Herramienta para la exploración de tendencias y detección de patrones epidemiológicos en Argentina

Morales Arturo Leonardo[1, 2, 3], Figueroa Marcelo[4] , Delrieux Claudio[3], Ramallo Virginia[1], Dipierri José Edgardo[4]

[1] Instituto Patagónico de Ciencias Sociales y Humanas, CENPAT-CONICET.
[2] Departamento de Informática Trelew, Facultad de Ingeniería, UNPSJB.
[3] Departamento de Ingeniería Eléctrica y de Computadoras, UNS.
[3] Unidad de Genética Médica. Hospital Materno Infantil. Jujuy, Argentina.
lmorales@cenpat-conicet.gob.ar

Abstract. Este trabajo tiene como objetivo presentar la primera etapa del desarrollo de una herramienta informática que analiza los datos del registro sistematizado de fallecimientos y sus causas (codificadas según CIE-9 y CIE-10) entre los años 1991 y 2017 (Ministerio de Salud de la Nación Argentina), combinándola con la información demográfica del Instituto Nacional de Estadísticas y Censos (INDEC). El objetivo fue permitir el ingreso de conjuntos de códigos de causas de óbito, definidos según el interés de investigación, y obtener rápidamente reportes epidemiológicos para el periodo de tiempo indicado. El entrecruzamiento de la información de fallecimientos con la información demográfica y su referenciación geográfica (departamental, provincial y regional) nos permite calcular, por cada año, las tasas, incidencia y mortalidad, y visualizarlas en los reportes, diferenciando por sexo, grupo etario o zona geográfica. Mediante la utilización de herramientas para el análisis de datos del lenguaje Python (Pandas, Geopandas, Seaborn, Ploomber) se confeccionó un proceso de datos con alta capacidad de mantenimiento, flexibilidad y evolución. Considerando la variabilidad de los posibles estudios epidemiológicos en Argentina (estudio de enfermedades poco frecuentes, respiratorias, congénitas, etc.), se prevé que una herramienta de este tipo podrá acelerar y enriquecer la exploración y el análisis de patrones.

Keywords: Epidemiología, Minería de datos, Visualización de la Información.

1 Introducción

El estudio de las causas de los fallecimientos constituye una fuente de información fundamental para la planificación y elaboración de políticas públicas de salud de un país. Durante la ejecución de dichos estudios, una parte del enfoque está dedicada a cuantificar los fallecimientos (por motivos generales y por causas específicas), conocer cómo evolucionan esas cantidades a lo largo del tiempo, verificar cómo es su relación con el tamaño poblacional y establecer dónde se producen y a qué edades, para luego, con esta base descriptiva constituida, poder elaborar hipótesis más robustas acerca del por qué se producen [1].

Las causas de los fallecimientos se nominan utilizando el sistema de Clasificación Internacional de Enfermedades o CIE (en inglés ICD, por International Classification of Disease). Este sistema es mantenido por la Organización Mundial de la Salud desde 1948 y en su décima revisión, CIE-10, es mucho más específico que su antecesor CIE-9, con una expansión que va desde los 17.000 códigos a aproximadamente 155.000 [2].

La detección de patrones de comportamiento de distintas enfermedades toma fuerza desde la perspectiva multidisciplinaria que proponen los estudios sindémicos. En ellos se busca comprender cómo interactúan (o no) las enfermedades, identificar si se presentan juntas o de forma secuencial en el tiempo y reconocer cuáles son los entornos sociales y ambientales en los que éstas acontecen [3].

El presente trabajo de investigación resume la primera etapa del desarrollo de una herramienta de análisis de datos, con bibliotecas de código abierto, que permita la visualización de la información y el comportamiento de las causas de fallecimientos en Argentina entre 1991 y 2017. Se presentan aquí las transformaciones aplicadas sobre los datos que, en conjunto y dispuestas de forma ordenada, conformaron un pipeline de fácil parametrización, evolución y extensibilidad hacia nuevos análisis y nuevos hallazgos.

2 Materiales y métodos

2.1 Datos de entrada

El registro de fallecimientos 1991-2017 se compone de un archivo tabulado por cada año, pudiendo variar de uno a otro el esquema definido para estructurar la información. Este registro fue obtenido a través de la Dirección de Estadísticas e Información de la Salud del Ministerio de Salud de la Argentina.

