COVID-19: Metodología de corrección de las series según fecha de inicio de los síntomas

ICOVID Chile

Última actualización: 24 de agosto de 2020

Contenidos

Introducción

Introducción

- Durante una epidemia, contar con la información oportuna sobre el número de casos nuevos es crucial.
- Sin embargo, la toma de conocimiento de su existencia por parte de las autoridades sanitarias generalmente está sujeta a demoras debido al período de incubación, el tiempo de consulta a un especialista, el tiempo al diagnóstico y el tiempo del reporte del diagnóstico, entre otros.
- Esto resulta en sesgo negativo en el número de casos nuevos observados en la actualidad.
- Esto también genera que existan sesgos importantes en la estimación de indicadores epidemiológicos de interés basados en los datos observados, tal como el número de reproducción efectivo.

Introducción

- Existe una extensa literatura sobre la estimación de casos ocurridos y no reportados en diferentes contextos, incluyendo a la industria aseguradora y más recientemente en el modelamiento de datos de epidemias.

- Aquí consideramos una generalización del modelo propuesto por McGough et al. (2020. Nowcasting by Bayesian Smoothing: A flexible, generalizable model for real-time epidemic tracking. PLOS Computational Biology. https://doi.org/10.1371/journal.pcbi.1007735).

Introducción

- El modelo propuesto por McGough et al. (2020) fue también implementado por Gonzalo Mena a datos chilenos agrupados.
- A diferencia del modelo propuesto por McGough et al. (2020), el modelo considerado aquí permite que la distribución del rezago cambie como función del tiempo y se usa un proceso de mayor orden para modelar la intensidad de la infección.

Datos Faltantes

Imputación de datos faltantes

- Se asumió un mecanismo aleatorio de generación de datos faltantes para las variables de interés.

- Se utilizó el procedimiento de imputación múltiple para corregir las series (Rudin, D. 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc., New York, USA. ISBN:9780471087052).

- Se generan 5 bases de datos completas.

Modelo de Estimación de Casos Presentes

Antecedentes

- Modelo originalmente desarrollado para datos agrupados en base a reportes epidemiológicos.

- Desarrollado y mejorado en base a discusiones de los integrantes del subgrupo de Datos de ICOVID Chile:
  - Cuadrado, Cristobal (Universidad de Chile).
  - Engel, Eduardo (Universidad de Chile).
  - Jara, Alejandro (Pontificia Universidad Católica).
  - Marshall, Guillermo (Pontificia Universidad Católica).
  - Quintana, Fernando (Pontificia Universidad Católica).
  - Zubizarreta, José (Universidad de Harvard).

Notación

- $Y_{i,j}$ es la cantidad de casos que inician síntomas el i-ésimo día, informados con j días de rezago.

- α_i es la intensidad del proceso (cantidad esperada de casos que inician síntomas el i-ésimo día).

- $(\theta_{k,0}, \ldots, \theta_{k,30})$ es el vector de probabilidades de la distribución de rezago para la k-ésima semana epidemiológica (se).

Modelo

- La cantidad de casos que inician síntomas el i-ésimo día, e informados con j días de rezago, se modela a través de una distribución de Poisson:

$$Y_{i,j} \mid \alpha_i, \theta_j \overset{ind.}{\sim} \text{Poisson}\left(\mu_{i,j}\right), i = 1, \ldots, I, j = 0, \ldots, 30,$$

$$\mu_{i,j} = \exp\{\alpha_i\} \times \theta_{\text{se}_i,j}.$$

- Intensidad de la pandemia es modelada como un proceso AR(2) Gaussiano:

$$\alpha_1 \sim N\left(0, \sigma^2\left(\frac{(1-\phi_2)}{(1+\phi_2)\left[(1-\phi_2)^2-\phi_2^2\right]}\right)\right),$$

$$\alpha_2 \mid \alpha_1 \sim N\left(\frac{\phi_1}{1-\phi_2} \times \alpha_1, \frac{\sigma^2}{1-\phi_2^2}\right),$$

$$\alpha_i \mid \alpha_{i-1}, \alpha_{i-2} \sim N\left(\phi_1 \times \alpha_{i-1} + \phi_2 \times \alpha_{i-2}, \sigma^2\right).$$

