Base de datos epidemiológica, socioeconómica y demográfica COVID-19 en Chile y el mundo

Jessica Pavani[2], Luis Gutiérrez[1], Gabriel Arriagada[2], Jaime Cerda[3], Leonardo Jofré[2], Danilo Garrido[2], Inés Varas[4], Oscar Ortiz[1], and Iván Gutiérrez[2]

[1] *Instituto de Ciencias Agroalimentarias, Animales y Ambientales – ICA3, Universidad de O'Higgins*
[2] *Facultad de Matemáticas, Departamento de Estadística, Pontificia Universidad Católica de Chile*
[3] *Facultad de Medicina, Departamento de Salud Pública, Pontificia Universidad Católica de Chile*
[4] *Núcleo Milenio Centro para el Descubrimiento de Estructuras en Datos Complejos (MiDaS)*

Resumen

La necesidad de tener datos para desarrollar estudios respecto al Covid-19 es una prioridad para los investigadores. Se desarrolla una base de datos que contiene información epidemiológica, socioeconómica, medidas sanitarias aplicadas e información demográfica, tanto para Chile como para una gran cantidad de países del mundo. La base de datos se utiliza como insumo para distintos modelos en el contexto del proyecto Epi-Covid.

1. Introducción

Para poder estudiar los efectos del COVID-19 en distintas variables de interés se requiere de bases de datos, que se encuentren actualizadas y contemplen la mayor cantidad de información posible. En Chile y el mundo se ha disponibilizado datos epidemiológicos de la pandemia del COVID-19, los cuales requieren de un proceso de estandarización y constante actualización. En el caso chileno, el Ministerio de Ciencia ha disponibilizado un repositorio de GitHub (6), en donde la información es de acceso público, con actualización periódica, en algunos casos diaria y en intervalos de 2 a 3 días. También, se disponibilizan algunos datos demográficos, provenientes del Censo de Población y Vivienda 2017 (4). A nivel mundial, existen distintas iniciativas que buscan generar repositorios de información para su uso en investigación relacionada al COVID-19. La universidad John Hopkins o la universidad de Oxford a través de su iniciativa Our World in Data (3) han disponibilizado información epidemiológica de la gran mayoría de los países del planeta.

Si bien existe abundante información de acceso público, esta generalmente no suele presentarse en un solo repositorio, acumulando información epidemiológica, socioeconómica y demográfica. Distintas disciplinas como la estadística, la economía y la medicina requieren de una base de datos que contenga la información necesaria para responder preguntas de interés respecto a la pandemia del COVID-19. Sin embargo, actualmente no existe una base de datos que aúne todos estos datos.

2. Objetivos

El objetivo de este trabajo consistió en crear una base de datos para Chile y el Mundo que recopilara información epidemiológica, socioeconómica y demográfica. La base de datos se actualiza automáticamente y recopila la información de fuentes oficiales, tanto en Chile como en el Mundo.

3. Metodología

Se desarrolló una metodología de ingesta y actualización automática, utilizando las herramientas **Docker** y **Amazon Web Services (AWS)**.

3.1. Bases de datos nacionales

Los datos provienen del *GitHub* del Ministerio de Ciencia y Tecnología de Chile. Mediante la creación de un contenedor en *Docker*, se agendó un trabajo que descarga toda la *raw-data* desde el repositorio del Ministerio de Ciencia y lo almacena en la plataforma *S3* de *AWS*. La *raw-data* almacenada, posteriormente es pre-procesada mediante la herramienta *Glue* de *AWS*, en el cual se establecen los esquemas, es decir, la estructura que llevara cada tabla, se estandarizan los datos en términos de tipo de datos, se corrigen errores de escritura y se guardan en formato *parquet*, con el propósito de poder ser consultados vía *SQL* en la herramienta *Athena* de *AWS*.

Si bien el repositorio en *GitHub* del Ministerio de Ciencia, cuenta con una gran cantidad de información relevante, esta no incluye información socioeconómica o demográfica de las distintas comunas en Chile. Para poder obtener dicha información, se incorporaron distintas variables tanto en tablas, como también se incorporaron datos georreferenciados, consultando las bases de datos del Censo 2017 (5), Observatorio Chileno de salud Pública (8) y Ministerio de Desarrollo Social (7).

