Hacia una medición de la dimensión real de la pandemia de COVID-19 en Chile en alta resolución espacial y temporal *

Gonzalo Mena
Statistics Department and Data Science Initiative, Harvard University

10 de Junio, 2020

Resumen

Conocer la verdadera dimensión de una epidemia en tiempo real y a escalas geográficas es fundamental para la correcta estimación de parámetros epidemiológicos, y así mejorar la toma de decisiones. Lamentablemente, los datos oficiales pueden entregar una medida sesgada de la realidad, debido a que, por ejemplo, el testeo no es aleatorio, puede haber un retraso de varios días entre el contagio y la confirmación a través de virología, etc. En este informe nos enfocamos en el problema del retraso. Primero, a nivel descriptivo, introducimos la *oportunidad*, un indicador de retraso que exhibe claros patrones de dependencia espacio-temporal, y que podía medir la saturación del sistema de salud. Segundo, a nivel inferencial, usando una reciente metodología bayesiana de *nowcasting* mostramos que es posible estimar con razonable precisión la cantidad de gente que presenta síntomas en un determinado día, a nivel nacional. Finalmente, comentamos las limitaciones tanto producto de nuestros métodos como de la escasez de datos publicados, y señalamos que tanto una mejor desagregación temporal del inicio de síntomas por comuna como los reportes de fallecidos por comuna serían muy útiles para suplementar este análisis.

1 Introducción

En situaciones tan desafiantes como la pandemia de COVID-19, es crucial tomar decisiones apropiadas con el fin de proteger a la población (Lipsitch and Santillana, 2019). En las sociedades avanzadas estas decisiones están idealmente basadas en evidencia (datos), los cuales son cuidadosamente curados, analizados y modelados por expertos en epidemiología, salud pública, estadística, matemática, etc. Lamentablemente, sesgos en la recolección de datos pueden tener un gran impacto en la estimación de los parámetros epidemiológicos que se usarán finalmente en la toma de decisiones Zhao et al. (2020); Donnat and Holmes (2020); Holmdahl and Buckee (2020).

Nuestro objetivo es corregir algunos de esos sesgos y así tener un mejor seguimiento de la dimensión real de la epidemia con la mayor resolución espacio-temporal posible, sin que eso comprometa la privacidad de los ciudadanos. En este informe nos enfocamos principalmente en la corrección de los retrasos en el reporte de casos. Ahí, el punto principal es que usualmente el diagnóstico de virología (PCR) puede ocurrir varios días después del contagio, y así, los reportes de casos diarios entregan en realidad una observación de un pasado más o menos reciente, dependiendo de cuánto se tarda en reportar un caso como tal. Dada la rapidez con la que avanza el contagio, cabe la posibilidad de que así las medidas que se tomen lleguen demasiado tarde para ser efectivas.

Aunque es imposible conocer el momento exacto del contagio, sí es posible al menos rastrear el momento de la aparición de síntomas, si es que estos eventos son también reportados. Específicamente, acá consideramos datos pareados donde existe correspondencia entre la fecha del diagnóstico y la fecha del inicio de síntomas. Esta correspondencia permite la inferencia de la distribución del tiempo que transcurre entre ambas fechas (el retraso), y basado en esta distribución, la estimación oportuna de la cantidad de casos cuyos síntomas se manifiestan por primera vez hoy, sin que necesariamente se hayan aún reportado. Este proceso de 'predicción del presente' se conoce como *nowcasting* (*ahoróstico*) epidemiológico (Bastos et al., 2019; McGough et al., 2020; Stoner and Economou, 2019), y puede significar un insumo valioso, debido a que permite una toma de decisiones más oportuna, previo a que se observe un

*Documento preliminar

abrupto incremento en los casos. Desde el punto de vista epidemiológico, la curva de contagios respecto al inicio de síntomas es una cantidad más fidedigna para la estimación de parámetros epidemiológicos usados en la toma de decisiones, como el R efectivo.