Modelo

- La distribución del rezago para cada semana epidemiológica es modelado a través de un proceso AR(1) con distribuciones marginales Dirichlet:

$$(\theta_{1,0}, \ldots, \theta_{1,30}) \sim \text{Dirichlet}((1{,}0, \ldots, 1{,}0)),$$

$$\boldsymbol{\theta}_k = (\theta_{k,0}, \ldots, \theta_{k,30}) \mid \boldsymbol{\theta}_{k-1} \sim \text{Dirichlet}(5{,}0 \times (\theta_{k-1,0}, \ldots, \theta_{k-1,30})).$$

Modelo

- El modelo se completó con las siguientes distribuciones a priori para los hiper-parámetros:

$$\phi_2 \sim U(-1, 1),$$

$$\phi_1 \mid \phi_2 \sim U(l, u),\ u = |1 - \phi_2|,\ l = -u,$$

$$\sigma \sim U(0, 100).$$

Ajuste del Modelo

- El modelo se aplicó para casos confirmados sintomáticos y no-sintomáticos, ya que la fecha de inicio de los síntomas se considera un proxy de la fecha de infección (desfasada por el periodo de incubación).
- El modelo se aplicó para cada una de las 5 bases imputadas con respecto a los datos faltantes.
- El modelo se aplicó para las diferentes divisiones territoriales de forma independiente.
- Para cada modelo se creo una cadena de Markov de tamaño 70.000.
- Las primeras 20.000 muestras son descartadas.
- Un ejemplo de código BUGS para la implementación del modelo se muestra a continuación.

Código de BUGS (Parte 0)

```
# Notación:
# N es el número de registros.
# nd es el número total de días de rezago.
# maxp es el número de días de la epidemia.
# fe[1:N] es el vector que contiene la fecha
#           de inicio de los síntomas.
# period[1:N] es el vector que contiene la
#           semana epidemiológica en la que
#           se recibe la información.
# fe.d[1:N] es el vector que contiene los días
#           de rezago del registro.
```

Código de BUGS (Parte 1)

```
model
{
# likelihood
  for(i in 1:N)
  {
     log(mu[i]) <- alpha[fe[i]] +
                      log(theta[period[i], fe.d[i]])
     count[i] ~ dpois(mu[i])
  }
```

Código de BUGS (Parte 2)

```
# AR(2) Gaussian process for the intensity of the process
  psi[2] ~ dunif(-1, 1)
  u <- abs(1 - psi[2])
  l <- -u
  psi[1] ~ dunif(l, u)
  mu.ar[1] <- 0
  mu.ar[2] <- (psi[1]/(1-psi[2]))*alpha[1]
  tau1 <- ((1+psi[2])/(1-psi[2]))*((1-psi[2])*(1-psi[2]) -
                    psi[1]*psi[1])*tau
  tau2 <- tau*(1 - psi[2]*psi[2])
```

Código de BUGS (Parte 3)

```
  alpha[1] ~ dnorm(mu.ar[1], tau1)
  alpha[2] ~ dnorm(mu.ar[2], tau2)
  for(i in 3:maxp)
  {
      mu.ar[i] <- psi[1]*alpha[i-1] + psi[2]*alpha[i-2]
      alpha[i] ~ dnorm(mu.ar[i],tau)
  }

# Variance of the AR(2) process
  tau <- pow(sigma, -2)
  sigma ~ dunif(0, 100)
```

Código de BUGS (Parte 4)

```
  # AR(1) Dirichlet process for the distribution of the delay
    for(j in 1:nd)
    {
        theta[1, j] <- w[1, j]/sum(w[1, 1:nd])
        w[1, j] ~ dgamma(a0[j], 1.0)

        for(k in 2:maxs)
        {
            theta[k, j] <- w[k-1, j]/sum(w[k-1, 1:nd])
            w[k, j] ~ dgamma(a1*theta[k-1,j], 1.0)
        }
    }
}
```

Series Simuladas

Series Simuladas

- La estimación de la carga y el número de reproducción efectivo se lleva a cabo en base a series simuladas del número de casos.
- La series son simuladas desde la distribución a posteriori predictiva del número de casos.
- La series están disponibles en el GitHub de MINCIENCIA.
- Se generan 40 muestras desde las distribución a posteriori predictiva del modelo en base a cada una de las 5 bases de datos imputadas, totalizando 200 realizaciones de cada serie temporal.