3.2. Bases de datos Mundiales

La data internacional epidemiológica fue obtenida del trabajo realizado por (2), en el cual se disponibiliza en formato estándar, la información recopilada de distintas entidades gubernamentales de los países del mundo. Se cuenta con información de más de 200 países, agregados a nivel nacional, de alrededor de 30 países a nivel regional y cerca de 10 países a nivel comunal. La data cuenta con las medidas sanitarias implementadas por los distintos países para combatir el Covid-19, además de contar con la data epidemiológica. Las medidas se indican tanto si la medida es recomendada, como si la medida es obligatoria.

De la data internacional agregada a nivel nacional, se crean tres tablas, para distinto uso según el requerimiento de las distintas aproximaciones desarrolladas en el proyecto. La primera tabla corresponde a la data replicada y disponibilizada en *Athena* con los datos informados acumulados diarios, la segunda tabla corresponde a los datos sin acumular agrupados por semana epidemiológica, y la tercera tabla corresponde a los datos sin acumular, desde el inicio de cada cuarentena.

La ingesta y actualización de los datos es automática, realizada a través de un *job* en *Glue* de *AWS*.
La data económica internacional proviene de tres fuentes de información: Banco Mundial (**?**), Programa de las Naciones Unidas para el Desarrollo (PNUD) (UND) y Our World in Data (OWID) de la universidad de Oxford.

La data económica incluye información de PIB per cápita, índice de desarrollo humano, porcentaje de la población sobre 65 años, porcentaje de pobreza extrema, densidad poblacional, entre otras variables.

De forma análoga a la data económica para Chile, la data económica internacional presenta periodos de actualización anuales o más distantes en el tiempo. Particularmente, para el caso del *PIB per cápita de europa*, se logró obtener una estimación desestacionalizada del PIB per cápita informada por la *Organización para la Cooperación y el Desarrollo Económico (OCDE)*, para sus países miembros.

4. Resultados

Los resultados se dividen en dos categorías:

- **Tablas**: Las tablas corresponden a la información originales[1], obtenida de las fuentes de información directamente o de la creación propia del proyecto.
- **Vistas**: Las vistas corresponden a una conjunción de tablas que pueden ser unidas y transformadas a partir de las tablas originales. Estas son creadas específicamente para ser utilizadas en el modelamiento, ya que son procesadas según los requerimientos necesarios de formato.

A continuación, se describen las tablas que fueron utilizadas directamente o como insumo para los productos.

[1] A que te refieres con información originales?, datos sin modificaciones de su contenido...

4.1. Tablas

- **maestra_comunas:** Contiene información socioeconómica y de desarrollo para todas las comunas de Chile, como porcentaje de la población en sectores económicos, nivel de escolaridad, índice de desarrollo socioeconómico y porcentaje de población en zona urbana o rural.
- **p19_activos_dc:** da cuenta del número de casos confirmados activos notificados en cada una de las comunas de Chile, según residencia, y concatena la historia de los informes epidemiológicos publicados por el Ministerio de Salud del país.
- **p26_nuevos_csintomas_dr:** da cuenta del número de casos confirmados nuevos por día según resultado del diagnóstico y que han presentado síntomas, por región de residencia, reportados por el Ministerio de Salud. Esta serie de tiempo incluye los casos nuevos denominados çon síntomas"por la autoridad sanitaria.
- **p27_nuevos_sin_sintomas_dr:** da cuenta del número de casos confirmados nuevos por día según resultado del diagnóstico y que son asintomáticos, por región de residencia, reportados por el Ministerio de Salud. Estos archivos incluyen datos a partir del 29 de abril. Esta serie de tiempo incluye los casos nuevos denominados "sin síntomas"por la autoridad sanitaria a partir de tal fecha.
- **p8_camasuci_dr:** da cuenta del número de pacientes en UCI, y que son casos confirmados por COVID-19, por región reportados diariamente por el Ministerio de Salud, desde el 01-04-2020.

