientific and Technical Research Council, Buenos Aires, Argentina. His research interests include image processing and artificial intelligence. Pablo Navarro received the Ph.D. degree in engineering from National Technological University, Buenos Aires, Argentina, in 2021. Contact him at pnavarro@cenpat-conicet.gob.ar

CELIA CINTAS is a Research Scientist at IBM Research Africa, Sandton, South Africa. She is a Member of the AI Science Team at the Kenya Lab. Her research interests include improving machine learning techniques to address challenges in global health, deep learning, and anomalous pattern detection. Cintas received the Ph.D. degree in computer science. Contact her at celia.cintas@ibm.com.

ROLANDO GONZÁLEZ-JOSÉ is currently a Principal Investigator at the Patagonian Institute of Human and Social Sciences (CONICET), Puerto Madryn, Argentina. He is also the Director of the National Patagonian Center for Research (CENPAT). His research was published in multidisciplinary journals such as *Nature, PNAS, Evolution, Nature Communications, Scientific*

Reports, and disciplinary journals such as *American Journal of Physical Anthropology, American Journal of Human Biology, Journal of Anatomy*, and *PloS Genetics*, among others. His research interests are focused on the evolution of modern cosmopolitan Latin American populations, with emphasis on the fine-scale structure of genetic and nongenetic variation. González-José received the Ph.D. degree in biological anthropology from the University of Barcelona, Barcelona, Spain, in 2003. He was the recipient of the Houssay Prize in social sciences in 2017, and the Argentinean Senate's honorific mention in science and technology in 2019. Contact him at rolando@cenpat-conicet.gob.ar.

VIRGINIA RAMALLO is a Physical Anthropologist, CNPq postdoctoral fellow with the Federal University of Rio Grande do Sul, Porto Alegre, Brazil. She is also a researcher at the Patagonian Institute of Human and Social Sciences (CONICET), Puerto Madryn, Argentina. Her interest is the study of interaction between the socio-cultural context and the biological history of Latin American populations, based on multiple data source: genealogies, genetic analysis, health records, census, and surname analysis. Ramallo received the Ph.D. degree in natural sciences from The La Plata National University, La Plata, Argentina. Contact her at ramallo@cenpat-conicet.gob.ar.

CLAUDIO DELRIEUX is a Fulbright Postdoctoral Fellow with the University of Denver, Denver, CO, USA. He is also Full Professor and PI with the Electric and Computer Engineering Department, Universidad Nacional del Sur, Bahía Blanca, Argentina, a fellow of the National Council of Science and Technology of Argentina (CONICET), and the Chair of the Imaging Sciences Laboratory. He is author of more than 90 Scopus indexed papers, and more than 100 refereed international conference papers. His current interests include image and video processing, computer graphics, scientific visualization, and artificial intelligence. Delrieux received the B.S. degree in electric engineering and the Ph.D. degree in computer science from the Universidad Nacional del Sur. Contact him at cad@uns.edu.ar.

Journal of *Imaging*

MDPI

Article

body2vec: 3D Point Cloud Reconstruction for Precise Anthropometry with Handheld Devices

Magda Alexandra Trujillo-Jiménez [1,2,*], Pablo Navarro [1,2,3], Bruno Pazos [1,2,3], Leonardo Morales [1,2,3], Virginia Ramallo [2], Carolina Paschetta [2], Soledad De Azevedo [2], Anahí Ruderman [2], Orlando Pérez [2], Claudio Delrieux [1] and Rolando Gonzalez-José [2]

1 Laboratorio de Ciencias de las Imágenes, Departamento de Ingeniería Eléctrica y Computadoras, Universidad Nacional del Sur, and CONICET, Bahía Blanca B8000, Argentina; pnavarro@cenpat-conicet.gob.ar (P.N.); bpazos@cenpat-conicet.gob.ar (B.P.); lmorales@cenpat-conicet.gob.ar (L.M.); cad@uns.edu.ar (C.D.)

2 Instituto Patagónico de Ciencias Sociales y Humanas, Centro Nacional Patagónico, CONICET, Puerto Madryn U9120, Argentina; ramallo@cenpat-conicet.gob.ar (V.R.); paschetta@cenpat-conicet.gob.ar (C.P.); deazevedo@cenpat-conicet.gob.ar (S.D.A.); ruderman@cenpat-conicet.gob.ar (A.R.); orlandoperez@cenpat-conicet.gob.ar (O.P.); rolando@cenpat-conicet.gob.ar (R.G.-J.)

3 Departamento de Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco, Trelew U9100, Argentina

* Correspondence: mtrujillo@cenpat-conicet.gob.ar; Tel.: +54-9291-4261021

Received: 31 July 2020; Accepted: 31 August 2020; Published: 11 September 2020

Abstract: Current point cloud extraction methods based on photogrammetry generate large amounts of spurious detections that hamper useful 3D mesh reconstructions or, even worse, the possibility of adequate measurements. Moreover, noise removal methods for point clouds are complex, slow and incapable to cope with semantic noise. In this work, we present `body2vec`, a model-based body segmentation tool that uses a specifically trained Neural Network architecture. `Body2vec` is capable to perform human body point cloud reconstruction from videos taken on hand-held devices (smartphones or tablets), achieving high quality anthropometric measurements. The main contribution of the proposed workflow is to perform a background removal step, thus avoiding the spurious points generation that is usual in photogrammetric reconstruction. A group of 60 persons were taped with a smartphone, and the corresponding point clouds were obtained automatically with standard photogrammetric methods. We used as a 3D silver standard the clean meshes obtained at the same time with LiDAR sensors post-processed and noise-filtered by expert anthropological biologists. Finally, we used as gold standard anthropometric measurements of the waist and hip of the same people, taken by expert anthropometrists. Applying our method to the raw videos significantly enhanced the quality of the results of the point cloud as compared with the LiDAR-based mesh, and of the anthropometric measurements as compared with the actual hip and waist perimeter measured by the anthropometrists. In both contexts, the resulting quality of `body2vec` is equivalent to the LiDAR reconstruction.

Keywords: deep learning; neural networks; structure from motion; 3D point cloud; anthropometry

1. Introduction

Reconstruction of 3D objects is among the many potential applications of Computer Vision models working together with Deep Learning techniques. This combined approach can contribute to solve common problems regarding the analysis of human body shape. In several contexts, including sports, health, and other contexts, there is a recurrent need to have accurate anthropometric

measurements (i.e., the shape, form, size, and several perimeters and volumes of the human body) [1–3]. This is the case in many clinical applications ranging from diagnostic, treatment, and follow-up of overweight-related conditions, to less frequent but important skeletal pathologies, such as scoliosis [4]. Obesity-related conditions constitute a specifically critical case, since overweight and obesity have become increasingly widespread, considered one of the main public health challenges of the 21st century [5]. Overweight is diagnosed and clinically treated after the assessment of anthropometric traits including weight, height, and several body perimeters [6]. These measurements are typically obtained by traditional manual methods, which are imprecise, require specific professional intervention, and may turn to be incomplete. For instance, in diagnosing obesity, a key indicator is the distribution of abdominal adipose tissue, which is an aspect of geometric shape rather than a relationship among classical anthropometric measures [7].

The acquisition of complete 3D models of human bodies, and its translation into data adequately represented for clinical and non-clinical practices, encloses several difficulties. First, even when the person can be constrained to remain motionless inside a full body scanner device, incomplete surface data is obtained generated by occlusions [8]. This leads to a loss of quality due to missing data in the occluded areas. Second, given their cost and complexity of use, short-range LiDAR scanners are still out of the scope of most physicians, specialists, and research groups. Finally, even though traditional anthropometric measurements are acknowledged to be imprecise in the assessment of overweight-related conditions, they still represent a cheap and easy way to gather information regarding the weight condition of an individual [9–13].

The study of human body shape needs to evolve from the classical somatotype/anthropometric approaches towards inexpensive and practical 3D technologies, and data in digital format. Apart from overweight related issues, clinical applications of 3D body scanning include the precise diagnosis of skeletal pathologies, such as scoliosis, and prosthetic design, among others. Non-clinical applications of 3D body scanning are also diverse. Bodybuilding, fitness, and high competition performance would largely benefit from an accurate and straightforward method to register and visualize body shape and its evolution throughout specific training or activities. Online outfit sales also can be enhanced by replacing the classical size conventions by using individual-specific avatars as the basis to select the proper outfit model and size. In addition, the ability to register data in digital format will enable novel scientific pursuits, for instance to develop population-wise body shape studies (e.g., stratified by ethnic group, age, geographical location, nourishing habits, etc.). Collectively, this would trigger a more refined capture of geometric data, better measurements, and full potential to compile large datasets in the screening of diverse populations. This will in turn enable a rapid, precise, and non-invasive quantification of human body shape [14–16].

Given the diversity of applications and the associated intrinsic complexity, body shape analyses require innovative technological approaches in order to improve the accuracy and precision in the acquisition, processing and analysis of digitized data. Among the first human body acquisition and tracking approaches, we can mention typical feature-engineering-based methods. Mikik et al. [17] for instance, applied body part localization procedures based on template fitting and region growing. These templates are computed from 2D silhouettes, using prior knowledge of average body part shapes (ellipsoids and cylinders) and dimensions using synchronized multiple cameras. Later, more elaborate generative methods enabled the automatic recovery of human shape and pose from images, which leverages learned deformation models using template meshes (from scanners) for graphics applications [18].

More recently, 3D surface scanners provided automated and accurate measurements of body shape adequate for clinically and anthropometric applications. Ng et al. [19] applied more sophisticated techniques based on anatomical landmark positioning, and measurements of idealized circumferences, areas, and volumes, which allow an assessment of the bodily fat distribution and its relationship with metabolic disorders, body mass index, other anthropometric indexes, and their relation to within and among ethnic groups diversity. In addition, in some applications in health and medicine, it is necessary

to segment body parts or volumes (arms, legs, head, torso) from 3D scanned data [20]. For this, fitting techniques are used to deform template models to recognize segment endpoints, determine their locations, and take measurements of them. In others applications, like support tools for fitness, exercise guidance, and wellness activities, it is common to use off-the-shelf 3D devices for body scanning, for example, Microsoft Kinect®. This allows a user to obtain morphometric data to evaluate the body shape and physical conditions of trainees with acceptable precision [21].

Close-range photogrammetry based on structure from motion (SfM) recently emerged as a viable alternative in several applications, ranging from industrial context to cultural heritage preservation, engineering, Earth sciences, and several other 3D imagery tasks. The photogrammetric procedure typically takes successive frames in a video, determines the correspondence between salient points in each frame, and infers the extrinsic camera parameters with which the actual 3D position of these points can be determined. SfM is advantageous for several reasons: it is low-cost, flexible, only requires widespread acquisition devices (for example smartphones or tablets) [22]. This allows quick and easy scanning and point cloud generation. These point clouds, when are computed from adequate video takes, may have only few spurious points, allowing useful characterization and geometric measurements in geographic and urban scales [23]. However, in very close range acquisition, as is the case in 3D body scanning and reconstruction, plain SfM is unable to achieve the precise and high-quality geometry required to obtain accurate measurements. This is mostly due to noise in the background that generate spurious photogrammetric determinations, which in turn deliver wrongly detected points that alter significantly the quality of the point cloud. In this condition, 3D reconstruction and geometric measurements with a noisy point cloud will produce results that are upright unusable. Further decimation and filtering algorithms are inadequate in this context since they distort the underlying body structure, thus leading to inaccurate 3D models. This situation can be mitigated by carefully using a clear background, even illumination, and high-end devices, all conditions that go against the very simplicity and inexpensiveness of SfM in other contexts.

In this work, we present `body2vec`, a model-based approach to background filtering in 3D body scanning. Our method pre-processes the acquisition frames using a convolutional neural network (CNN) that identifies the region of interest (human silhouettes) and filters out the background, thus leading to much cleaner point clouds and more precise subsequent meshes. The final result is both robust with respect to acquisition conditions (background, illumination, and video quality) and accurate enough to produce quality anthropometric measurements. Thus it can be successfully applied to videos taken with low cost devices like smartphones or tablets. Applying our method resulted in an error reduction of almost an order of magnitude (measured as the mean absolute distance to the LiDAR mesh taken as a silver standard). Finally, we estimated the hip and waist perimeters using a very simple fit, comparing the results of the LiDAR-based mesh with `body2vec` against the actual anthropometrists' measurements taken as gold standard, achieving 1.23 cm less mean error in average in the hip measurement and 3.21 cm higher error in the waist measurement.

2. Materials and Methods

In this section, (i) we describe the data collection in detail, including raw video takes from smartphones, LiDAR based, and anthropometric; (ii) we present `BRemNet` (Background Removal Network), a human body identification and segmentation model that is applied for video background removal; (iii) we generate SfM point clouds from the raw and clean videos, generate a registration thereof to the meshed LiDAR acquisition, and measure the respecive registration errors; and (iv) we evaluate anthropometric measurements from the LiDAR mesh, and the raw and clean point clouds, and compare them to the measurements performed by anthropometrists. An overview of the complete process can be seen in Figure 1.