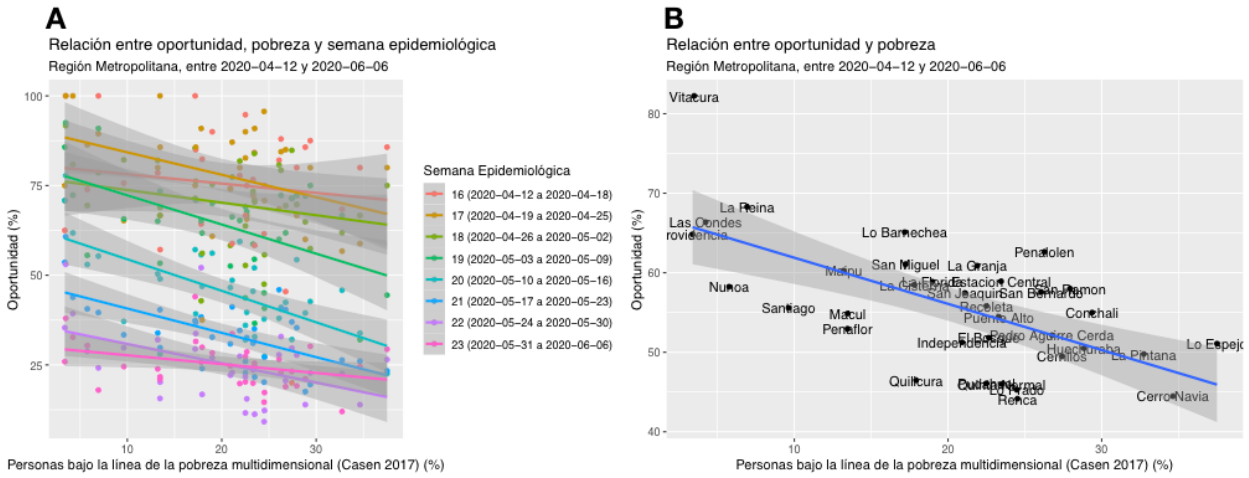

Figura 1: *Oportunidad* como función del tiempo y la pobreza multidimensional como proxy de vulnerabilidad social. **A** La *oportunidad* varía de acuerdo al nivel de pobreza y la semana epidemiológica. De hecho, se observan efectos importantes de estas dos componentes en una regresión del tipo $T_{t,s} = \gamma_0 + \gamma_1 t + \gamma_2 p_s$ donde $T_{t,s}$ es la *oportunidad* en la semana t y comuna s y p_s es la pobreza, con un con $R^2 \approx 0.62$. **B** Adicionalmente, al promediar los valores de la semana epidemiológica se observa una fuerte relación lineal (negativa) entre pobreza y *oportunidad*. Ahí, $R \approx -0.6$.

2 Resultados

Nuestro análisis considera aspectos descriptivos e inferenciales. A nivel descriptivo, mostramos que una simple cantidad, la *oportunidad* (*timeliness*) (Stoner et al., 2019) puede ser un indicador útil de la saturación del sistema de salud. Definimos la *oportunidad* como la proporción de los casos cuyos síntomas empezaron en la misma semana que se reportan (ver sección 4 para más detalles). Los resultados se resumen en las Figuras 1 y 2. La Figuras 1-**A** y 2 muestran un claro patrón de dependencia espacio-temporal en la *oportunidad*: esta ha decrecido consitentemente en las últimas semanas en la Región Metropolitana, presumiblemente como consecuencia del incremento de la intensidad de la pandemia en esta región. Aún más, la Figura 1(**A-B**) muestran que los efectos espaciales quedan bien capturados a través de un indicador *local* de vulnerabilidad social, la pobreza multidimensional (Casen, 2017).

Además, a nivel inferencial, mostramos en la Figura 3 que los datos permiten hacer un *nowcasting* razonable a nivel nacional. Sin embargo, para ser útiles en la toma de decisiones, el *nowcast* debería ejecutarse a nivel regional o comunal. Lamentablemente, ahí, los datos de la aparición de síntomas se encuentran sólo por semana epidemiológica y no a resolución diaria, lo que dificulta las inferencia. Aún así, estos datos muestran un interesante patrón (Figura 2): existe una heterogeneidad en la distribución del retraso (medido de manera gruesa) dependiendo de la comuna, y ésta correlaciona con el nivel de pobreza. La existencia de esta correlación indica que el seguimiento de las curvas por inicio de síntomas debe considerar el contexto sociodemográfico, y que a nivel descriptivo, puede ser oportuno reportar cantidades que den cuenta de la distribución del retraso en distintos lugares.