- **p38_fallecidos_dc:** da cuenta del número de casos fallecidos en cada una de las comunas de Chile según su residencia, y concatena la historia de los informes epidemiológicos publicados por el Ministerio de Salud del país. Se entiende por casos fallecidos a las muertes confirmadas debido a COVID-19 y que se encuentran debidamente registradas en la base de datos del Registro Civil e Identificación. Para estos casos, la comuna de residencia fue obtenida desde la plataforma EPIVIGILA, o bien, se asignó la comuna de circunscripción donde fue inscrita la defunción en los casos sin comuna de residencia.
- **p7_pcr_dr:** da cuenta del número de exámenes PCR realizados por región reportados diariamente por el Ministerio de Salud, desde el 09-04-2020.
- **censo:** da cuenta de la cantidad de personas y hogares por comuna, de todas las comunas del país. Además, desglosa la cantidad de población por grupo etareo, inmigrantes, mujeres, y miembros pertenecientes a pueblos originarios. Para el caso de las viviendas, desglosa por viviendas particulares y colectivas.
- **ips_2019:** da cuenta del **Índice de prioridad social** de las comunas pertenecientes a la Región Metropolitana, calculado por el Ministerio de Desarrollo Social.
- **indice_ruralidad:** da cuenta del índice de ruralidad calculado por el Ministerio de Desarrollo Social, y extendido para todas las comunas de Chile.
- **idse:** contiene información socioeconómica para 322 comunas de Chile. Contiene índices de educación, alcantarillado, pobreza, esperanza de vida y otros, los cuales combina para crear el **Índice de desarrollo socioeconómico (IDSE)**.
- **p29_cuarentena:** contiene la identificación y características de las zonas de cuarentena establecidas por el Plan de Acción por Coronavirus del Gobierno de Chile. Las zonas de cuarentena se establecen como una medida sanitaria en una extensión territorial definida que implica que las personas deben permanecer en sus domicilios habituales hasta que la autoridad disponga lo contrario. Los criterios para la definir la cuarentena son: Velocidad de Propagación de la Enfermedad, Densidad de casos por km2, Perfil etáreo de la población del territorio (adultos mayores y personas con enfermedades crónicas), Vulnerabilidad Social.
- **p1_casos_tot_acum_dc:** da cuenta de los casos confirmados y probables notificados (desde el 19 de junio, informe #27 se incluyen los casos probables) en cada una de las comunas de Chile, según residencia, y concatena la historia de los informes epidemiológicos publicados por el Ministerio de Salud del país.
- **data_mundial_s:** Da cuenta de los casos semanales confirmados sin acumular, recuperados sin acumular, fallecidos sin acumular, hospitalizados, tests PCR aplicados, y diversas medidas sanitarias aplicadas por los gobiernos (cierre de escuelas, cierre de espacios de trabajo, cancelación de eventos, restricciones de reunión, restricciones de transporte público, cuarentenas, restricciones de desplazamiento interno, cierre de fronteras,

presencia de políticas de difusión de información, política de testeo, rastreo de casos), para todos los países del mundo a nivel país, partiendo desde la primera semana epidemiológica (semana en que se manifiesta el primer caso positivo).

4.2. Información de variables

Las datos puedes ser clasificados en cuatro categorías:

- Datos epidemiológicos
- Datos socioeconómicos
- Medidas sanitarias
- Datos demográficos

Además, para cada categoría se cuenta con información específica para Chile (tabla 1), de donde la información proviene de fuentes nacionales y creación propia. También se cuenta con información mundial (tabla 2). En ambas tablas se describe la fuente, el periodo de actualización y el nivel de resolución de los datos.