La segunda fuente de información está conformada por las proyecciones de población para el mismo periodo 1991-2017. Se obtuvo a partir de los informes de estimaciones y proyecciones de población para el total del país 1950-2015 (según el Censo Nacional de Población, Hogares y Viviendas 1991) en combinación con el informe análogo para el período 2010-2040 (según el Censo Nacional de Población, Hogares y Viviendas 2010). Para los años en donde la información se solapa, es decir de 2010 a 2017, se optó por preservar la estimación del informe más reciente.

2.2 Procesamiento

Los principales indicadores calculados fueron incidencia y tasas de muerte por causas específicas (CSMR del inglés cause-specific mortality rate), entendida como cantidad de muertes debido a causas específicas por cada 1000 muertes en un área geográfica. En este trabajo agrupamos registros en distintas áreas geográficas según la división político-administrativa argentina. La incidencia es expresada como cantidad de muertes por causas específicas por cada 1000 habitantes y se calculó utilizando las proyecciones de población del INDEC.

Para la implementación del pipeline se utilizó la biblioteca *Ploomber*, en su versión 0.20, del lenguaje de programación *Python*. Este paquete permite realizar tareas de minería de datos a partir de la definición de las transformaciones que queremos aplicar sobre los datos. Cada tarea se implementa entonces en forma de funciones *Python*. Así,

el pipeline se construye de forma clara, trazable y mantenible. Su definición se realiza incorporando tareas a un archivo en formato *yaml*, las que generan como resultado un producto en formato *csv*, *parquet*, *pdf*, entre otros. Estas tareas implementadas como funciones reciben además un medio para poder trabajar sobre productos de otras tareas. Luego, *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 cuando realicemos algún cambio en su implementación. En posteriores ejecuciones del pipeline entero, se reutilizan los productos ya procesados. Como última mención acerca de esta herramienta, destacamos que *Ploomber* nos permite indicar parámetros de entrada al pipeline. En nuestro caso, los principales parámetros fueron los códigos de las causas específicas y los criterios de inclusión de cada grupo etario.

3 Implementación del pipeline

Conociendo la manera dispuesta por la biblioteca *Ploomber* para la implementación de un pipeline, procedemos a describir brevemente las transformaciones aplicadas sobre los datos. La Fig. 1 presenta una vista general del procesamiento.

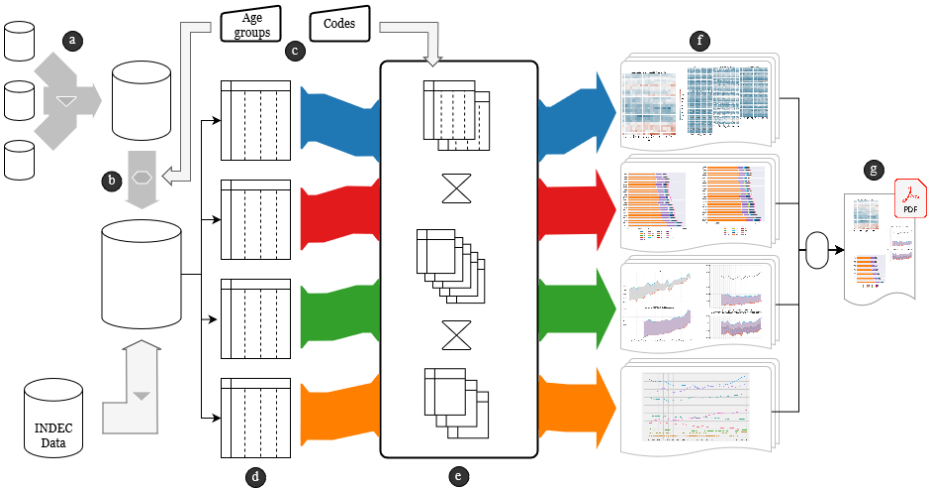

Fig. 1. Pipeline de visualización de la información. Describe las transformaciones aplicadas sobre los datos y la obtención de las visualizaciones que hacen parte del reporte obtenido.