3 Discusión

3.1 Limitaciones

Nuestro objetivo a nivel descriptivo es promover el uso de la *oportunidad* como medida válida para capturar el retraso. Lo hicimos mostrando que exhibe patrones de correlación coherentes y con sentido. Sin embargo, un análisis más concienzudo es necesario para excluir la posiblidad de que la relación entre

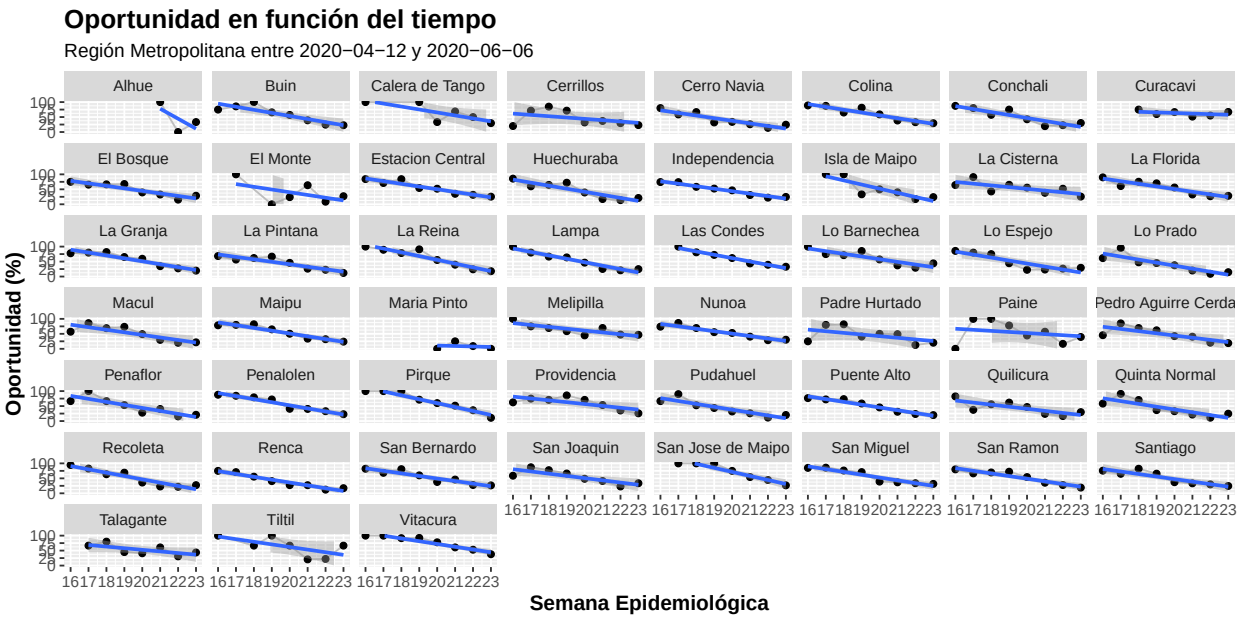

Figura 2: La *oportunidad* ha mostrado un consistente decrecimiento en las últimas 4 semanas en la Región Metropolitana, consistente con el aumento de su intensidad. Este decrecimiento queda bien representado por una recta en la gran mayoría de las comunas, aunque alrededor de la semana 18 hubo un breve hiato donde la *oportunidad* subió en muchas comunas (por ejemplo, Cerrillos).

oportunidad y pobreza esté mediada por otros factores, por ejemplo, la incidencia. Sin embargo, análisis preliminares sugieren que esto no es el caso.

Existen diversas limitaciones en el método de *nowcasting* que aplicamos: en primer lugar, nuestro método de imputación puede llevar a sesgos de estimación. En segundo lugar, el supuesto de que la distribución del retraso no cambia en el tiempo es muy simplista, y de hecho contradice la evidencia de las Figuras 1 y 2 (se debe notar que la *oportunidad* es un funcional de la distribución del retraso, aproximadamente la probabilidad de que el retraso es menor o igual a 7 días). Sin embargo, puede ser útil como una primera aproximación. Otra limitación es que los intervalos de credibilidad pueden ser aún muy grandes como para ser una cantidad útil. Típicamente se requiere una base de datos importante para hacer nowcasting precisos pero lamentablemente, sólo existe información de inicio de síntomas para casos reportados desde el 29 de marzo de 2020.