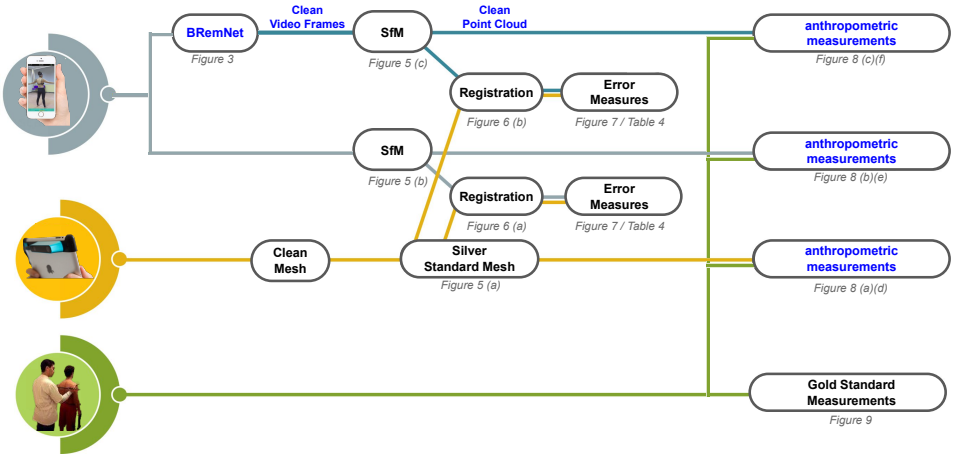


Figure 1. Overview of workflow. Inputs: Raw video in gray, LiDAR-Scanner in yellow, and Classical anthropometry in green.

2.1. Data Collection

Smartphone videos and 3D body scans were taken from 60 volunteers (38 females, 22 males; average age = 39; sd = 12) within the facilities of the Puerto Madryn Regional Hospital. All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the procedure was approved by the Ethics Committee of the Puerto Madryn Regional Hospital under protocol number 19/17 (approved 4 September 2018). Even though higher quality SfM reconstructions could be obtained with high resolution photography, in this context this would be inadequate given the acquisition time required. However, given the rapid evolution of smartphones, it is foreseeable that very high resolution video will be feasible in the near future. It is worth noting that the problem addressed in this paper is not concerned with low resolution acquisition, but of low quality SfM reconstruction due to semantic noise, which is not related to resolution.

Videos were recorded in a single take, completely surrounding the volunteer while they stood in underwear or tight clothes with their arms extended and legs shoulder-width apart (see a typical frame in Figure 2a). The takes were about 35 s long, in MPEG-4, 1920×1080 @ 30 fps. At the same time, a 3D body scan was obtained using the first version of the Structure^{TM}sensor scanner (more detailed description of this data collection are published in Reference [24]). This latter acquisition generates a high-quality point cloud and subsequent 3D mesh, which will be taken as the reference for our video-based 3D reconstruction. Structure^{TM}sensor scanner was our best choice to achieve LiDAR quality with a handheld device able to perform quick captures, with harmless sensing technology, and an affordable price. Finally, anthropometric measurements were acquired by trained domain experts using the standard protocol, including total height (using the Seca 206 mechanical measuring tape, Seca GmBH & Co Kg, Hamburg, Germany), total weight, and body composition (muscle mass, fat-free, and body fat mass and percentages) estimated with a bioimpedance scale (Tanita BC 1100F), and hip and waist circumferences, using ergonomic measuring tape Seca 201 (Seca GmBH & Co Kg, Hamburg, Germany).

2.2. Segmentation Model

As already stated, close-range photogrammetry based on SfM will likely generate low quality point clouds which will require extensive subsequent filtering that will hamper the geometric accuracy of the acquisition. Our strategy is to perform a background removal in all the video frames prior to applying SfM, taking into advantage the fact that the foreground is always a human figure, which implies that a specific semantic segmentation model can be developed using machine learning.

We used the *Mask R-CNN* architecture as a baseline for identification and segmentation of human bodies. This method attempts to identify the pixel-level regions of each body instance in an image. In contrast to semantic segmentation, instance segmentation not only distinguishes semantics, but also different body instances. The model was trained to learn a pixel-level mask for a single class person. Below, we describe the architecture and functioning of the underlying model.

Mask R-CNN is a fully convolutional network (FCN) designed to help locate objects at pixel level and for semantic segmentation [25]. The underlying model is optimized over prior proposals for a multi-task loss function that combines the losses of classification, bounding box localization and segmentation mask $\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}}$. $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{box}}$. The loss function encourages the network to map each pixel to a point in feature space in a way such that pixels belonging to the same instance lie close together, while different instances are separated by a wide margin Reference [26]. $\mathcal{L}_{\text{mask}}$ is defined as the average binary cross-entropy loss, only including the k-th mask if the region is associated with the ground truth class k, where Y_{ij} is the label of a cell (i, j) in the true mask for the region; $\hat{y}_{ij}^{k}$ is the predicted value of the same cell in the mask learned for the ground truth class k (see Equation (1)).

$$\mathcal{L}_{\text{mask}} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} \left[y_{ij} \log \hat{y}_{ij}^{k} + (1 - y_{ij}) \log(1 - \hat{y}_{ij}^{k}) \right]. \tag{1}$$

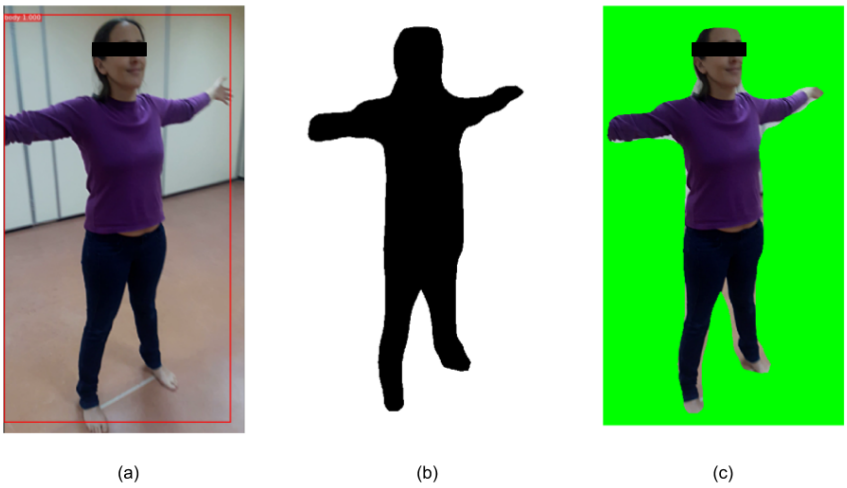


Figure 2. Intermediate results of `BRemNet`. (**a**) Bounding box, (**b**) Mask, and (**c**) Chroma.

We developed `BRemNet`, a further refinement of Mask R-CNN with the aim of pre-processing per frame the videos taken specifically for photogrammetric 3D body reconstruction (see Figure 3). As with Mask R-CNN, we use an RPN, but we add a binary classifier, and a background removal and chroma coding step. The RPN generates a set of bounding boxes which may contain the human body within the video frame. These boxes are refined using the Mask R-CNN regression model (see Figure 2a). The binary classifier was trained to label pixels as foreground/background using the pre-trained weights of the Microsoft Common Objects in COntext (MS COCO) [27] containing the labeled person. We prepared a training dataset with 200 frames with different bodies in different frame locations. These frames were manually annotated using the VGG Image Annotator [28]. The result of this step is then a binary mask containing the silhouette of the human body present in the frame (see Figure 2b). The mask is used in the final background removal and chroma coding step (see Figure 2c). After this processing, the video takes are converted to a set of about 500 frames in which the foreground (the human body) remained unchanged and the background was set into green, which reduces the error introduced in the subsequent SfM step.

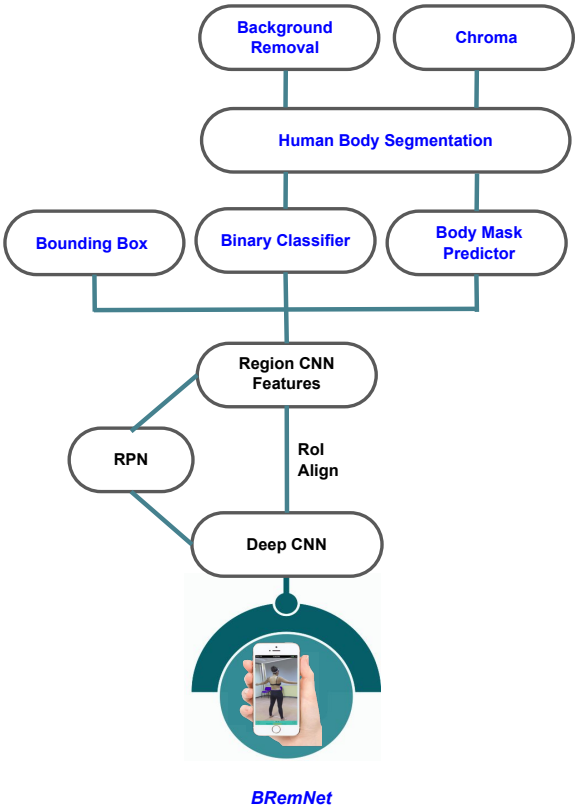


Figure 3. `BRemNet` architecture.

2.3. 3D Reconstruction and Measurement

Structure from Motion photogrammetry provides point clouds computed from multi-views or video takes. These point clouds in some cases can be close in quality to those generated by LiDAR sensors. For this reason, their use is steadily gaining popularity both in long range and in close range applications. Recently, several open-source libraries and applications were released to process SfM in different contexts. In particular, Visual SFM [29] implements the specific case of frame sequence matching, which is adequate for a frame sequence from a video. This procedure is performed both with the original video takes, and with the same takes pre-processed by the segmentation model described in the preceding subsection.

Our long term goal is to extract highly accurate and precise anthropometric measurements from the point clouds generated from videos taken with smartphones or similar devices. In this work, we focus on the abdominal perimeter, which is one of the most representative values related to overweight and similar conditions that currently require frequent and precise assessment in large populations given the current obesity epidemics. For this, we expect to determine the accuracy and precision of the SfM-based assessment, as compared with the LiDAR-based and with the direct measurements performed by trained anthropometrists. The procedure to this avail is to select the points that correspond to the navel height in the subject (the place where the anthropometrists take the actual abdominal perimeter), fit these points to an ellipse, in which the perimeter is the final estimation produced by the model. This navel-height point selection, ellipse fitting, and perimeter measurement is performed on the three point clouds available for the same individual (LiDAR-based, unprocessed video take, processed video take).

3. Results

Below, we evaluate the 3D body reconstruction quality results achieved using BRemNet for masking the human silhouette in the video takes. First, we assess the quality of the segmentation mask with respect to manually segmented masks. Then, we evaluate the enhancements with respect to the point clouds generated using SfM from the raw videos. Finally, we compare the accuracy and precision of the point clouds obtained after masking with respect to LiDAR-based when proposed model is employed, using the anthropometrist's measurements as a gold standard.

3.1. Mask Segmentation

Four quality metrics were applied to evaluate pixel-wise segmentation using BRemNet: Hamming Loss metric, Jaccard index, F1-measure, and Accuracy, against manually segmented masks on 20 frames randomly chosen. We considered the RoI determined by the minimax rectangle of the manually segmented mask in each case. Pixel positive condition then arises when pixels belong to the manual mask, and thus true and false positives, and true and false negatives are defined accordingly. In addition, we compared these results of BRemNet model with the segmentations generated by Mask R-CNN (see Table 1).

Table 1. Mask segmentation metrics.

Measure	Mean		Standard Deviation		Min		Max	
	BRemNet	Mask R-CNN	BRemNet	Mask R-CNN	BRemNet	Mask R-CNN	BRemNet	Mask R-CNN
Hamming loss	0.04149	0.04734	0.00559	0.00693	0.03428	0.03889	0.05578	0.06111
Jaccard	0.86457	0.84577	0.01913	0.02830	0.83373	0.80468	0.90798	0.89994
F-measure	0.92726	0.91620	0.01096	0.01655	0.90933	0.89177	0.95177	0.94733
Accuracy	0.95851	0.95266	0.00559	0.00693	0.94422	0.93889	0.96572	0.96111
FPR	0.02929	0.03226	0.04308	0.04979	0.03844	0.03938	0.03918	0.04252
FNR	0.07050	0.08385	0.09068	0.11291	0.05415	0.06017	0.06929	0.08992

Even though BRemNet performed better than Mask R-CNN in all the quality metrics, the improvement is only marginal. However, the true advantage of BRemNet is related to the consistency of the resulting mask during all the frames of a given take. The segmentation quality of Mask R-CNN is strongly dependent on the cleanness of the background and the stability of the take. For instance, some objects in the background together with part of the actual subject can be misleadingly identified as another object (e.g., a dog), resulting in a frame with a higher amount of false negatives.