Los conjuntos de datos anuales que conforman la fuente de información principal (registro de fallecimientos) presentan esquemas distintos a lo largo del período. Es decir, la conformación de los campos es distinta entre períodos, en cantidad y denominación. Para la unificación (Fig. 1a) se realizó la interpretación de cada columna en cada registro anual para obtener un conjunto de datos uniforme con los siguientes campos: "*provincia_id*", "*departamento_id*", "*codigo_defuncion*", "*sexo*", "*edad*", "*unidad_edad*", "*año*", "*mes_defuncion*".

Con la primera estandarización de los datos crudos, se procedió a generar campos que faciliten nuestros análisis, puntualmente en tres ejes: sexo, grupo etario, referenciación geográfica. Esta limpieza de datos es referenciada en la Fig. 1b. Se estandarizó la columna sexo con valores 1, 2, 3 según sexo femenino, masculino o sin definir, respectivamente. Se generó una nueva columna grupo etario utilizando "*unidad_edad*" (que indica si el dato edad fue medido en días, meses o años) y "*edad*" (que corresponde a cuántos días, o cuántos meses, o cuántos años, según corresponda). *Ploomber* permite la definición de distintos parámetros a utilizarse en la ejecución del pipeline (Fig. 1c). Se definió entonces el parámetro de grupos etarios a considerar, especificando un rango de años con su etiqueta correspondiente, por ejemplo, la etiqueta "*0 - 5*" para el rango de cero a cinco años, "*6 - 15*" para el rango de seis a quince o "*>85*" para registros con edades mayores a 85 años. Para el eje de la referenciación geográfica se limpiaron los códigos de provincia y departamento para obtener nombres y códigos estandarizados y válidos según la recomendación del Ministerio del Interior de la República Argentina. A partir del código de provincia se asignó un identificador y un nombre de región correspondiente, para así permitir el filtrado y la conformación de unidades administrativas argentinas en cuatro niveles distintos (nacional, regional, provincial y departamental). Podríamos pensar esta selección de registros en los niveles descritos como cuatro vistas de la información que podemos comenzar a manipular (Fig. 1d). Cada vista posee sus características propias y darán como resultado distintos subproductos. Por ejemplo,cuando se analiza la vista de registros para toda la Argentina, se filtran aquellos que contengan un código de fallecimiento dentro del conjunto de códigos indicados por parámetro. Con estas nuevas sub-vistas podemos obtener valores totales y valores relativos a las causas investigadas, respectivamente, y generar visualizaciones en este punto o continuar trabajando los datos para obtener incidencia o CSMRs. Procesos análogos se ejecutan para las regiones, las provincias y los departamentos. Esta compleja mecánica, en la cual se re-utilizan procedimientos, (como por ejemplo la obtención de los valores de incidencia, o valores de tasa de muerte por causas específicas) está representada en la caja blanca en la Fig 1e.

Las visualizaciones obtenidas son diversas, dependientes de los interrogantes que se deseen responder y cada una conlleva un tratamiento distinto de los datos (Fig 1f). Por ejemplo, el pipeline genera gráficos de líneas con los valores de incidencias anuales, con el fin de conocer su tendencia. Este interrogante se analiza para unidades administrativas específicas, ya sea el país completo o una región o provincia, y además es refinado separando los registros según el sexo o según el grupo etario o según ambas vistas combinadas y generando nuevas versiones de los gráficos de líneas.

En la búsqueda de revelar la mayor cantidad de patrones posibles, el pipeline también confecciona gráficos con barras apiladas, que reflejan la proporción de códigos de fallecimientos para cada año. Eventualmente el conjunto de códigos especificados por parámetro puede ser muy amplio, por lo cual las barras expresan las proporciones de categorías de códigos. Estas categorías son definidas de acuerdo al manual del sistema CIE y la cantidad de códigos en cada categoría es dependiente del conjunto especificado por parámetro. Distintas visualizaciones con proporciones de esta índole se obtienen diferenciando áreas geográficas, sexo y/o grupos etarios.

Tareas un tanto más complejas se presentaron al trabajar con la jerarquía superior de la división administrativa argentina, con áreas geográficas incluidas dentro de otras. Por ejemplo, para todo el país se graficó un mapa de calor utilizando como datos de

entrada los valores anuales de CSMR de cada una de las provincias, agrupadas a su vez según las regiones a las que pertenecen. Asimismo, un nuevo mapa de calor más detallado se obtiene con los valores de CSMR anuales de cada departamento de una región, agrupados por provincia. Nuevas formas de visualizaciones similares resultaron más detalladas aún al considerar los campos sexo y/o grupo etario para su diseño y creación.