3.2 Trabajo futuro

Nuestro objetivo principal es extender la aplicación del método de nowcasting a nivel comunal. Los resultados de las Figuras 1 y 2 son prometedores en tanto que muestran que efectivamente hay una señal robusta y heterogénea espacio-temporalmente respecto a la diferencia entre el inicio de síntomas y el reporte. La mayor limitación es la granularidad de los datos: a nivel comunal el inicio de síntomas se encuentra sólo disponible por semana epidemiológica, lo que nos obliga a artificialmente introducir la resolución de días mediante un muestreo artificial, el que inevitablemente introducirá más incertidumbre en las predicciones. En ese sentido, disponer de datos de inicio de síntomas desagregados por día y por comuna sería de extrema utilidad, así como también lo sería tener dicha información a nivel nacional, y no sólo a través de los histograms que fueron traducidos mediante visión computacional, introduciendo quizás aún más error.

Dos fuentes adicionales de informacións serían muy útiles para robustecer este análisis. En primer lugar, la cantidad de fallecidos cada día por comuna permitiría estimar la cantidad de casos totales de manera más fidedigna (Russell et al., 2020; Lu et al., 2020) e incluir explícitamente el subreporte como una componente de nuestro modelo de incidencia y *nowcasting*, como en Stoner et al. (2019). Adicionalmente, disponer de información fidedigna de testeo por comuna sería igualmente útil para tales fines.