Similar situations arise when the subject moves during a take, or if the camera movement along the subject is uneven. We explored the per-frame differences among BRemNet and Mask R-CNN in ten randomly chosen videos, using Jaccard index. The proportion of frames with a similarity less than 0.8 ranges from 3.23% to 38.57%, and the Jaccard index of the least similar frames can drop down to 0.37 (see Table 2 and Figure 4). Although not significant in the average, these Mask R-CNN bad frames result in a poor SfM point cloud reconstruction, since parts of the photogrammetric information will be wrongly inferred. In these takes, we selected the frames with larger disparity between Mask R-CNN and BRemNet (in this case the minimum Jaccard index among both segmentations). In these ten frames (one for each video), we manually segmented the expected mask and established the quality measures of Mask R-CNN and BRemNet (see Table 3). This analysis confirms that Mask R-CNN is prone to worst cases that may hamper the resulting point cloud, while BRemNet performs in a much more stable manner.

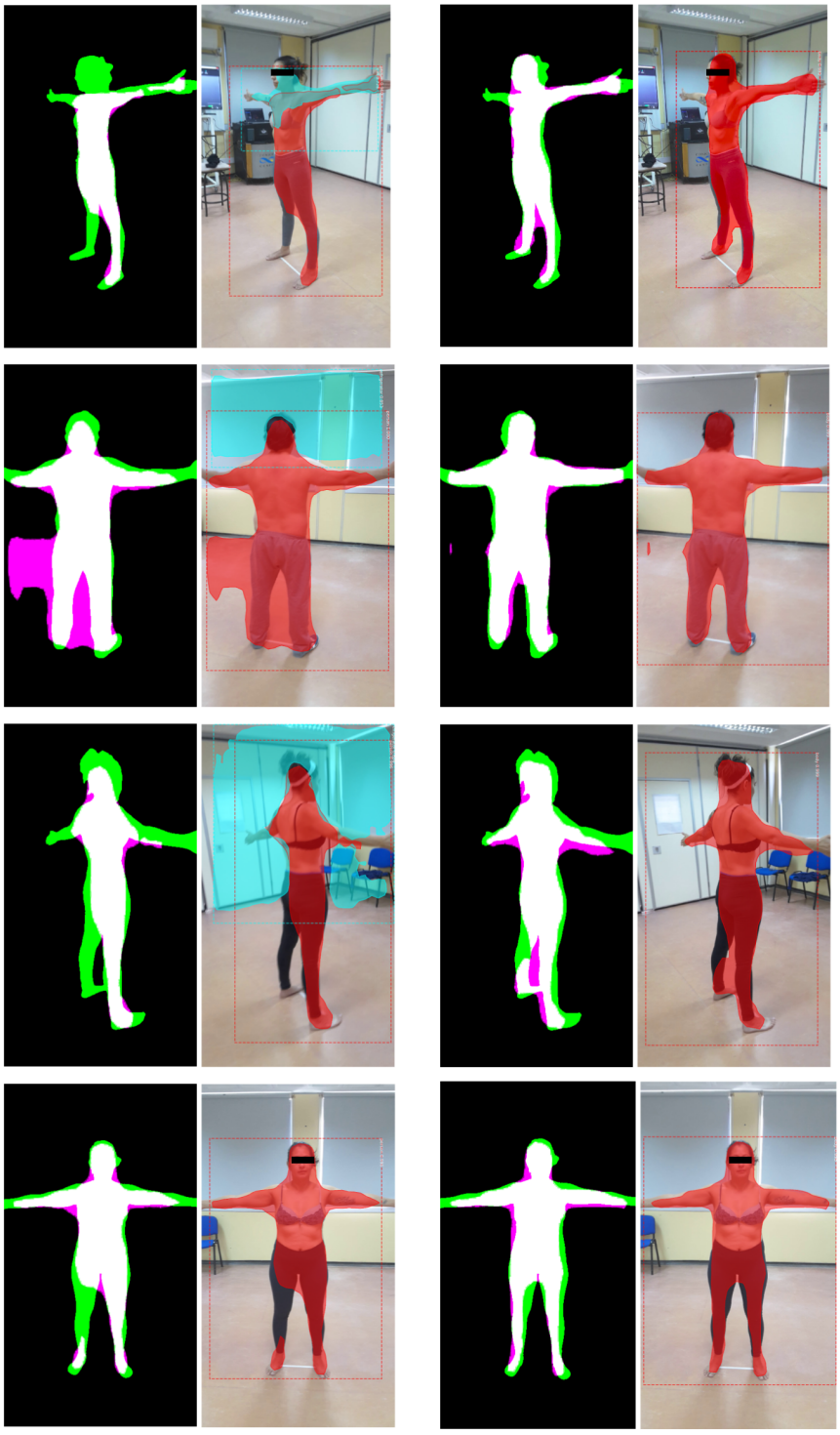


Figure 4. Segmentation examples: Mask R-CNN (left) and `BRemNet` (right). The segmented mask is superimposed in red to the actual frames. In Mask R-CNN, masks of other identified objects are superimposed in cyan. In masks, true positives are in white, false positives are in magenta, and false negatives are in green.

Table 2. Mask R-CNN vs. `BRemNet` video segmentation test (n = 10).

	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6	Video 7	Video 8	Video 9	Video 10
min Jaccard	0.6000	0.6606	0.7035	0.5057	0.5942	0.5967	0.3723	0.6195	0.4757	0.4541
Jaccard < 0.8	25.57%	11.11%	3.23%	29.73%	12.24%	30.43%	21.43%	21.16%	34.56%	38.57%
max FN	110,746	67,498	77,635	96,063	254,296	59,932	152,988	59,032	87,076	84,309

Table 3. Mask segmentation metrics in the frames with min Jaccard.

Measure	Mean		Standard Deviation		Min		Max	
	`BRemNet`	Mask R-CNN	`BRemNet`	Mask R-CNN	`BRemNet`	Mask R-CNN	`BRemNet`	Mask R-CNN
Jaccard	0.73543	0.28999	0.11578	0.13299	0.51644	0.06965	0.84800	0.45189
F-measure	0.84264	0.43349	0.08188	0.17417	0.68112	0.13023	0.91775	0.62249
FPR	0.15062	0.50837	0.17849	0.50000	0.10318	0.56022	0.13559	0.51065
FNR	0.15077	0.63428	0.17222	0.74144	0.09727	0.74409	0.13376	0.73075

3.2. Segmented Point Cloud Evaluation

As mentioned in Section 2.3, the frames segmented by `BRemNet` were used to compute clean point clouds using SfM. On average, these point clouds are composed of 37.265 points each. This represents almost an order of magnitude less than the size of the resulting point clouds computed with the raw video (i.e., without prior use of `BRemNet`), which in average were 229.659 points in size. On the other hand, LiDAR-based mesh models were previously processed with Laplacian smoothing and hole closing using the algorithms presented in Reference [30]. We thus obtained three full-body 3D models, the LiDAR-based mesh, and the raw and clean SfM-based point clouds (see Figure 5). Assuming the LiDAR-based as a silver standard, we used CloudCompare [31] to compare the raw and clean point clouds to the mesh. We measured Root Mean Square Error (RMSE), Mean Distance, and Standard Deviation using Iterative Closest Point (ICP) as registration method, after the proper alignment between the mesh and the point cloud using PCA (see Figure 6). As a registration method, ICP implements nearest neighbour reconstruction using Euclidean distance to estimate the closest point between two entities. Since it is an iterative process, the registration error slowly decreases during this procedure.

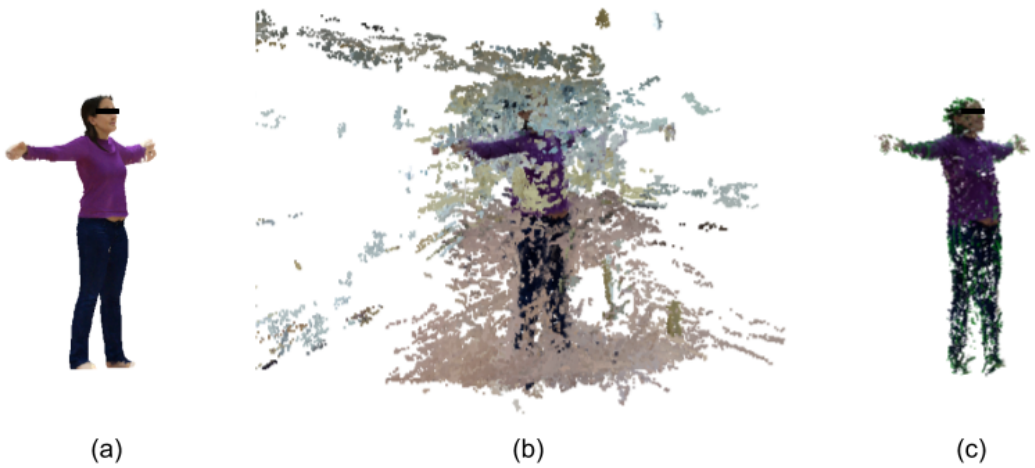


Figure 5. The three 3D full-body reconstructions: (**a**) LiDAR-based mesh cropped, (**b**) structure from motion (SfM)-based point cloud from raw video, and (**c**) point cloud from clean `BRemNet`-filtered video.

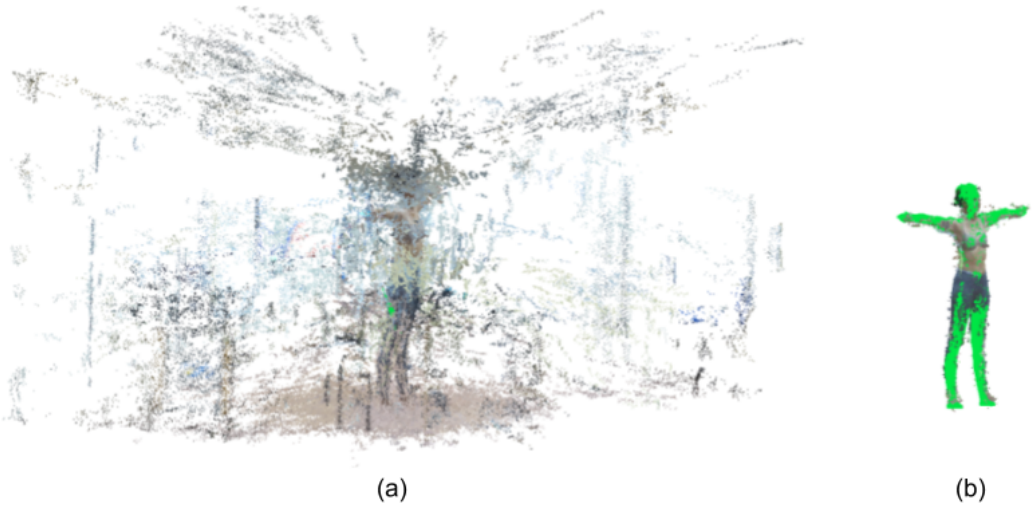


Figure 6. Automatic registration between (**a**) SfM-based point cloud from raw video and (**b**) `BRemNet`-segmented point cloud with LiDAR-based mesh (green).

We computed the Root Mean Square Error (RMSE) between the mesh and the two point clouds in the 60 3D models. The mean RMSE of the raw point cloud was 12.11 cm, while error was reduced in the clean cloud to 2.02 cm. We also computed the Mean Distance (MD), i.e., the mean of distances between each point in the cloud to the nearest triangle in the mesh. In the raw point cloud the MD was 6.28 cm, while, in the clean cloud, the MD dropped to 0.04 cm. Finally, we computed the Standard Deviation (SD) of the MD, which in the raw point cloud was 10.4 cm and in the clean one was 1.9 cm. All these results are shown for the 60 full-body reconstructions in Figure 7.

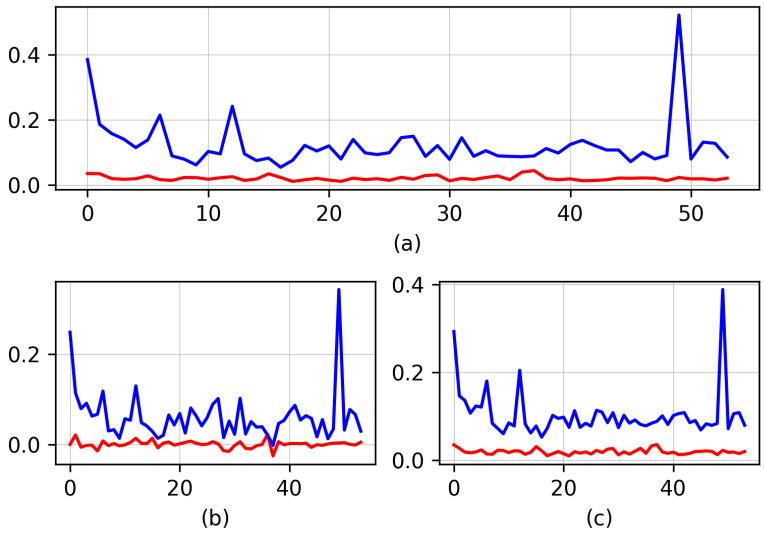


Figure 7. Comparison between the raw point cloud (blue) and the clean one (red) against the LiDAR-based mesh. (**a**) Root Mean Square Error (RMSE), (**b**) Mean Distance, and (**c**) Standard Deviation, all measurements in cm. The x axis represents the volunteer number.