Como último paso, con todas las visualizaciones que se hayan especificado, se obtiene un reporte completo en formato PDF (Fig 1g).

Tabla 1. Ejemplos de parámetros del mismo pipeline definidos para distintos proyectos. Estos valores son especificados en un archivo de extensión yaml denominado *env.yaml*.

Proyecto	Nombre del parámetro	Tipo de dato	Ejemplo de valores
Dengue	**causes_codes**	Lista de cadenas	["A91X", "A90X"]

Dengue	**age_groups**	Diccionario	{
			"0 - 5": [0, 5],
			"6 - 15": [6, 15],
			...,
			"76 - 85": [76, 85],
			">85": [86, 100]
			}

Alzheimer	**causes_codes**	Lista de cadenas	["G300", "G301", "G308", "G309", "2941", "3310"]

Alzheimer	**age_groups**	Diccionario	{
			"Temprano 1": [31, 60],
			"Temprano 2": [61, 65],
			"Tardío 1": [66, 85],
			"Tardío 2": [86, 100]
			}

Enfermedades poco frecuentes #1	**causes_codes**	Lista de cadenas	["C150", "C151", …, "Q991", "Q992", "T572", "T883"]
Enfermedades poco frecuentes #2	**causes_codes**	Lista de cadenas	["A500", "B004", "D179", "L722", "Q820", …, "R161"]

4 Resultados

Se obtuvo un pipeline modularizado, trazable y fácilmente parametrizable. Las primeras dos características resultan fundamentales para nuestros casos de estudio, ya que nos permite incorporar nuevas funcionalidades (en este trabajo, nuevas transformaciones de los datos y visualizaciones de la información a partir de ellas) de forma simple y clara dentro del flujo de ejecución de tareas. Esto es posible definiendo

nuevas funciones *Python* que trabajen sobre los datos, conectando dichas funcionalidades con los productos y las entradas de otras tareas, y resultando este proceso rápidamente verificable con la función de graficación de la secuencia de pasos ofrecida por *Ploomber*. La fácil parametrización reside en el mantenimiento de un solo punto que contiene todos los parámetros que serán reutilizables por cualquier tarea dentro del pipeline. Nos permite la conformación de conjuntos diversos (y fácilmente modificables) de códigos de causas de fallecimiento y variar la manera en que conformamos los grupos etarios, según las características y el enfoque del análisis. La Tabla 1 muestra ejemplos de parámetros que han sido utilizados en distintos abordajes con el mismo pipeline. El código fuente del proyecto es mantenido por los autores de este trabajo y la versión pública de su repositorio se encuentra disponible en el enlace: https://github.com/LeoMorales/epidemiology-pipeline. En su versión actual, el proyecto admite la carga de los parámetros mediante un archivo de parámetros siguiendo la forma dispuesta por la librería *Ploomber*. La creación de las interfaces adecuadas para facilitar esta carga, será parte de los siguientes pasos.

5 Conclusiones

La implementación de este pipeline conforma una primera iteración de un trabajo de analítica de datos que busca facilitar el análisis de la distribución y la tendencia de fallecimientos por causas específicas en estudios epidemiológicos. Entendiendo que las enfermedades no ocurren al azar en una población, sino que se presentan sólo cuando existe la acumulación correcta de factores de riesgo o determinantes en un individuo [4], se busca acelerar la fase de exploración y prueba de supuestos, para dar lugar a la siguiente etapa, aún más relevante, referida a la explicitación, explicación y prevención de los factores determinantes.

Referencias

1. Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., ... & Remuzzi, G. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. The lancet, 380(9859), 2095-2128.
2. Hirsch, J. A., Nicola, G., McGinty, G., Liu, R. W., Barr, R. M., Chittle, M. D., & Manchikanti, L. (2016). ICD-10: history and context. American Journal of Neuroradiology, 37(4), 596-599.
3. Singer, M., Bulled, N., Ostrach, B.M., & Mendenhall, E. (2017). Syndemics and the biosocial conception of health. The Lancet, 389, 941-950..
4. Dicker, R., Coronado, F., Koo, D., & Parrish, R. G. (2006). Principles of epidemiology in public health practice. Atlanta GA: US Department of Health and Human Services, 512.