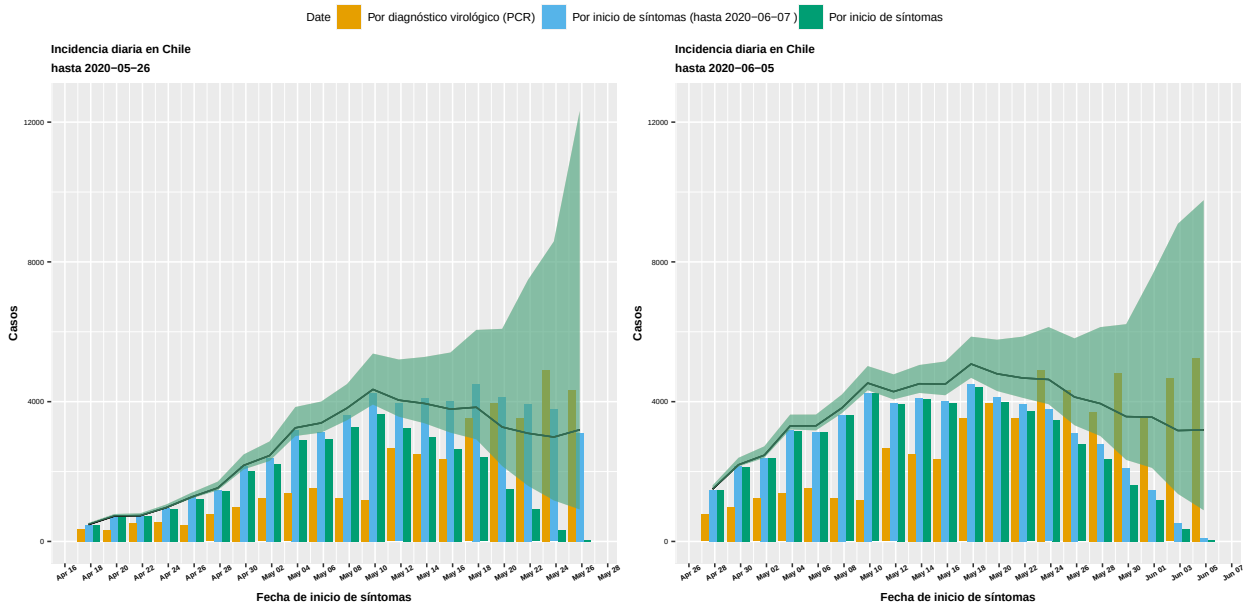

Figura 3: Resultados del método NobBS aplicados a los datos chilenos, ejecutados con datos hasta el 26 de mayo (**A**) y 03 de junio (**B**), con una resolución de dos días. Las barras verdes representan los síntomas cuyos casos han sido reportados en cada día hasta el día de la ejecución del algoritmo, las celestes representan la misma cantidad, pero usando los datos más recientes de los que se dispone registros (2020-06-04) y las barras beige representan la definición más usual de casos diarios, por su confirmación virológica. Las estimaciones puntuales de Y_t de NoBBS se muestran en la línea negra cuyas sombras grises representan intervalos de credibilidad del 95% de la distribución posterior de Y_t. En **A** se observa que las predicciones de NoBBS parecen ser precisas ya que los intervalos típicamente cubren a los valores que eventualmente se han realizado (celestes). En **B** es imposible decidir porque las barras verdes aún no terminan de realizarse. En cualquier caso, los intervalos de credibilidad tienen un tamaño razonable, lo que junto con la precisión sugieren la pertinencia de este análisis. Además, las estimaciones parecen consistentes con el posible efecto en la disminución (moderada) de casos debido a la cuarentena total en la RM en la segunda quincena de Mayo. Ciertamente la validación de esta hipótesis require una mayor investigación.

4 Métodos

4.1 Datos

Todos nuestros análisis se basan en reportes de fechas de inicios de síntomas para casos reportados. Usamos principalmente dos fuentes basadas en información oficial, detalladas a continuación.

4.1.1 Datos de alta resolución temporal (diaria) y baja resolución espacial (país)

Utilizamos datos de los informes epidemiológicos oficiales (Minsal, 2020b) que desde el 29 de Marzo incluyen un histograma de la distribución de la fecha de inicio de síntomas de todos los casos reportados hasta la elaboración del informe. En la mayoría de los casos éstos se reportan cada 3 o 4 días. Adicionalmente, entre el 13 y 27 de abril de 2020 estos gráficos aparecieron también con frecuencia diaria en los llamados 'informes ejecutivos situación epidemiológica' (Minsal, 2020a). Estos datos, disponibles hasta ahora en los informes del minsal en forma de histogramas, han sido transcritos sistemáticamente por Perez (2020) a formato numérico y se encuentran disponibles para uso público. La transcripción se realiza mediante un algoritmo de visión computacional que estima la altura de las barras del histograma, y está sujeto a errores. Para tomar en cuenta esos posibles esos errores tomamos la *mediana* de entre todos los informes de los casos cuyos síntomas comenzaron ese día.

Usando estos datos finalmente construimos una serie de datos pareados, donde cada caso tiene asignado una fecha de inicio de síntomas y fecha de reporte. Debido a que los reportes no se producen

diariamente, es necesario imputar la fecha real de reporte para todos los casos que se producen entre dos reportes consecutivos. En nuestra implementación actual, imputamos muestreando de una estimación simple de la distribución del retraso, usando los datos del periodo donde este informe se producía diariamente (entre el 12 y 28 de Abril). Finalmente, subsampleamos a intervalos de dos días usando el promedio como el representante de cada intervalo. Empíricamente vemos que este suavizado incrementa la precisión del *nowcasting* (ver 4.2), presumiblemente como consecuencia de que las estimaciones son sensibles a valores que no pueden ser calculados con fidedignamente a nivel diario, debido a la insuficiente periodicidad de los informes.

4.1.2 Datos de resolución baja resolución temporal (semanal) y alta resolución espacial (comuna)

Utilizamos también los datos de fecha de inicio de síntomas histórico (producto 15) en el repositorio del ministerio de ciencias. Estos corresponden a la transcripción de los informes de vigilancia epidemiológica, que desde el 13 de abril de 2020 muestran la cantidad de casos cuyo inicio de síntomas empezó en cada una de las semanas epidemiológicas, por comuna. Estos datos se reportan típicamente dos veces por semana (Lunes y Viernes). Usamos la convención de que el reporte con fecha r corresponde a la SE t si la mayor parte de los días entre los reportes $r-1$ y r estuvieron en la semana t. Debido a que los informes se publican Lunes y Viernes, esto se traduce de que cada semana epidemiológica es documentada en dos informes: el primero es el del Viernes y el segundo es el del Lunes de la semana epidemiológica siguiente. Nuestro análisis de *oportunidad* d se basa en este tipo de datos, y detalles se explican a continuación.

