Epidemiological, socioeconomic, and demographic COVID-19 database for Chile and globally

Oscar Ortiz[1,5], Danilo Garrido[2], Leonardo Jofré[2], Iván Gutiérrez[2], Jessica Pavani[2], Inés Varas[4], Luis Gutiérrez[1], Gabriel Arriagada[2], and Jaime Cerda[3]

[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)*
[5] *Facultad de Ingeniería, Escuela de Ingeniería, Pontificia Universidad Católica de Chile*

Abstract

The need for data to develop studies on COVID-19 is a priority for researchers. We generated a database containing epidemiological, socioeconomic, applied sanitary measures, and demographic information, both for Chile and many countries in the world. The database could be used as input for different models in the context of the Epi-Covid project.

1 Introduction

Databases are required to study the effects of policy responses on COVID-19, which should be up-to-date and include as much information as possible. In Chile and the world, epidemiological data of the COVID-19 pandemic have are available, requiring standardization and constant updating. In the Chilean case, the Ministry of Science, Technology, Knowledge, and Innovation has made available a GitHub repository (7), where the information is publicly accessible, with periodic updates, in some cases daily and at intervals of 2 to 3 days. Also, some demographic data is available from the Population and Housing Census 2017 (5). Worldwide, different initiatives seek to generate repositories of information for use in research related to COVID-19. The Johns Hopkins University or the University of Oxford, through their initiative Our World in Data (4), have available epidemiological information from the vast majority of countries on the planet.

Although there is a lot of information available to the public, it is generally not presented in a single repository, accumulating epidemiological, socioeconomic, and demographic information. Different disciplines such as statistics, economics, and medicine require a database that contains the information necessary to answer questions of interest regarding the COVID-19 pandemic. However, there is currently no database that brings together all this data.

2 Goal

The objective of this work was to create a database for Chile and the World that would collect epidemiological, socioeconomic, and demographic information associated with COVID-19. The database is automatically updated and compiles data from official sources, both in Chile and globally.

3 Methodology

An automatic ingest and update methodology was developed, using the tools **Docker** and **Amazon Web Services (AWS)**.

3.1 Data for Chile

The data comes from the GitHub of the Chilean Ministry of Science and Technology. We scheduled a job that downloads all the raw data from the Ministry of Science repository and stores it on AWS S3. Later, we processed

raw data with AWS Glue, setting the schema and structure of the databases. We create the structure that will carry each table, set data type, correct writing errors, and save it in parquet format for SQL compatibility.

Although the GitHub repository of the Ministry of Science has a large amount of relevant information, it does not include the socioeconomic or demographic data of cities in Chile. We solved this issue, joining GitHub data with socioeconomic or demographic data obtained from the 2017 Census (6), Chilean Public Health Observatory (9) and Ministry of Social Development (8).

3.2 Data globally

We obtained international epidemiological data from the COVID-19 Data Hub (3), which is already in a standard format, and it is available for more than 200 countries, aggregated at the national level, about 30 countries at the regional level, and about ten countries at the communal level. The data includes info about policy responses implemented by governments to combat Covid-19, in addition to epidemiological data.

From the international data added at the national level, we created three tables for different approaches developed in the project. The first table corresponds to the data replicated and available in AWS Athena with the daily accumulated data. The second table corresponds to the non-accumulated data grouped by epidemiological week. The third table corresponds to the non-accumulated data from the start of each quarantine.

We automated data wrangling creating a job in AWS Glue.
International economic data comes from three information sources: World Bank (2), United Nations Development Programme (UND) and Our World in Data (OWID) from the University of Oxford.
The economic data includes information about GDP per capita, human development index, percentage of the population over 65, percentage of extreme poverty, population density, among other variables.

Analogous to the economic data for Chile, the international economic data presents annual update periods or more distant in time. In particular, for the case of the GDP per capita of Europe, a seasonally adjusted estimate of GDP per capita reported by the Organization for Economic Cooperation and Development (OECD) can be obtained for its member countries.

4 Results

The results are divided into two categories:

- **Tables**: The tables correspond to the information obtained from the information sources directly or from the project's own creation.
- **Views**: The views correspond to a conjunction of tables that can be joined and transformed from the original tables. These are created specifically to be used in modeling, since they are processed according to the necessary format requirements.

Next, the tables that were used directly or as input for the products are described.