3.3. Abdominal Perimeter Measurements

With the three resulting full-body reconstructions, we performed two anthtropometric measurements, namely hip and waist perimeters. Along with body-mass index, hip and waist perimeters constitute the most widely anthropometric measurements used to detect and analyze overweight conditions. These results were then compared to the actual measurements taken by anthropometrists with the traditional instruments. First, the 3D model is scaled to the measured anthropometric height of the dataset and the centroid is obtained. Around the mean height (on the Y axis), an RoI is considered that takes slices of 1 cm height and orthogonal to the Y axis (i.e., parallel to the floor). Points within each of the slices are fitted to ellipses [32], of which we calculate the perimeter using the Ramanujan approximation. According to the anthropometrists' practice, the hip is the region with the largest perimeter slightly below the middle height of the subject. With our procedure, in the three full-body reconstructions we searched the slice in which this condition arise, and use the perimeter of the fitted ellipse as a predictor of the hip perimeter. In the waist, however, there is no single consensus as to where the actual perimeter should be measured in the subjects, which is usually taken slightly below the navel. Lacking this specific shape trait, we adopted a criterion which is to search the slice above the middle height of the subject whose fit was the closest to the actual anthropometrist's measurement (see Figure 8).

We evaluated the error of these two measurements for the three full-body reconstructions against the actual measurements performed by anthropometrists. The absolute error was calculated from each waist and hip measurement respect to measurement taken by anthropometrists. In Table 4, we show the mean and SD of the two measurements in the 60 subjects. Finally, we performed a least squares regression between the actual anthropometrists' measurements and the estimated measurements performed on the clean point cloud computed with the `BremNet`-filtered videos (see Figure 9).

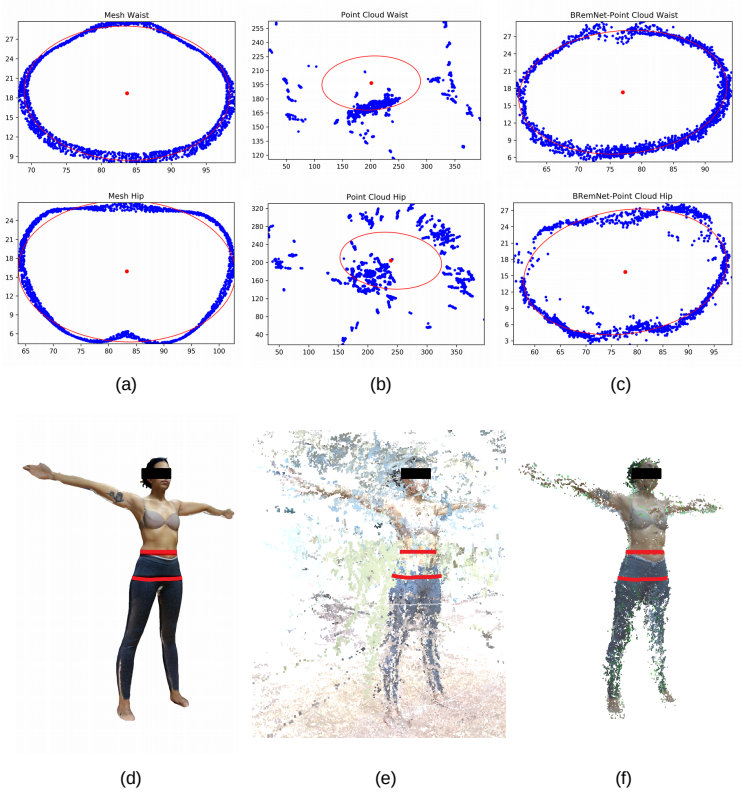


Figure 8. Waist and hip approximation. (**a**,**d**) LiDAR-based mesh, (**b**,**e**) Unsegmented point cloud (scale 1:6), (**c**,**f**) `BRemNet`-segmented point cloud.

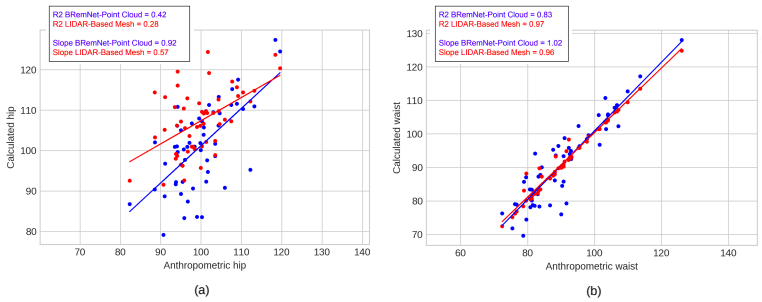


Figure 9. Linear regression of the estimated waist and hip against the actual measurements. `BRemNet`-Point cloud in blue and LiDAR-based mesh in red. (**a**) Waist and (**b**) hip.

Table 4. Waist and hip measurement error.

	Mean Error (cm)		**Standard Deviation (cm)**	
	Hip	**Waist**	**Hip**	**Waist**
LiDAR-based meshes	7.935	**0.910**	6.864	**1.808**
Unsegmented point clouds.	271.708	302.718	87.375	123.548
`BRemNet`-segmented point clouds	**6.701**	4.128	**4.419**	3.148

4. Discussion and Conclusions

We performed geometric reconstructions of the waist and hip of 60 subjects, for which we have also the actual anthropometric ground truth, for both the LiDAR-based and the masked SfM-based point cloud. Our final goal was to assess the applicability of 3D body reconstruction based on handheld video acquisition for anthropometric purposes. The proposed approach increases significantly the quality and robustness of the SfM-point cloud. Masked point clouds have 83.31% less RMSE, 99.32% less mean distance, and 81.72% less standard deviation as compared with the LiDAR-based silver standard. Regressions of both anthropometric measures using both technologies over the 60 subjects show an equivalently acceptable quality. Hip reconstruction is less accurate, since both LiDAR and SfM acquire a tight anatomic surface of the subjects, while the anthropometrists' tape evens out the buttocks separation. This aspect is currently being considered in our reconstruction model.

These preliminary results are quite promising, since there is still room for optimization. In particular, the `BRemNet` model used transfer learning from a general purpose machine vision net, but it can be retrained with a larger set of manually segmented masks to have a higher accuracy than the one shown in Figure 4. In addition, as mentioned above, for hip and waist perimeters (and other anthropometric measures) a better anatomical model can lead to better estimations than the one used in this paper. Finally, with a larger sample set, a more stable and refined regression model can be established to yield final estimations that takes into account other aspects of the subjects' information apart from the acquisition (e.g., somatotype, gender, ethnicity, etc.).

This contribution is related to achieve useful *measurements* only, not to produce 3D reconstruction of good quality, which are clearly better performed with LiDAR acquisition. However, these measurements can be used to feed quite realistic and personalized avatars (e.g., using the computational bodybuilding model), triggering a significant set of possible applications in human health, online outfit retail, and sports, to mention just a few. In particular, as already presented elsewhere [33], it will be possible to verify that the morphometry of the human body is a robust predictor of biomedical phenotypes related to conditions, such as obesity and overweight, as well as validating its usefulness in routine clinical practice.

Author Contributions: M.A.T.-J. and C.D. conceived the original idea. M.A.T.-J., P.N., B.P. and L.M. made the collection of videos and 3D images. V.R., C.P., S.D.A., A.R. and O.P. took the anthropometric data. M.A.T.-J. developed the deep learning network architecture and conducted the experiments. M.A.T.-J. and C.D. analyzed the results. M.A.T.-J. C.D. and R.G.-J. wrote the manuscript. All authors revised and agreed to submit the current version of the manuscript.

Funding: This research was funded by CONICET Grant PIP 2015-2017 ID 11220150100878CO CONICET D.111/16.

Acknowledgments: Authors would like to thank the Centro Nacional Patagónico CCT-CENPAT, Primary Health Care Centers CAPS Favaloro and CAPS Fontana of the Puerto Madryn Regional Hospital for allowing us to use this places, and for its hospitality during the data collection phase. Besides, would like to thank the volunteer participants who took part in this research.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LiDAR	Light Detection and Ranging o Laser Imaging Detection and Ranging
SfM	Structure from motion

References

1. Santos, D.A.; Dawson, J.A.; Matias, C.N.; Rocha, P.M.; Minderico, C.S.; Allison, D.B.; Sardinha, L.B.; Silva, A.M. Reference values for body composition and anthropometric measurements in athletes. *PLoS ONE* **2014**, *9*, e97846. [CrossRef] [PubMed]
2. Maessen, M.F.; Eijsvogels, T.M.; Verheggen, R.J.; Hopman, M.T.; Verbeek, A.L.; de Vegt, F. Entering a new era of body indices: The feasibility of a body shape index and body roundness index to identify cardiovascular health status. *PLoS ONE* **2014**, *9*, e107212. [CrossRef] [PubMed]
3. Zakaria, N.; Gupta, D. *Anthropometry, Apparel Sizing and Design*; Woodhead Publishing: Sawston, UK, 2019.
4. Schmitz, A.; Gäbel, H.; Weiss, H.; Schmitt, O. Anthropometric 3D-body scanning in idiopathic scoliosis. *Zeitschrift fur Orthopadie und ihre Grenzgebiete* **2002**, *140*, 632. [CrossRef] [PubMed]
5. World Health Organization. *Obesity: Preventing and Managing the Global Epidemic*; World Health Organization: Geneva, Switzerland, 2000.
6. Ruderman, A.; Pérez, L.O.; Adhikari, K.; Navarro, P.; Ramallo, V.; Gallo, C.; Poletti, G.; Bedoya, G.; Bortolini, M.C.; Acuña-Alonzo, V.; et al. Obesity, genomic ancestry, and socioeconomic variables in Latin American mestizos. *Am. J. Hum. Biol.* **2019**, *31*, e23278. [CrossRef]
7. Navarro, P.; Ramallo, V.; Cintas, C.; Ruderman, A.; de Azevedo, S.; Paschetta, C.; Pérez, O.; Pazos, B.; Delrieux, C.; González-José, R. Body shape: Implications in the study of obesity and related traits. *Am. J. Human Biol.* **2020**, *32*, e23323. [CrossRef]
8. Anguelov, D.; Srinivasan, P.; Koller, D.; Thrun, S.; Rodgers, J.; Davis, J. SCAPE: Shape Completion and Animation of People. *ACM Trans. Graph.* **2005**, *24*, 408–416. [CrossRef]
9. Ulijaszek, S.J.; Kerr, D.A. Anthropometric measurement error and the assessment of nutritional status. *Br. J. Nutr.* **1999**, *82*, 165–177. [CrossRef]
10. Gordon, C.C.; Bradtmiller, B. Interobserver error in a large scale anthropometric survey. *Am. J. Hum. Biol.* **1992**, *4*, 253–263. [CrossRef]
11. Perini, T.A.; de Oliveira, G.L.; Ornellas, J.d.S.; de Oliveira, F.P. Technical error of measurement in anthropometry. *Rev. Bras. Med. Esporte* **2005**, *11*, 81–85. [CrossRef]
12. Grellety, E.; Golden, M.H. The effect of random error on diagnostic accuracy illustrated with the anthropometric diagnosis of malnutrition. *PLoS ONE* **2016**, *11*, e0168585. [CrossRef]
13. Krishan, K.; Kanchan, T. Measurement error in anthropometric studies and its significance in forensic casework. *Ann. Med. Health Sci. Res.* **2016**, *6*, 62. [CrossRef]
14. Daniell, N.; Olds, T.; Tomkinson, G. Volumetric differences in body shape among adults with differing body mass index values: An analysis using three-dimensional body scans. *Am. J. Hum. Biol.* **2014**, *26*, 156–163. [CrossRef] [PubMed]
15. Jaeschke, L.; Steinbrecher, A.; Pischon, T. Measurement of waist and hip circumference with a body surface scanner: Feasibility, validity, reliability, and correlations with markers of the metabolic syndrome. *PLoS ONE* **2015**, *10*, e0119430. [CrossRef] [PubMed]
16. Medina-Inojosa, J.; Somers, V.K.; Ngwa, T.; Hinshaw, L.; Lopez-Jimenez, F. Reliability of a 3D body scanner for anthropometric measurements of central obesity. *Obes. Open Access* **2016**, *2*. [CrossRef]
17. Mikić, I.; Trivedi, M.; Hunter, E.; Cosman, P. Human body model acquisition and tracking using voxel data. *Int. J. Comput. Vis.* **2003**, *53*, 199–223. [CrossRef]