Oportunidad Definimos la *oportunidad* como la proporción de los casos reportados durante una SE que presentaron síntomas dentro de esa misma semana. Para las Figuras 1 y 2 consideramos sólo los informes del Lunes, debido a que los del Viernes típicamente continenen muy pocos casos cuyos síntomas iniciaron síntomas dentro de la misma semana del informe (de acuerdo a nuestra definición en el párrafo de arriba) y de esa manera debilitan la señal. Aún así, los efectos mostrados siguen estando presentes si se consideran los informes de los Viernes también. Además, en la Figura 1 excluimos las comunas con porcentage urbano menor a 90%, debido a que presentan patrones muy distintos al resto. Nuevamente, la inclusión o exclusión de estas comunas no influía mayormente en la validez de nuestros resultados.

4.2 Nowcasting

Denotamos por Y_t la cantidad de casos cuyos síntomas empezaron el día (o periodo) t. Debido a inevitables retrasos en el diagnóstico éstos son finalmente reportados de la siguiente manera

$$Y_t = \sum_{d=0}^{D} Y_{t,d}, \tag{1}$$

Donde $Y_{t,d}$ es la cantidad de casos cuyos síntomas iniciaron el día t y fueron reportados con un retraso de d días. D es el tiempo máximo que puede tardarse un reporte. El problema de *nowcasting* en epidemiología corresponde a estimar Y_t dados los reportes de casos hasta t y sus correspondientes inicio de síntoma. Estos no se conocen del todo, ya que en el día t sólo se conoce el triángulo $Y_{t',d}$ donde $t'+d \leq t$. Nuestra inferencia se basa en el reciente trabajo de McGough et al. (2020), originalmente aplicado en datos históricos de Dengue en Puerto Rico e Influenza en Estados Unidos. La implementación de este trabajo se hizo pública en el paquete `NobBS` *(Nowcasting by Bayesian Smoothing)* de `R`.

`NobBS` se basa en un modelamiento Bayesiano del mecanismo generativo de los datos $Y_{t,d}$. Específicamente, se asume que

$$Y_{t,d} \sim \text{NegativeBinomial}\left(r, p_{t,d}\right), \quad p_{t,d} = \frac{r}{r+\lambda_{t,d}}, \tag{2}$$

$$\lambda_{t,d} = \log(\alpha_t) + \beta_d, \tag{3}$$

$$\alpha_t \sim \mathcal{N}(\alpha_{t-1}, \sigma^2), \tag{4}$$

$$\beta_d \sim \text{Dirichlet}(\theta). \tag{5}$$

La ecuación (2) es la verosimilitud (*likelihood*), y conecta las observaciones con el mecanismo generativo. El uso de una distribución Binomial negativa es un paso más de complejidad sobre el supuesto de usual de que las observaciones tienen una distribución Poissoniana. Así, el parámetro r y permite representar la sobre o sub dispersión de los resultados. La ecuación (3) muestra el parámetro $\lambda_{t,d}$ que subyace a los valores observados, y que sigue una estructura de producto entre el mecanismo de incidencia y el de

retraso: específicamente, la incidencia real está representada por el parámetro α_t y se asume que tiene una estructura suave, lo cual se representa mediante la ecuación (4) que muestra que el *prior* es un paseo aleatorio, el cuál tiene un efecto de regularización en las fluctuaciones entre $t-1$ y t. Por otra parte, el parámetro β_d representa justamente la distribución del delay, para la que se asume un prior de *Dirichlet* (ecuación (5)). Así, la estructura de producto significa que los casos Y_t representados mediante la variable latente α_t se 'reparten' a través de β. La versión completa del modelo involucra otros hyper-parametros, los cuales no se muestran acá por el bien de la claridad.

Dado el anterior modelo generativo bayesiano, el paquete `NobBS` realiza inferencia de la distribución posterior de los parámetros (en particular, de α, β e Y_t) mediante Markov Chain Monte Carlo (MCMC) (Robert and Casella, 2013; Diaconis, 2009), implementada a través del lenguage de programación probabilsta BUGS/JAGS (Plummer et al., 2003). Disponer de esta distribución posterior permite fácilmente construir intervalos de confianza o credibilidad (como la Figura 3), hacer diagnósticos, análisis de sensibilidad, etc.