Chile			
Datos epidemiológicos			
	Fuente	**Actualización**	**Resolución**
Casos nuevos	Ministerio de Ciencia	Día, Semana	Región, Comuna
Casos activos	Ministerio de Ciencia	Día	Comuna
Casos recuperados	Ministerio de Ciencia	Día	País
Test PCR	Ministerio de Ciencia	Día	Región
Fallecidos	Ministerio de Ciencia	Día	Comuna
Positividad	Ministerio de Ciencia	Día	Comuna
Camas UCI	Ministerio de Ciencia	Día	Región
Vacunación	Ministerio de Ciencia	Día	Comuna
Datos socioeconómicos			
	Fuente	**Actualización**	**Resolución**
Índice de prioridad social (2019)	Ministerio de Desarrollo Social	Fijo	Comuna
Indice de desarrollo socioeconómico	Ochisap	Fijo	Comuna
Presencia aeropuerto	Epi-Covid	Fijo	Comuna
Presencia puerto comercial	Epi-Covid	Fijo	Comuna
Años de escolaridad	Ochisap	Fijo	Comuna
Porcentaje viviendas con alcantarillado	Ochisap	Fijo	Comuna
Esperanza de vida al nacer	Ochisap	Fijo	Comuna
Proporción población en sectores productivos	Ochisap	Fijo	Comuna
Medidas sanitarias			
	Fuente	**Actualización**	**Resolución**
Plan paso a paso	Ministerio de Ciencia	Día	Comuna
Cuarentena	Ministerio de Ciencia	Día	Comuna
Datos demográficos			
	Fuente	**Actualización**	**Resolución**
Población (2020)	Censo 2017 (proyección)	Fijo	Comuna
Índice de ruralidad (2019)	Ministerio de Desarrollo Social	Fijo	Comuna
Índice de movilidad	Ministerio de Ciencia	Día	Comuna
Capital regional	Biblioteca del Congreso Nacional	Fijo	Región
Capital provincial	Biblioteca del Congreso Nacional	Fijo	Provincia

Tabla 1: Datos recopilados a partir de fuentes nacionales.

Mundo			
Datos epidemiológicos			
	Fuente	**Actualización**	**Resolución**
Casos nuevos	Our world in data	Día	País
Casos activos	Our world in data	Día	País
Casos recuperados	Our world in data	Día	País
Fallecidos	Our world in data	Día	País
Vacunación	Our world in data	Día	País
Datos socioeconómicos			
	Fuente	**Actualización**	**Resolución**
PIB per cápita	Banco Mundial	Fijo	País
Índice de desarrollo humano	PNUD	Fijo	País
Medidas sanitarias			
	Fuente	**Actualización**	**Resolución**
Cuarentena	Our world in data	Día	País
Cierre de escuelas	Our world in data	Día	País
Cierre de fronteras	Our world in data	Día	País
Cierre de espacios de trabajo	Our world in data	Día	País
Restricciones en transporte público	Our world in data	Día	País
Restricciones de reunión	Our world in data	Día	País
Cancelación de eventos	Our world in data	Día	País
Trazabilidad de casos	Our world in data	Día	País
Datos demográficos			
	Fuente	**Actualización**	**Resolución**
Población (2020)	Our world in data	Fijo	País

Tabla 2: Datos mundiales recopilados de fuentes internacionales.

5. Conclusiones preliminares

Se crean bases de datos que contienen información de distintas dimensiones. Estos datos fueron procesados, chequeados y disponibilizados para su utilización en modelos estadísticos que buscan responder preguntas de investigación respecto a la pandemia del Covid-19. Se crearon datos tanto en detalle para el caso de Chile, como datos para los países del mundo, a nivel nacional y con actualización diaria.

La actualización periódica de la base de datos, permite un análisis con datos actualizados, permitiendo realizar análisis de toda ventana de tiempo desde que los datos han sido disponibilizados.

Referencias

[UND] Human Development Report Office Statistical Data API | Human Development Reports.

[2] Guidotti, E. and Ardia, D. (2020). Covid-19 data hub. *Journal of Open Source Software*, 5(51):2376.

[3] Hasell, J., Mathieu, E., Beltekian, D., Macdonald, B., Giattino, C., Ortiz-Ospina, E., Roser, M., and Ritchie, H. (2020). A cross-country database of covid-19 testing. *Scientific data*, 7(1):1–7.

[4] Instituto Nacional de Estadísticas (2017a). Microdatos censo 2017.

[5] Instituto Nacional de Estadísticas (2017b). Servicio de mapas del censo 2017.

[6] Ministerio de Ciencia (2020). Datos-covid19.

[7] Ministerio de Desarrollo Social (2019). Índice de ruralidad comunal 2019.

[8] Observatorio Chileno de Salud Pública (2013). Nivel socioeconómico y de salud de las comunas de chile.