18. Balan, A.O.; Sigal, L.; Black, M.J.; Davis, J.E.; Haussecker, H.W. Detailed human shape and pose from images. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
19. Ng, B.; Hinton, B.; Fan, B.; Kanaya, A.; Shepherd, J. Clinical anthropometrics and body composition from 3D whole-body surface scans. *Eur. J. Clin. Nutr.* **2016**, *70*, 1265–1270. [CrossRef]
20. Chiu, C.Y.; Pease, D.L.; Fawkner, S.; Sanders, R.H. Automated body volume acquisitions from 3D structured-light scanning. *Comput. Biol. Med.* **2018**, *101*, 112–119. [CrossRef]
21. Saba, M.; Sorrentino, F.; Muntoni, A.; Casti, S.; Cherchi, G.; Carcangiu, A.; Corda, F.; Murru, A.; Spano, L.D.; Scateni, R.; et al. A Seamless Pipeline for the Acquisition of the Body Shape: The Virtuoso Case Study. In Proceedings of the Eurographics Italian Chapter Conference, Catania, Italy, 11–12 September 2017; pp. 71–80.
22. Westoby, M.J.; Brasington, J.; Glasser, N.F.; Hambrey, M.J.; Reynolds, J.M. 'Structure-from-Motion' photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology* **2012**, *179*, 300–314. [CrossRef]
23. Genchi, S.A.; Vitale, A.J.; Perillo, G.M.; Delrieux, C.A. Structure-from-motion approach for characterization of bioerosion patterns using UAV imagery. *Sensors* **2015**, *15*, 3593–3609. [CrossRef]
24. Paschetta, C.; Ramallo, V.; Teodoroff, T.; Navarro, P.; Pazos, B.; Trujillo Jiménez, M.A.; Morales, L.; Pérez, O.; De Azevedo, S.; González-José, R. *RAICES: Una Experiencia de Muestreo Patagónico*, 1st ed.; Libro de Resúmenes de las Decimocuartas Jornadas Nacionales de Antropología Biológica: Buenos Aires, Argentine, 2019; Volume 1. (In Spanish)
25. Wei, X.S.; Xie, C.W.; Wu, J.; Shen, C. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Patter Recognit.* **2018**, *76*, 704–714. [CrossRef]
26. De Brabandere, B.; Neven, D.; Van Gool, L. Semantic instance segmentation with a discriminative loss function. *arXiv* **2017**, arXiv:1708.02551.
27. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zürich, Switzerland, 6–12 September 2014; pp. 740–755.
28. Dutta, A.; Gupta, A.; Zissermann, A. VGG Image Annotator (VIA). 2016. Available online: http://www.robots.ox.ac.uk/vgg/software/via (accessed on 1 September 2020).
29. Wu, C. VisualSFM: A Visual Structure from Motion System. 2011. Available online: http://www.cs.washington.edu/homes/ccwu/vsfm (accessed on 1 September 2020).
30. Cignoni, P.; Callieri, M.; Corsini, M.; Dellepiane, M.; Ganovelli, F.; Ranzuglia, G. Meshlab: An open-source mesh processing tool. In Proceedings of the Eurographics Italian Chapter Conference, Salerno, Italy, 2–4 July 2008; Volume 2008, pp. 129–136.
31. Girardeau-Montaut, D. CloudCompare. 2016. Available online: http://pcp2019.ifp.uni-stuttgart.de/presentations/04-CloudCompare_PCP_2019_public.pdf (accessed on 1 September 2020).
32. Halır, R.; Flusser, J. Numerically stable direct least squares fitting of ellipses. In Proceedings of the 6th International Conference in Central Europe on Computer Graphics and Visualization (WSCG), Citeseer, Bory, Czech Republic, 9–13 February 1998; Volume 98, pp. 125–132.
33. Utkualp, N.; Ercan, I. Anthropometric measurements usage in medical sciences. *BioMed Res. Int.* **2015**, *2015*, 404261. [CrossRef]

Anexo B

Ética

Dirección de Salud

Puerto Madryn, 2 de agosto de 2021

Nota N°010 /2021

Dr. Damián Leonardo Taire
Hospital Zonal de Puerto Madryn
"Dr. Andrés Ísola"
S/D _____

Ref.: Remite Informe Bioético

De nuestra mayor consideración:

Me dirijo a Ud. en nombre del Subcomité de Ética de la Investigación del Comité de Bioética del Área Programática Norte, con el objeto de hacerle llegar el Informe Bioético sobre el protocolo titulado *"Detección de un patrón espacial para la incidencia de infecciones respiratorias asociado con indicadores socioeconómicos y ambientales en el Área Programática Norte"*, que oportunamente nos hiciera llegar para su aprobación.

Consideramos especialmente relevante, de interés científico y social para la salud de nuestra comunidad, el proyecto que se propone desarrollar en su calidad de investigador responsable, acompañado por un equipo que incluye investigadores del HZPM y del IPCSH/CeNPat/CONICET. Sin que ello signifique contradictorio con lo expresado precedentemente, a su pedido de evaluación hemos considerado oportuno incluir, en el Informe Bioético que se adjunta, algunas observaciones y sugerencias que hacen a nuestra tarea en relación con la ética de la investigación, y que esperamos que se incorporen a una segunda versión del Protocolo con el objeto de poder brindar el aval solicitado.

Sin otro particular, saludo a Ud. atentamente,

Lic. María del Carmen Cid
Coordinadora
Comité de Bioética Área Programática Norte

RECIBIDO EL 2 8 2021.
Area Programática Norte
MINISTERIO DE SALUD

INFORME BIOÉTICO

Protocolo de Investigación

Detección de un patrón espacial para la incidencia de infecciones respiratorias asociado con indicadores socioeconómicos y ambientales en el Área Programática Norte

Promotor del Protocolo

Dr. Damián Leonardo Taire (HZPM – Investigador visitante IPCSH/CeNPat/ CONICET

Investigador responsable

Dr. Damián Leonardo Taire (Médico Pediatra especialista en Neumonología Infantil y Medicina Legal HZPM)

Equipo de investigadores

Dra. Sabrina Fernández (Neumonóloga adultos HZPM)

Dra. Laura Aman (Médica Generalista HZPM)

Dra. Silvina Leiva (Médica Pediatra HZPM)

Dra. Virginia Ramallo (Antropóloga Biológica, Investigadora Adjunta CONICET)

Dr. Rolando González José (Antropólogo Biológico, Investigador Principal CONICET)

Dra. Anahí Ruderman (Bióloga, CONICET)

Lic. Bruno Alfredo Pazos (Licenciado en Sistemas, Becario Doctoral CONICET)

Afiliaciones institucionales

Hospital Zonal "Dr. Andrés Ísola", Puerto Madryn, Chubut

Instituto Patagónico de Ciencias Sociales y Humanas (IPCSH)/CeNPat/CONICET

Fecha de presentación del Protocolo al Comité de Bioética

01/07/2021 Versión 1

Introducción

En la fecha indicada precedentemente se recibe la solicitud de aval bioético por parte del Dr. Damián Leonardo Taire, responsable del equipo de investigación, integrado por profesionales del Hospital Zonal de Puerto Madryn y del CeNPat/CONICET para el protocolo cuyo título se consigna al inicio de este Informe.

El Protocolo desarrolla los siguientes componentes: Resumen, Información del Proyecto, Justificación y Objetivos del Estudio, Diseño de la Investigación, Bibliografía.

No incluye Anexos.

Apreciaciones generales

Desde la perspectiva bioética entendemos que el proyecto constituye una interesante iniciativa tendiente a aportar nuevos conocimientos acerca de las características y necesidades de la población, con el objeto de lograr una gestión eficaz de la salud y ajustadas respuestas sanitarias por parte del Sistema.

Los investigadores consideran que una manera de sistematizar ese conocimiento, para que se transforme en herramienta de decisión y anticipación, son los estudios epidemiológicos, relevados a través de series temporales periódicas. Con tal motivo se centran en dos enfermedades del sistema respiratorio, tuberculosis y COVID-19, con el fin de describir la frecuencia de ambas y explorar la distribución espacial y demográfica de los casos, en relación a diversas variables socioeconómicas.

Se trata de un estudio observacional descriptivo correlacional, por cuanto explora la ocurrencia de un fenómeno, analizando su distribución espacial y demográfica en relación a diversos factores de interés (variables socioeconómicas) con la finalidad de conformar una base de datos epidemiológica retrospectiva sobre las enfermedades respiratorias mencionadas.

Las bases de datos que se utilizarán como fuentes para el estudio son las correspondientes a:

a) Sistema Nacional de Vigilancia de la Salud (SNVS).
b) Sistema Integrado de Salud Argentina (SISA)
c) Seguimiento de casos de COVID-19 en el Área Programática Norte por parte del Grupo de Análisis Espacial en Respuesta al COVID-19 (GAER).

d) Aplicación Informática desarrollada por la Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB).
e) Censos Nacionales
f) Encuestas Permanentes de Hogares
g) Padrones electorales y otras fuentes de información similares.

El Protocolo manifiesta encuadrar el estudio en la Ley Nacional N° 25.326 De Protección de Datos Personales, expresando que todas las fases de la investigación amparan la privacidad de los individuos, garantizando de forma total, completa y garantizada la anonimización de los datos.

El Comité de Bioética considera que llevar adelante la investigación proyectada implica un aporte valioso en el campo de la salud a nivel local. Por consiguiente, se expresa positivamente en relación a su validez social, científica y ética.

No obstante lo expresado, y a pedido del equipo de investigación, se consignarán una serie de observaciones a contemplar para poder brindar el aval bioético.

Observaciones bioéticas sobre el Protocolo

Según la Guía para Investigaciones con Seres Humanos (Res. N° 1480/11 MSN), un Protocolo de estudio no requiere la aprobación de un Comité de Ética de la Investigación (CEI) cuando en ella no participan seres humanos o cuando se utiliza información de tipo pública, *siempre que no se identifique a los individuos de ningún modo* (Res. N° 1480/11; Sección A2, P5: Excepciones al requerimiento de revisión por un CEI, incisos a y b). Ello implica que *no existe posibilidad alguna de identificar a los individuos que han sido parte de los datos.*

Por consideraciones similares, se exceptúa la obligación de obtención del Consentimiento Informado en los casos en que se utilizan datos o muestras no vinculables, o información de conocimiento público; es decir, cuando no es posible establecer la identidad de las personas cuyos datos se utilizan, por lo tanto *los investigadores no pueden contactarlas para obtener su consentimiento* (Res. N° 1480/11 MSN; Sección A3, P8: Excepciones, incisos a, b y c)

Atentas a lo mencionado precedentemente, y dado que el Equipo de Investigación ha solicitado el aval bioético para el presente estudio, sugerimos la incorporación al Protocolo de un apartado de Consideraciones Bioéticas que contemple los siguientes aspectos:

1. Declaración expresa de que el protocolo de investigación se ajusta estrictamente a los requerimientos de la Res. N° 1480/11 MSN citadas en párrafos anteriores, y que aun

así, el Equipo de Investigación ha solicitado la revisión del mismo por el Subcomité de Ética de la Investigación del Área Programática Norte, quien ha elaborado el presente Informe.

2. Consistentemente con las normas a las que hemos hecho referencia, sería deseable efectuar algunas aclaraciones respecto de los datos a obtener a partir de la aplicación Bulsarapp de cartografía demográfica a través de apellidos con la que cuenta el GIBEH.

3. De la misma manera, nos parece oportuno que el Protocolo detalle cuáles son las variables socioeconómicas que se prevé vincular a los casos de tuberculosis y COVID-19.

4. Dado el tipo de estudio, y la inexistencia de personas físicas en carácter de participantes del mismo, se explique por qué razones resulta relevante la aclaración de que el CONICET opere como titular de un seguro de daños, tal como aparece mencionado en el Resumen.

5. Se incorpore explícitamente al marco normativo, además de la Ley N° 25.326 De protección de datos personales, la mencionada Guía para Investigaciones con Seres Humanos (Res. N° 1480/11 MSN) y las Pautas Éticas Internacionales para la Investigación relacionada con la Salud con Seres Humanos (Pautas CIOMS).

6. Mención expresa de que toda modificación o enmienda al Protocolo, en caso de haberlas, será comunicada al Comité de Bioética del Área Programática Norte.

7. Atentas a que el estudio manejará datos de población vulnerable, expresar qué medidas tiene previstas el Equipo de Investigación para evitar la estigmatización de dichos colectivos, especialmente en las instancias de comunicación pública de resultados y conclusiones.