4.3 Código

Todos los análisis y código mostrados en este informe están disponbiles bajo petición al autor.

5 Agradecimientos

Agradecemos a Joaquin Fontbona (CMM-Dim-UChile) por discusiones que motivaron el estudio del retraso de reporte como un posible indicador espacio-temporal de saturación local de la capacidad de respuesta del sistema de salud a la pandemia. Tambien agradecemos a Jorge Pérez (DCC-IMFD-Uchile), Héctor Ramínrez (CMM-DIM-UChile), Alejandro Maass (CMM-DIM-Uchile), Raquel Jiménez (Boston University), Pamela Martinez (Harvard), Gonzalo Contador (Worcester Polytechnic Institute), Mauricio Vargas, Mauricio Santillana (Harvard), Fred Lu (Stanford), Oliver Stoner (Exeter), Theo Economou (Exeter) y Pablo Martínez (Harvard) por sus valiosos comentarios.

References

Bastos, L. S., Economou, T., Gomes, M. F., Villela, D. A., Coelho, F. C., Cruz, O. G., Stoner, O., Bailey, T., and Codeço, C. T. (2019). A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in medicine*, 38(22):4363–4377.

Casen, E. (2017). Observatorio social. *Previsión social: sintesis de resultados. Santiago de Chile: Ministerio de Desarrollo Social y Familia*.

Diaconis, P. (2009). The markov chain monte carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205.

Donnat, C. and Holmes, S. (2020). Modeling the heterogeneity in covid-19's reproductive number and its impact on predictive scenarios. *arXiv preprint arXiv:2004.05272*.

Holmdahl, I. and Buckee, C. (2020). Wrong but useful—what covid-19 epidemiologic models can and cannot tell us. *New England Journal of Medicine*.

Lipsitch, M. and Santillana, M. (2019). Enhancing situational awareness to prevent infectious disease outbreaks from becoming catastrophic. *Global Catastrophic Biological Risks*, pages 59–74.

Lu, F. S., Nguyen, A., Link, N., and Santillana, M. (2020). Estimating the prevalence of covid-19 in the united states: three complementary approaches.

McGough, S. F., Johansson, M. A., Lipsitch, M., and Menzies, N. A. (2020). Nowcasting by bayesian smoothing: A flexible, generalizable model for real-time epidemic tracking. *PLoS computational biology*, 16(4):e1007735.

Minsal (2020a). Informe ejecutivos de situación eepidemiológica del 1 de marzo al 27 de abril. *Departamento de Epidemiología, Ministerio de salud, Chile*. http://epi.minsal.cl/wp-content/uploads/2020/04/Informe_45_COVID_19_Chile.pdf.

Minsal (2020b). Informe epidemiológico n22, enfermedad por sars-cov2. *Departamento de Epidemiología, Ministerio de salud, Chile*. https://cdn.digital.gob.cl/public_files/Campa%C3%B1as/Corona-Virus/Reportes/Informe_EPI_010620.pdf.

Perez, J. (2020). Compilación de datos covid-19. *Comunicación personal*. https://docs.google.com/spreadsheets/d/1mLx2L8nMaRZu0Sy4lyFniDewl6jDcgnxB_d0lHG-boc/edit#gid=1170070241.

Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Russell, T. W., Hellewell, J., Abbott, S., Jarvis, C., van Zandvoort, K., nCov working group, C., Flasche, S., Kucharski, A., et al. (2020). Using a delay-adjusted case fatality ratio to estimate under-reporting. *Centre for Mathematical Modeling of Infectious Diseases Repository*.

Stoner, O. and Economou, T. (2019). Multivariate hierarchical frameworks for modeling delayed reporting in count data. *Biometrics*.

Stoner, O., Economou, T., and Drummond Marques da Silva, G. (2019). A hierarchical framework for correcting under-reporting in count data. *Journal of the American Statistical Association*, pages 1–17.

Zhao, Q., Ju, N., and Bacallado, S. (2020). Bets: The dangers of selection bias in early analyses of the coronavirus disease (covid-19) pandemic. *arXiv preprint arXiv:2004.07743*.