4.1 Tables

- **maestra_comunas:** It contains socioeconomic and development information for all communes in Chile, as a percentage of the population in economic sectors, level of education, socioeconomic development index, and percentage of population in urban or rural areas.
- **p19_activos_dc:** It gives an account of the number of confirmed active cases notified in each of the communes of Chile, according to residence, and concatenates the history of the epidemiological reports published by the country's Ministry of Health.
- **p26_nuevos_csintomas_dr:** It gives an account of the number of new confirmed cases per day according to the result of the diagnosis and that have presented symptoms, by region of residence, reported by the Ministry of Health. This time series includes new cases called "with symptoms" by the health authority.

- **p27_nuevos_sin_sintomas_dr:** It gives an account of the number of new confirmed cases per day according to the result of the diagnosis and that are asymptomatic, by region of residence, reported by the Ministry of Health. These files include data as of April 29. This time series includes new cases called "without symptoms" by the health authority as of that date.
- **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 Views

4.3 Data properties

The data can be classified into four categories:

- Epidemiological data
- Socio-economic data

- Policy responses
- demographic data

In addition, for each category there is specific information for Chile (table 1), from where the information comes from national sources and own creation. World information is also available (table 2). Both tables describe the source, the update period and the level of resolution of the data.

4.3.1 Data for Chile

Chile				
Datos epidemiológicos				
	Fuente	**Tipo de dato**	**Actualización**	**Resolución**
Casos nuevos	Ministerio de Ciencia	Dinámico	Diaria (2 - 3 días)	Región, Comuna
Casos activos	Ministerio de Ciencia	Dinámico	Diaria (2 - 3 días)	Comuna
Casos recuperados	Ministerio de Ciencia	Dinámico	Diaria	País
Test PCR	Ministerio de Ciencia	Dinámico	Diaria	Región
Fallecidos	Ministerio de Ciencia	Dinámico	Diaria (2 - 3 días)	Comuna
Positividad	Ministerio de Ciencia	Dinámico	Diaria	Comuna
Camas UCI	Ministerio de Ciencia	Dinámico	Diaria	Región
Vacunación	Ministerio de Ciencia	Dinámico	Diaria	Comuna
Datos socioeconómicos				
	Fuente	**Tipo de dato**	**Actualización**	**Resolución**
Índice de prioridad social (2019)	Ministerio de Desarrollo Social	Estático	Fijo	Comuna
Indice de desarrollo socioeconómico	Ochisap	Estático	Fijo	Comuna
Presencia aeropuerto	Epi-Covid	Estático	Fijo	Comuna
Presencia puerto comercial	Epi-Covid	Estático	Fijo	Comuna
Años de escolaridad	Ochisap	Estático	Fijo	Comuna
Porcentaje viviendas con alcantarillado	Ochisap	Estático	Fijo	Comuna
Esperanza de vida al nacer	Ochisap	Estático	Fijo	Comuna
Proporción población en sectores productivos	Ochisap	Estático	Fijo	Comuna
Medidas sanitarias				
	Fuente	**Tipo de dato**	**Actualización**	**Resolución**
Plan paso a paso	Ministerio de Ciencia	Dinámico	Diaria	Comuna
Cuarentena	Ministerio de Ciencia	Dinámico	Diaria	Comuna
Datos demográficos				
	Fuente	**Tipo de dato**	**Actualización**	**Resolución**
Población (2020)	Censo 2017 (proyección)	Estático	Fijo	Comuna
Índice de ruralidad (2019)	Ministerio de Desarrollo Social	Estático	Fijo	Comuna
Índice de movilidad	Ministerio de Ciencia	Dinámico	Diaria	Comuna
Capital regional	Biblioteca del Congreso Nacional	Estático	Fijo	Región
Capital provincial	Biblioteca del Congreso Nacional	Estático	Fijo	Provincia

Table 1: Datos recopilados a partir de fuentes nacionales.

4.3.2 Data globally

5 Preliminary conclusions

Databases containing information of different dimensions are created. These data were processed, checked and available for use in statistical models that seek to answer research questions regarding the Covid-19 pandemic. Data were created both in detail for the case of Chile, as well as data for the countries of the world, at the national level and with daily updates.

The periodic update of the database allows an analysis with updated data, allowing an analysis of the entire time window since the data has been available.

References

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

[2] Banco Mundial (2021). Databank.

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

Table 2: Datos mundiales recopilados de fuentes internacionales.

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

[4] 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.

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

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

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

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

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