8. Destino de los datos una vez finalizado el estudio.

9. Plan de publicación de resultados.

10. Inclusión como Anexo del CV del Responsable del equipo de investigación, Dr. Damián Leandro Taire.

Conclusión

El Comité de Bioética del Área Programática Norte solicita la incorporación de las observaciones sugeridas, a los fines de poder brindar el aval bioético solicitado.

Bibliografía de referencia

ASOCIACIÓN MÉDICA MUNDIAL/AMM (2013): *Declaración de Helsinki. Principios éticos para las investigaciones médicas en seres humanos*. Consultada en: https://www.wma.net/es/policies-post/declaracion-de-helsinki-de-la-amm-principioseticos-para-las-investigaciones-medicas-en-seres-humanos/ ello.pdf

CIOMS (2017): *Pautas éticas internacionales para la investigación relacionada con la salud con seres humanos*. Consejo de Organizaciones Internacionales de las Ciencias Médicas/ Organización Mundial de la Salud. Consultada en: https://cioms.ch/wp-content/uploads/2018/01/CIOMS-EthicalGuideline_SP_WEB.pdf

SALUD INVESTIGA / MSN (2011): *Res. N° 1480/11. Guía para investigaciones en salud humana. Buenos Aires, Ministerio de Salud/Presidencia de la Nación*. Consultada en: https://salud.misiones.gob.ar/wp-content/uploads/2017/07/Guia-inv-Salud-Humana.pdf

UNESCO (2005): *Declaración universal sobre Bioética y Derechos Humanos*. Consultada en: http://portal.unesco.org/es/ev.php-URL_ID=31058&URL_DO=DO_TOPIC&URL_SECTION=201.html

Bonita R, Beaglehole, R y Kjellström T (2008): *Epidemiología básica*. Washington D.C., OPS.

Fathalla, M (2004): *Guía práctica de investigación en salud. Publicación científica y técnica N° 620*. Washington D. C, OPS

Piezas Legales

Ley N° 25.326 De Protección de Datos Personales.

Puerto Madryn, 26 de julio de 2020

Cdra. María de los Ángeles Madariaga
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

Arq. Elena Keller
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

Dra. Rosa Díaz
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

Mg. Claudia Verónica Espinosa
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

Lic. Graciela Perrone
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

Dra. Nancy Edith Settón
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

Lic. María del Carmen Cid
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

Puerto Madryn, 12 de agosto de 2021

Nota N° 011/2021

Sr. Investigador HZPM
IPCSH/CCT/CeNPat/CONICET
Dr. Damián Leonardo Taire
S/D _____

Ref.: Aval Bioético Proyecto de Investigación

Tenemos el agrado de dirigirnos a Ud. con el objeto de informar que como resultado de la lectura y análisis del Protocolo titulado "*Detección de un patrón espacial para la incidencia de infecciones respiratorias asociado con indicadores socioeconómicos y ambientales en el Área Programática Norte*" - *Versión 2, de fecha 10/08/2021,* el Subcomité de Ética de la Investigación del Comité de Bioética del Área Programática Norte acuerda **brindar el AVAL BIOÉTICO solicitado**, por considerar que tanto el Protocolo del Estudio, como la presentación del CV del Promotor del mismo, han dado tratamiento adecuado a las observaciones y pedidos de ampliación de la información, realizados mediante Informe Bioético del 26/07/2021, que le hiciéramos llegar por Nota N° 010/2021 del 02/08/2021.

El Comité de Bioética considera sumamente valiosa la iniciativa promovida por el Equipo de Investigación autor del presente estudio, interpretando que el mismo aportará conocimiento original y de impacto beneficioso para la salud de la comunidad local, transferible a otras experiencias que pudieran ser realizadas a partir de sus resultados y conclusiones. Por consiguiente, el Comité reconoce su validez científica y social, mientras que simultáneamente, y como objeto de su incumbencia, entiende que el mismo respeta los estándares éticos de la investigación que han sido consagrados en foros nacionales e internacionales en las últimas décadas.

Sin otro particular, nos es muy grato saludar atentamente, deseando que puedan desarrollar su tarea con todo éxito.

Cdora. María de los Ángeles Madariaga
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

Arq. Elena Keller
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

RECIBIDO EL 17.8.21
Area Programática Norte
MINISTERIO DE SALUD

Psic. Graciela Perrone
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

Dra. Nancy Edith Settón
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

Dra. Nora Díaz Tumini
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

Mg. Claudia Verónica Espinosa
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

Lic. María del Carmen Cid
Subcomité de Ética de la Investigación
Comité de Bioética
Área Programática Norte

Ministerio de Salud
Provincia del Chubut

HOSPITAL ZONAL PUERTO
MADRYN
Dr. ANDRES ISOLA
ROBERTO GOMEZ 383-TEL 0280
4453030 FAX 0280 4451010

Puerto Madryn, 15 de Septiembre de 2021.-

Al

Dr. Damián Taire

Hospital Andrés Isola

S / D

Ref.: Nota N° 0762/21 – HZPM

De mi mayor consideración:

Por la presente el Comité de Docencia e Investigación del Hospital Andrés Isola, de la ciudad de Puerto Madryn, se expide respecto del Proyecto: **"Detección de un patrón espacial para la incidencia de infecciones respiratorias asociado con indicadores socioeconómicos y ambientales en el Área Programática Norte"** considerando que el mismo se enmarca dentro de un problema de salud real y frecuente, entendiendo que puede aportar información valiosa para la toma de decisiones en nuestra ciudad.

Dada la metodología de análisis planteada, se sugiere que los investigadores médicos, acompañen el proceso de análisis para que los resultados cumplan con la plausabilidad biológica.

Asimismo solicitamos participar, tanto el Comité como el equipo directivo en los avances y resultado obtenidos.-

Finalmente felicitamos a los participantes,

Sin otro particular saludo atte.-

MESA DE ENTRADA
RECIBIDO
22 Sep
HOSPITAL ZONAL DE PTO MADRYN

Dra. Miriam Tufaro
Presidente Comité Interdisciplinario
de Docencia e Investigación
Hospital Dr Andrés Isola

Comité Interdisciplinario de
Docencia e Investigación
Hospital Zonal Andrés R. Isola
Puerto Madryn - Chubut

Bibliografía

George Redmonds, Turi King, and David Hey. *Surnames, DNA, and Family History*. OUP Oxford, 2011.

Rolf Foerster. Acerca de los nombres de las personas (üy) entre los mapuches. otra vuelta de tuerca. *Revista Chilena de Antropología*, (21), 2010.

CR Guglielmino, G Zei, and LL Cavalli-Sforza. Genetic and cultural transmission in sicily as revealed by names and surnames. *Human biology*, pages 607–627, 1991.

Susanna C Manrubia and Damián H Zanette. At the boundary between biological and cultural evolution: The origin of surname distributions. *Journal of Theoretical Biology*, 216(4):461–477, 2002.

Franz Manni, Bruno Toupance, Audrey Sabbagh, and Evelyne Heyer. New method for surname studies of ancient patrilineal population structures, and possible application to improvement of y-chromosome sampling. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 126 (2):214–228, 2005.

Emma Laura Alfaro, María Ester Albeck, and José Edgardo Dipierri. Apellidos en casabindo entre los siglos xvii y xx: Continuidades y cambio. *Andes*, (16):147–165, 2005.

Ranajit Chakraborty, Sara A Barton, Robert E Ferrell, and William J Schull. Ethnicity determination by names among the aymara of chile and bolivia. *Human Biology*, pages 159–177, 1989.

José Edgardo Dipierri, E Alfaro, and Ignacio Bejarano. Surnames, abo system and miscegenation in highlands population of province of jujuy (northwest argentine). *Homo*, 50 (1):14–20, 1999.

Pablo Mateos. A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place*, 13(4):243–263, 2007.

George H Darwin. Marriages between first cousins in england and their effects. *Journal of the Statistical Society of London*, 38(2):153–184, 1875.

Daniela Torres-Hernández, Tatiana Fletcher-Toledo, Roberth Alirio Ortiz-Martínez, and María Amparo Acosta-Aragón. La endogamia como causa de consanguinidad y su asociación con anomalías congénitas. *Medicina & Laboratorio*, 25(1):409–418, 2023.

Alan H Bittles and Michael L Black. Consanguinity, human evolution, and complex diseases. *Proceedings of the National Academy of Sciences*, 107(suppl_1):1779–1786, 2010.

James F Crow and Arthur P Mange. Measurement of inbreeding from the frequency of marriages between persons of the same surname. *Eugenics Quarterly*, 12(4):199–203, 1965.

Norikazu Yasuda and Toshiyuki Furusho. Random and nonrandom inbreeding revealed from isonymy study. i. small cities of japan. *American Journal of Human Genetics*, 23 (3):303, 1971.

Luigi Luca Cavalli-Sforza and Walter Fred Bodmer. *The genetics of human populations*. Courier Corporation, 1999.

Norikazu Yasuda and NE Morton. Studies on human population structure. In *Third International Congress of Human Genetics*, pages 249–265. Johns Hopkins Press Baltimore, 1967.

Nobuyoshi Yasuda, Luigi Luca Cavalli-Sforza, Maurice Skolnick, and Antonio Moroni. The evolution of surnames: an analysis of their distribution and extinction. *Theoretical population biology*, 5(1):123–142, 1974.

Motoo Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

Alberto Piazza, Sabina Rendine, Gianna Zei, Antonio Moroni, and Luigi Luca Cavalli-Sforza. Migration rates of human populations from surname distributions. *Nature*, 329 (6141):714–716, 1987.

John H Relethford. Isonymy and population structure of irish isolates during the 1890s. *Journal of biosocial science*, 14(2):241–247, 1982.

Jean Pierre Dugène and Frederic Bauduer. Marriage patterns in the western french pyrenees during the 1800–1899 period: data from the village of béost (ossau valley, béarn). *Bulletins et Memoires de la Societe d'Anthropologie de Paris*, 25:118–126, 2013.

Gabriel W Lasker. The occurrence of identical (isonymous) surnames in various relationships in pedigrees: a preliminary analysis of the relation of surname combinations to inbreeding. *American Journal of Human Genetics*, 20(3):250, 1968.

Alvaro Rodríguez Larralde and Italo Barrai. Estudio genético demográfico del estado zulia, venezuela, a través de isonimia. *Acta cient. venez*, pages 134–43, 1998.

Luigi L Cavalli-Sforza and Anthony WF Edwards. Phylogenetic analysis. models and estimation procedures. *American journal of human genetics*, 19(3 Pt 1):233, 1967.

Masatoshi Nei. Analysis of gene diversity in subdivided populations. *Proceedings of the national academy of sciences*, 70(12):3321–3323, 1973.

José E Dipierri, EL Alfaro, Chiara Scapoli, Elisabetta Mamolini, A Rodriguez-Larralde, and Italo Barrai. Surnames in argentina: A population study through isonymy. *American Journal of Physical Anthropology*, 128(1):199–209, 2005a.

Ronald A Fisher, A Steven Corbet, and Carrington B Williams. 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*, pages 42–58, 1943.

A Rodríguez-Larralde, G Formica, Chiara Scapoli, María Beretta, Elisabetta Mamolini, and Italo Barrai. Microevolution in perugia: isonymy 1890–1990. *Annals of human biology*, 20(3):261–274, 1993.

A Rodriguez-Larralde. Estimadores de aislamiento en base a distribución de apellidos. *XXXVI Convención Anual de AsoVAC, Valencia, Venezuela*, 1986.

Wendy R Fox and Gabriel W Lasker. The distribution of surname frequencies. *International Statistical Review/Revue Internationale de Statistique*, pages 81–87, 1983.

L Tarskaia, GI El'Chinova, C Scapoli, E Mamolini, A Carrieri, A Rodriguez-Larralde, and I Barrai. Surnames in siberia: a study of the population of yakutia through isonymy. *American Journal of Physical Anthropology*, 138(2):190–198, 2009.

Yan Liu, Liujun Chen, Yida Yuan, and Jiawei Chen. A study of surnames in china through isonymy. *American Journal of Physical Anthropology*, 148(3):341–350, 2012.

Italo Barrai, G Formica, Chiara Scapoli, M Beretta, Elisabetta Mamolini, S Volinia, Roberto Barale, P Ambrosino, and F Fontana. Microevolution in ferrara: isonymy 1890–1990. *Annals of human biology*, 19(4):371–385, 1992.

Augusto César Cardoso-dos Santos, Virginia Ramallo, Marcelo Zagonel-Oliveira, Maurício Roberto Veronez, Pablo Navarro, Isabella L Monlleó, Victor Hugo Valiati, José Edgardo Dipierri, and Lavinia Schuler-Faccini. An invincible memory: what surname analysis tells us about history, health and population medical genetics in the brazilian northeast. *Journal of Biosocial Science*, 53(2):183–198, 2021.

Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2): 17–23, 1950.

Sergio Rey, Dani Arribas-Bel, and Levi John Wolf. *Geographic data science with python*. CRC Press, 2023.

Luc Anselin. Local indicators of spatial association—lisa. *Geographical analysis*, 27(2): 93–115, 1995.

Luc Anselin, Ibnu Syabri, Oleg Smirnov, et al. Visualizing multivariate spatial correlation with dynamically linked windows. In *Proceedings, CSISS Workshop on New Tools for Spatial Data Analysis, Santa Barbara, CA*, 2002.

Samuel Karlin and James McGregor. The number of mutant forms maintained in a population. In *Proceedings of the Fifth Berkeley Symposium on mathematics, Statistics and probability*, volume 4, pages 415–438, 1967.

Gianna Zei, R Guglielmino Matessi, Enzo Siri, Antonio Moroni, and L Cavalli-Sforza. Surnames in sardinia: I. fit of frequency distributions for neutral alleles and genetic population structure. *Annals of Human Genetics*, 47(4):329–352, 1983.

Sewall Wright. Isolation by distance. *Genetics*, 28(2):114, 1943.

C Scapoli, H Goebl, S Sobota, E Mamolini, A Rodriguez-Larralde, and I Barrai. Surnames and dialects in france: Population structure and cultural evolution. *Journal of theoretical biology*, 237(1):75–86, 2005.

Masatoshi Nei and Yoko Imaizumi. Genetic structure of human populations. *Heredity*, 21: 183–190, 1966.

Chiara Scapoli, Elisabetta Mamolini, Alberto Carrieri, Alvaro Rodriguez-Larralde, and Italo Barrai. Surnames in western europe: A comparison of the subcontinental populations through isonymy. *Theoretical Population Biology*, 71(1):37–48, 2007.

Italo Barrai, Alvaro Rodriguez-Larralde, Elisabetta Mamolini, Franz Manni, and Chiara Scapoli. Isonymy structure of usa population. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 114(2):109–123, 2001.

Junyong Meng, Haisheng Chen, Xingxing Liang, and Jungang Yan. The empirical study of the spatial distribution of chinese surnames. In *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 398–403. IEEE, 2016.

Xiaohui Fan, Yan Liu, Yida Yuan, Jiawei Chen, and Liujun Chen. A surname-based index of migration intensity and its application in china. *Physica A: Statistical Mechanics and its Applications*, 626:129034, 2023.

Jiawei Chen, Liujun Chen, Yan Liu, Xiaomeng Li, Yida Yuan, and Yougui Wang. An index of chinese surname distribution and its implications for population dynamics. *American Journal of Physical Anthropology*, 169(4):608–618, 2019.

E Azevedo, NE Morton, C Miki, and Shirley Yee. Distance and kinship in northeastern brazil. *American Journal of Human Genetics*, 21(1):1, 1969.

Alvaro Rodriguez-Larralde, José Dipierri, Emma Alfaro Gomez, Chiara Scapoli, Elisabetta Mamolini, Germano Salvatorelli, Sonia De Lorenzi, Alberto Carrieri, and Italo Barrai. Surnames in bolivia: A study of the population of bolivia through isonymy. *American journal of physical anthropology*, 144(2):177–184, 2011.

José Dipierri, Alvaro Rodriguez-Larralde, Emma Alfaro, Chiara Scapoli, Elisabetta Mamolini, Germano Salvatorelli, Graziano Caramori, Sonia De Lorenzi, Massimo Sandri, Alberto Carrieri, et al. A study of the population of paraguay through isonymy. *Annals of human genetics*, 75(6):678–687, 2011.

I Barrai, A Rodriguez-Larralde, J Dipierri, E Alfaro, N Acevedo, E Mamolini, M Sandri, A Carrieri, and C Scapoli. Surnames in chile: a study of the population of chile through isonymy. *American journal of physical anthropology*, 147(3):380–388, 2012.

Edwin Francisco Herrera Paz, Chiara Scapoli, Elisabetta Mamolini, Massimo Sandri, Alberto Carrieri, Alvaro Rodriguez-Larralde, and Italo Barrai. Surnames in honduras: A study of the population of honduras through isonymy. *Annals of human genetics*, 78 (3):165–177, 2014.

A Carrieri, M Sans, JE Dipierri, E Alfaro, E Mamolini, M Sandri, A Rodríguez-Larralde, C Scapoli, and I Barrai. The structure and migration patterns of the population of uruguay through isonymy. *Journal of Biosocial Science*, 52(2):300–314, 2020.

Alvaro Rodríguez-Larralde, Jorge Morales, and Italo Barrai. Surname frequency and the isonymy structure of venezuela. *American Journal of Human Biology: The Official Journal of the Human Biology Association*, 12(3):352–362, 2000.

JE Dipierri, A Rodríguez-Larralde, EL Alfaro, and I Barrai. Isonymic structure of the argentine northwest. *Annals of Human Biology*, 34(4):498–503, 2007.

José Edgardo Dipierri, Alvaro Rodríguez Larralde, Emma Laura Alfaro, Alberto Andrade, Estela Chaves, and Italo Barrai. Distribución de apellidos y migración en el noroeste argentino. *Antropo*, 10:35–50, 2005b.

Rubén A Bronberg, José E Dipierri, Emma L Alfaro, Italo Barrai, Alvaro Rodríguez-Larralde, Eduardo E Castilla, Vincenza Colonna, Greta Rodríguez-Arroyo, and Graciela Bailliet. Isonymy structure of buenos aires city. *Human biology*, 81(4):447–461, 2009.

Sonia E Colantonio, Gabriel W Lasker, Bernice A Kaplan, and Vicente Fuster. Use of surname models in human population biology: a review of recent developments. *Human Biology*, pages 785–807, 2003.

P.E.N. Codigo Electoral Nacional. Codigo electoral nacional - ley 19945/1972, 1972. URL https://www.argentina.gob.ar/normativa/nacional/ley-19945-19442/actualizacion. Last accessed 1 February 2024.

Ministerio de Justicia. Infoleg - proteccion de los datos personales, 2000. URL https://servicios.infoleg.gob.ar/infolegInternet/anexos/60000-64999/64790/norma.htm. Last accessed 1 February 2024.

Instituto Geográfico Nacional. Capas sig, 2019. URL https://www.ign.gob.ar/NuestrasActividades/InformacionGeoespacial/CapasSIG. Last accessed 1 February 2024.

Ministerio De Modernización. Api del servicio de normalización de datos geográficos de argentina, 2016. URL https://datosgobar.github.io/georef-ar-api/. Last accessed 1 February 2024.

Dirección Nacional de Datos e Información Pública. Guía para la identificación y uso de entidades interoperables, 2016. URL https://datosgobar.github.io/paquete-apertura-datos/guia-interoperables/#guia-para-la-identificacion-y-uso-de-entidades-interoperables. Last accessed 1 February 2024.

Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. .º'Reilly Media, Inc.", 2012.

Leonardo Morales, Pablo Navarro, Celia Cintas, Rolando Gonzalez-Jose, Virginia Ramallo, and Claudio Delrieux. Bulsarapp: Interactive visual analysis for surname trend exploration. *IEEE Computer Graphics and Applications*, 42(4):28–39, 2021.

Leonardo Monasterio. Surnames and ancestry in brazil. *PloS one*, 12(5):e0176890, 2017.

María Ester Albeck, Emma Laura Alfaro, José Edgardo Dipierri, and Estela Raquel Chaves. Los apellidos de salta en el siglo xxi: origen geo-lingüítico, diversidad y frecuencia. *Andes*, 28(2):00–00, 2017.

William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175, page 14. Las Vegas, NV, 1994.

Alan M MacEachren and John H Ganter. A pattern identification approach to cartographic visualization. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 27(2):64–81, 1990.

Ana María Foschiatti. El movimiento natural de la población. *Geográfica digital*, 8(16): 1–40, 2011.

Gustavo Busso. Migración interna y desarrollo territorial en argentina a inicios del siglo xxi. brechas e impactos sociodemográficos de la migración interna interprovincial. In *IX Jornadas Argentinas de Estudios de Población*. Asociación de Estudios de Población de la Argentina, 2007.

Guillermo Ángel Velázquez and Sebastián Gómez Lende. Dinámica migratoria: coyuntura y estructura en la argentina de fines del xx. *Amérique Latine Histoire et Mémoire. Les Cahiers ALHIM. Les Cahiers ALHIM*, (9), 2004.

D Castro de Guerra, A Rodriguez-Larralde, and J Pinto-Cisternas. Distribución de los apellidos y estructura de población en algunas poblaciones de origen negro de la zona costera norcentral de venezuela. *Acta Cient. Venez*, 41:241–249, 1990.

Richard Alford. Naming and identity: A cross-cultural study of personal naming practices. 1987.

Robert Reuven Sokal, RM Harding, Gabriel Ward Lasker, and CGN Mascie-Taylor. A spatial analysis of 100 surnames in england and wales. *Annals of Human Biology*, 19 (5):445–476, 1992.

CGN Mascie-Taylor and Gabriel W Lasker. Geographical distribution of common surnames in england and wales. *Annals of human biology*, 12(5):397–401, 1985.

Bruno Mourrieras, Pierre Darlu, Joëlle Hochez, and Serge Hazout. Surname distribution in france: A distance analysis by a distorted geographical map. *Annals of human biology*, 22(3):183–198, 1995.

Brian D Gushulak and Douglas W MacPherson. The basic principles of migration health: population mobility and gaps in disease prevalence. *Emerging themes in epidemiology*, 3:1–11, 2006.

Tom Strachan and Andrew Read. *Human molecular genetics*. Garland Science, 2018.

Elizabeth Pennisi. The human genome, 2001.

OMIM. An online catalog of human genes and genetic disorders, 1966-2024. URL https://www.omim.org/. Last accessed 1 February 2024.

Mike Fleckenstein, Lorraine Fellows, Mike Fleckenstein, and Lorraine Fellows. Data analytics. *Modern data strategy*, pages 133–142, 2018.

Fabio Nelli. *Python data analytics: Data analysis and science using PANDAs, Matplotlib and the Python Programming Language*. Apress, 2015.

Jiawei Han, Micheline Kamber, and Jian Pei. Data mining concepts and techniques third edition. *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*, 2012.

Gordon S Linoff and Michael JA Berry. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 2011.

Hadley Wickham. Tidy data. *Journal of Statistical Software*, 59(10):1–23, 2014. doi: 10.18637/jss.v059.i10. URL https://www.jstatsoft.org/index.php/jss/article/view/v059i10.

Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. Nia-aa research framework: toward a biological definition of alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, 2018.

Philip Scheltens, Bart De Strooper, Miia Kivipelto, Henne Holstege, Gael Chételat, Charlotte E Teunissen, Jeffrey Cummings, and Wiesje M van der Flier. Alzheimer's disease. *The Lancet*, 397(10284):1577–1590, 2021.

Jesús Andrade-Guerrero, Alberto Santiago-Balmaseda, Paola Jeronimo-Aguilar, Isaac Vargas-Rodríguez, Ana Ruth Cadena-Suárez, Carlos Sánchez-Garibay, Glustein Pozo-Molina, Claudia Fabiola Méndez-Catalá, Maria-del-Carmen Cardenas-Aguayo, Sofía Diaz-Cintra, et al. Alzheimer's disease: An updated overview of its genetics. *International Journal of Molecular Sciences*, 24(4):3754, 2023.

Thomas D Bird, Thomas H Lampe, Ellen J Nemens, Gary W Miner, SM Sumi, and Gerard D Schellenberg. Familial alzheimer's disease in american descendants of the volga germans: probable genetic founder effect. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 23(1):25–31, 1988.

J Otto Pohl. Volk auf dem weg: Transnational migration of the russian-germans from 1763 to the present day. *Studies in Ethnicity and Nationalism*, 9(2):267–286, 2009.

Wagner Raúl A. Alemanes del wolga en argentina, 2007. URL http://www.alemanesdelwolga.com.ar/. Last accessed 1 February 2024.

F Pagés Larraya, Lina Grasso, and Gonzalo Marí. Prevalencia de las demencias del tipo alzheimer, demencias vasculares y otras demencias del dsm-iv y del icd-10 en la república argentina. *Revista Neurológica Argentina*, 29:148–153, 2004.

Carlos M Melcon, Leonardo Bartoloni, Marcelo Katz, Rodrigo Del Mónaco, Carlos A Mangone, Mario O Melcon, and Ricardo F Allegri. Propuesta de un registro centralizado de casos con deterioro cognitivo en argentina (redecar) basado en el sistema nacional de vigilancia epidemiológica. *Neurología Argentina*, 2(3):161–166, 2010.

Patricio Chrem Méndez, Ismael Calandri, Federico Nahas, María Julieta Russo, Ignacio Demey, María Eugenia Martín, María Florencia Clarens, Paula Harris, Fernanda Tapajoz, Jorge Campos, et al. Argentina-alzheimer's disease neuroimaging initiative (arg-adni): neuropsychological evolution profile after one-year follow up. *Arquivos de Neuro-Psiquiatria*, 76:231–240, 2018.

Tatiana Itzcovich, Patricio Chrem-Méndez, Silvia Vázquez, Micaela Barbieri-Kennedy, Matías Niikado, Horacio Martinetto, Ricardo Allegri, Gustavo Sevlever, and Ezequiel I Surace. A novel mutation in psen1 (p. t119i) in an argentine family with early-and late-onset alzheimer's disease. *Neurobiology of Aging*, 85:155–e9, 2020.

Matías Jonás García and Ana Comesaña. Prevalence of neurocognitive disorders in a rural area of argentina. *Revista de la Facultad de Ciencias Medicas (Cordoba, Argentina)*, 78(4):347–352, 2021.

Sergio J Rey and Luc Anselin. Pysal: A python library of spatial analytical methods. In *Handbook of applied spatial analysis: Software tools, methods and applications*, pages 175–193. Springer, 2009.

Thomas D Bird, EM Nemens, D Nochlin, SM Sumi, EM Wijsman, and Gerald D Schellenberg. Familial alzheimer's disease in germans from russia: a model of genetic heterogeneity in alzheimer's disease. In *Heterogeneity of Alzheimer's Disease*, pages 118–129. Springer, 1992.

Jorge J Llibre-Guerra, Yan Li, Ricardo F Allegri, Patricio Chrem Mendez, Ezequiel I Surace, Juan J Llibre-Rodriguez, Ana Luisa Sosa, Carmen Aláez-Verson, Erika-Mariana Longoria, Alberto Tellez, et al. Dominantly inherited alzheimer's disease in latin america: Genetic heterogeneity and clinical phenotypes. *Alzheimer's & Dementia*, 17(4): 653–664, 2021.

Nadia Dehghani, Jose Bras, and Rita Guerreiro. How understudied populations have contributed to our understanding of alzheimer's disease genetics. *Brain*, 144(4):1067–1081, 2021.

H Bickel. Dementia syndrome and alzheimer disease: an assessment of morbidity and annual incidence in germany. *Gesundheitswesen (Bundesverband der Arzte des Offentlichen Gesundheitsdienstes (Germany))*, 62(4):211–218, 2000.

Uta Ziegler and Gabriele Doblhammer. Prävalenz und inzidenz von demenz in deutschland–eine studie auf basis von daten der gesetzlichen krankenversicherungen von 2002. *Das Gesundheitswesen*, 71(05):281–290, 2009.

Geneviève Chêne, Alexa Beiser, Rhoda Au, Sarah R Preis, Philip A Wolf, Carole Dufouil, and Sudha Seshadri. Gender and incidence of dementia in the framingham heart study from mid-adult life. *Alzheimer's & Dementia*, 11(3):310–320, 2015.

Marina Muzzio, Josefina MB Motti, Paula B Paz Sepulveda, Muh-ching Yee, Thomas Cooke, María R Santos, Virginia Ramallo, Emma L Alfaro, Jose E Dipierri, Graciela Bailliet, et al. Population structure in argentina. *PLoS One*, 13(5):e0196325, 2018.

Pierre Luisi, Angelina García, Juan Manuel Berros, Josefina MB Motti, Darío A Demarchi, Emma Alfaro, Eliana Aquilano, Carina Argüelles, Sergio Avena, Graciela Bailliet, et al. Fine-scale genomic analyses of admixed individuals reveal unrecognized genetic ancestry components in argentina. *PloS one*, 15(7):e0233808, 2020.

2020 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 16(3):391–460, 2020. doi: https://doi.org/10.1002/alz.12068. URL https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1002/alz.12068.

Randall J Bateman, Paul S Aisen, Bart De Strooper, Nick C Fox, Cynthia A Lemere, John M Ringman, Stephen Salloway, Reisa A Sperling, Manfred Windisch, and Chengjie Xiong. Autosomal-dominant alzheimer's disease: a review and proposal for the prevention of alzheimer's disease. *Alzheimer's research & therapy*, 3(1):1–13, 2011.

Igor Akushevich, Julia Kravchenko, Arseniy Yashkin, P Murali Doraiswamy, Carl V Hill, Alzheimer's Disease, and Related Dementia Health Disparities Collaborative Group. Expanding the scope of health disparities research in alzheimer's disease and related dementias: Recommendations from the "leveraging existing data and analytic methods for health disparities research related to aging and alzheimer's disease and related dementias" workshop series. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 15(1):e12415, 2023.

Organización Panamericana de la Salud. Envejecimiento y cambios demográficos, 2017. URL https://web.archive.org/web/20210126061933/https://www.paho.org/salud-en-las-americas-2017/?post_type=post_t_es&p=314&lang=es. Last accessed 26 January 2021.

Gustavo Álvarez, Alicia Gómez, and María Fernanda Olmos. Pobreza y comportamiento demográfico en argentina: La heterogeneidad de la privación y sus manifestaciones. *Papeles de población*, 13(51):77–110, 2007.

Francisco Gatto. Crecimiento económico y desigualdades territoriales en argentina. *En: Crisis, recuperación y nuevos dilemas. La economía argentina, 2002-2007-LC/W. 165-2007-p. 307-356*, 2007.

PODER EJECUTIVO NACIONAL (P.E.N.). Ley 19945/1972 - codigo electoral nacional, 2012. URL https://www.argentina.gob.ar/normativa/nacional/ley-19945-19442/actualizacion. Last accessed 1 February 2021.

Guadalupe Soto-Estrada, Laura Moreno-Altamirano, and Daniel Pahua Díaz. Panorama epidemiológico de méxico, principales causas de morbilidad y mortalidad. *Revista de la Facultad de Medicina (México)*, 59(6):8–22, 2016.

Mariana Sanmartino. Pensar problemas complejos desde un enfoque social: transición epidemiológica y enfermedad de chagas. *Contribuciones desde Coatepec*, 15, 2016.

Merill Singer and Scott Clair. Syndemics and public health: Reconceptualizing disease in bio-social context. *Medical anthropology quarterly*, 17(4):423–441, 2003.

Merrill Singer. *Introduction to syndemics: A critical systems approach to public and community health*. John Wiley & Sons, 2009.

Hannah Blencowe, Simon Cousens, Fiorella Bianchi Jassir, Lale Say, Doris Chou, Colin Mathers, Dan Hogan, Suhail Shiekh, Zeshan U Qureshi, Danzhen You, et al. National, regional, and worldwide estimates of stillbirth rates in 2015, with trends from 2000: a systematic analysis. *The Lancet Global Health*, 4(2):e98–e108, 2016.

Simon Cousens, Hannah Blencowe, Cynthia Stanton, Doris Chou, Saifuddin Ahmed, Laura Steinhardt, Andreea A Creanga, Özge Tunçalp, Zohra Patel Balsara, Shivam Gupta, et al. National, regional, and worldwide estimates of stillbirth rates in 2009 with trends since 1995: a systematic analysis. *The Lancet*, 377(9774):1319–1330, 2011.

Joao Paulo Souza and Rajiv Bahl. Purpose study: understanding the burden of stillbirths in south asia. *The Lancet Global Health*, 10(7):e930–e931, 2022.

World Health Organization et al. The who application of icd-10 to deaths during the perinatal period: Icd-pm. 2016.

Donna L Hoyert and Elizabeth CW Gregory. Cause of fetal death: data from the fetal death report, 2014. 2016.

Hugo Behm. Determinantes económicos y sociales de la mortalidad en américa latina. *Salud colectiva*, 7:231–253, 2011.

Alfredo Bolsi, Pablo Paolasso, and Fernando Longhi. El norte grande argentino entre el progreso y la pobreza. *Población & sociedad*, (12-13):227–283, 2005.

Victoria Mazzeo. La mortalidad infantil en argentina. análisis de sus cambios y de las diferencias regionales. *Población y Desarrollo-Argonautas y caminantes*, 10:9–20, 2014.

Valeria F Chapur, Emma L Alfaro, Rubén Bronberg, and José E Dipierri. Relación de la mortalidad infantil con la altura geográfica en el noroeste argentino. *Archivos argentinos de pediatría*, 115(5):462–469, 2017.

Abdel R Omram. The epidemiologic transition: a theory of the epidemiology of population change. *Bulletin of the World Health Organization*, 79(2):161–170, 2001.

Aldo Rosano, Lorenzo D Botto, Beverley Botting, and Pierpaolo Mastroiacovo. Infant mortality and congenital anomalies from 1950 to 1994: an international perspective. *Journal of Epidemiology & Community Health*, 54(9):660–666, 2000.

Rubén Adrian Bronberg, Valeria Fernanda Chapur, and José Edgardo Dipierri. Tendencia secular (1980-2018) de las muertes infantiles por malformaciones congénitas en argentina. *Revista de la Facultad de Ciencias Médicas*, 78(3):287, 2021.

Elvira B Calvo and Ana Biglieri. Impacto de la fortificación con ácido fólico sobre el estado nutricional en mujeres y la prevalencia de defectos del tubo neural. *Archivos argentinos de pediatría*, 106(6):492–498, 2008.

Ruben Bronberg, Jorge Martinez, Leonardo Morales, Anahi Ruderman, Damian Taire, Virginia Ramallo, and Jose Dipierri. "prevalence and secular trend of neural tube defects in fetal deaths in argentina, 1994–2019". *Birth Defects Research*, 115(18):1737–1745,

2023. doi: https://doi.org/10.1002/bdr2.2248. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bdr2.2248.

Foro de las Sociedades Respiratorias Internacionales. El impacto global de la enfermedad respiratoria. *Ciudad de México: Asociación Latinoamericana de Tórax*, 2017.

Sociedad Argentina de Pediatría and Comités Subcomisiones. Recomendaciones para el manejo de las infecciones respiratorias agudas bajas en menores de 2 años. resumen ejecutivo. *Arch Argent Pediatr*, 113(4):373–374, 2015.

H Cody Meissner. Viral bronchiolitis in children. *New England Journal of Medicine*, 374 (1):62–72, 2016.

Pablo Obando-Pacheco, Antonio José Justicia-Grande, Irene Rivero-Calle, Carmen Rodríguez-Tenreiro, Peter Sly, Octavio Ramilo, Asunción Mejías, Eugenio Baraldi, Nikolaos G Papadopoulos, Harish Nair, et al. Respiratory syncytial virus seasonality: a global overview. *The Journal of infectious diseases*, 217(9):1356–1364, 2018.

Cannizzaro, A., & Nuñez de la Rosa, D. El covid-19 según el clima, 2020. URL https://cenpat.conicet.gov.ar/el-covid-19-segun-el-clima/#:~:text=Los%20estudios%20sugieren%20que%20los,de%20este%20agente%20infeccioso%20nuevo. Last accessed 1 February 2024.

Harish Nair, Eric AF Simões, Igor Rudan, Bradford D Gessner, Eduardo Azziz-Baumgartner, Jian Shayne F Zhang, Daniel R Feikin, Grant A Mackenzie, Jennifer C Moïïsi, Anna Roca, et al. Global and regional burden of hospital admissions for severe acute lower respiratory infections in young children in 2010: a systematic analysis. *The Lancet*, 381(9875):1380–1390, 2013.

Virginia E Pitzer, Cécile Viboud, Wladimir J Alonso, Tanya Wilcox, C Jessica Metcalf, Claudia A Steiner, Amber K Haynes, and Bryan T Grenfell. Environmental drivers of the spatiotemporal dynamics of respiratory syncytial virus in the united states. *PLoS pathogens*, 11(1):e1004591, 2015.

Pablo Barneche, Agustina Bugallo, Hilario Ferrea, Marcia Ilarregui, Carolina Monterde, Ma Virginia Pérez, Tamara Santa María, Sebastián Serrano, and Karina Angeletti. Mé-

todos de medición de la pobreza. conceptos y aplicaciones en américa latina. *Entrelíneas de la Política Económica*, 26(4):31–41, 2010.

Juan Carlos Feres and Xavier Mancero. *El método de las necesidades básicas insatisfechas (NBI) y sus aplicaciones en América Latina*. Cepal, 2001.

Sergio Andés Kaminker. Segregación residencial y proyectos de ciudad: Puerto madryn como espacio en disputa. 2015.

Sergio Andrés Kaminker. Segregación residencial en puerto madryn, chubut (1991-2010): formas y efectos de una urbanización acelerada en una ciudad intermedia de la patagonia central. 2016.

Michael Marmot. Social determinants of health inequalities. *The lancet*, 365(9464): 1099–1104, 2005.

Paula A Braveman. Swimming against the tide: challenges in pursuing health equity today. *Academic medicine*, 94(2):170–171, 2019.