Una comparación entre métodos econométricos y de Machine Learning para predecir la incidencia del COVID-19 en Perú

Marzo, 2020

Índice

- Objetivos
- Métodos
  - Modelo epidemiológico SIR
  - Aproximación econométrica: Controles Sintéticos
  - Métodos de Machine Learning
- Conclusiones

Índice

Objetivos

- Determinar si existen diferencias significativas entre las proyecciones de infectados utilizando métodos epidemiológicos, métodos econométricos tradicionales y métodos de Machine Learning.

- Diseñar un método que disminuya el error en las estimaciones y determinar parámetros de política como la tasa reproductiva básica (R_0) y la tása óptima de testeo de COVID-19.

Índice

Metodología

- El modelo formalmente se define como un sistema de ecuaciones diferenciales:

$$\frac{dS}{dt} = -\beta \times I \times S$$

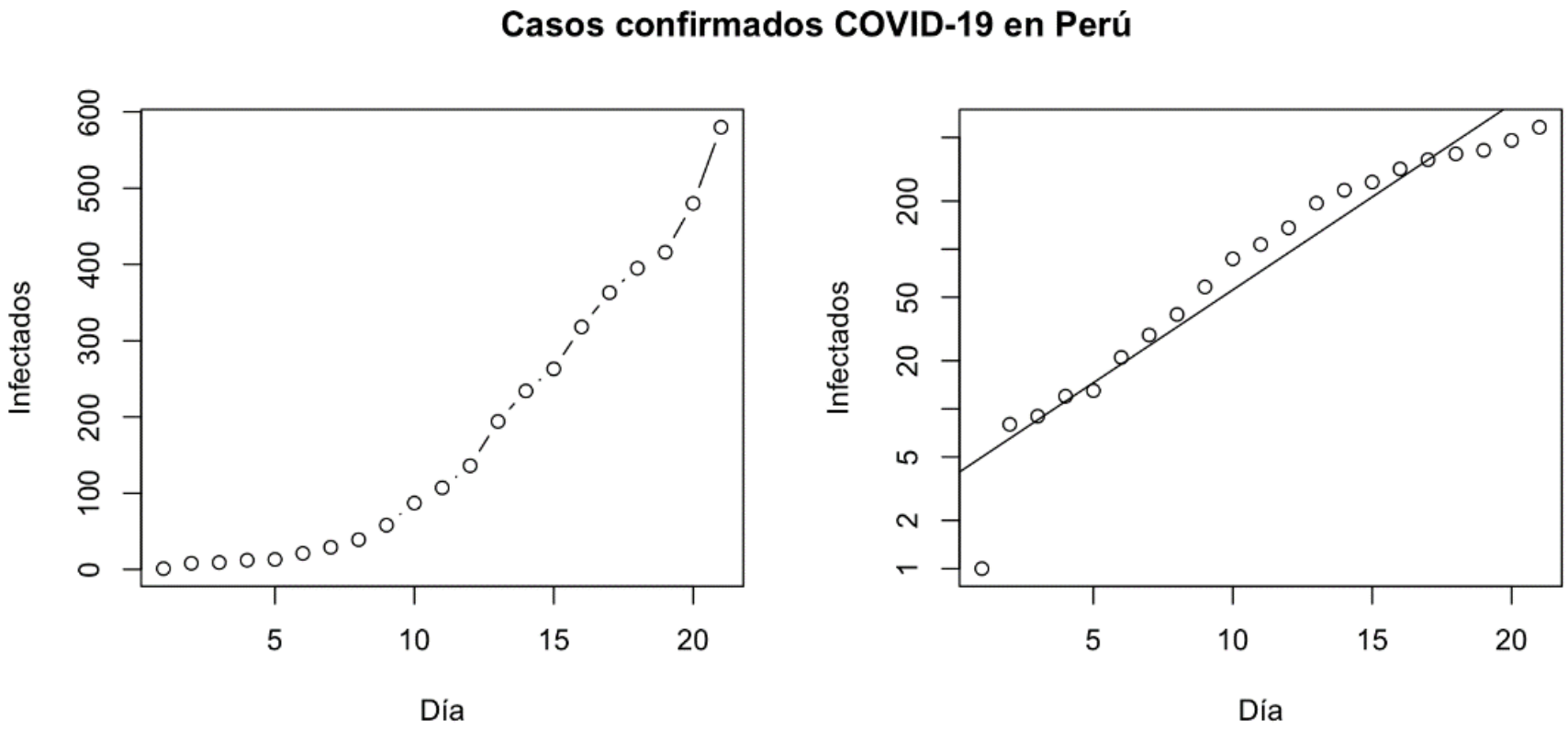
$$\frac{dI}{dt} = \beta \times I \times S - \gamma \times I$$

$$\frac{dR}{dt} = \gamma \times I$$

- El ratio de contagio básico (R_0) para este sistema se define como:

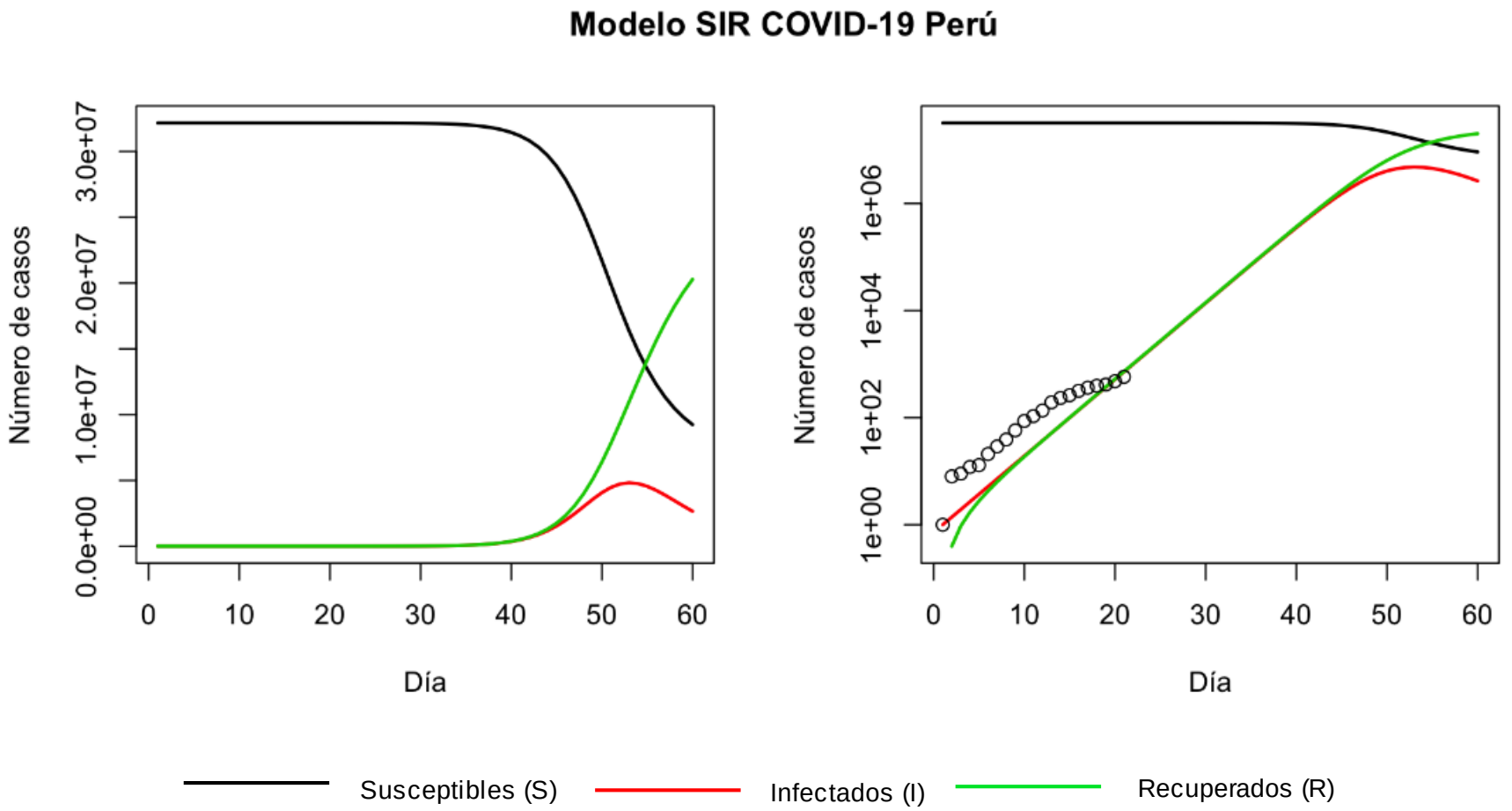
$$R_0 = \frac{\beta}{\gamma} N$$

Data



Fuente: Ministerio de Salud (2020)

Resultados


Modelo SIR COVID-19 Perú
Número de casos
0.0e+00
1.0e+07
2.0e+07
3.0e+07
0
10
20
30
40
50
60
Día
Número de casos
1e+00
1e+02
1e+04
1e+06
0
10
20
30
40
50
60
Día
Susceptibles (S)
Infectados (I)
Recuperados (R)

Resultados

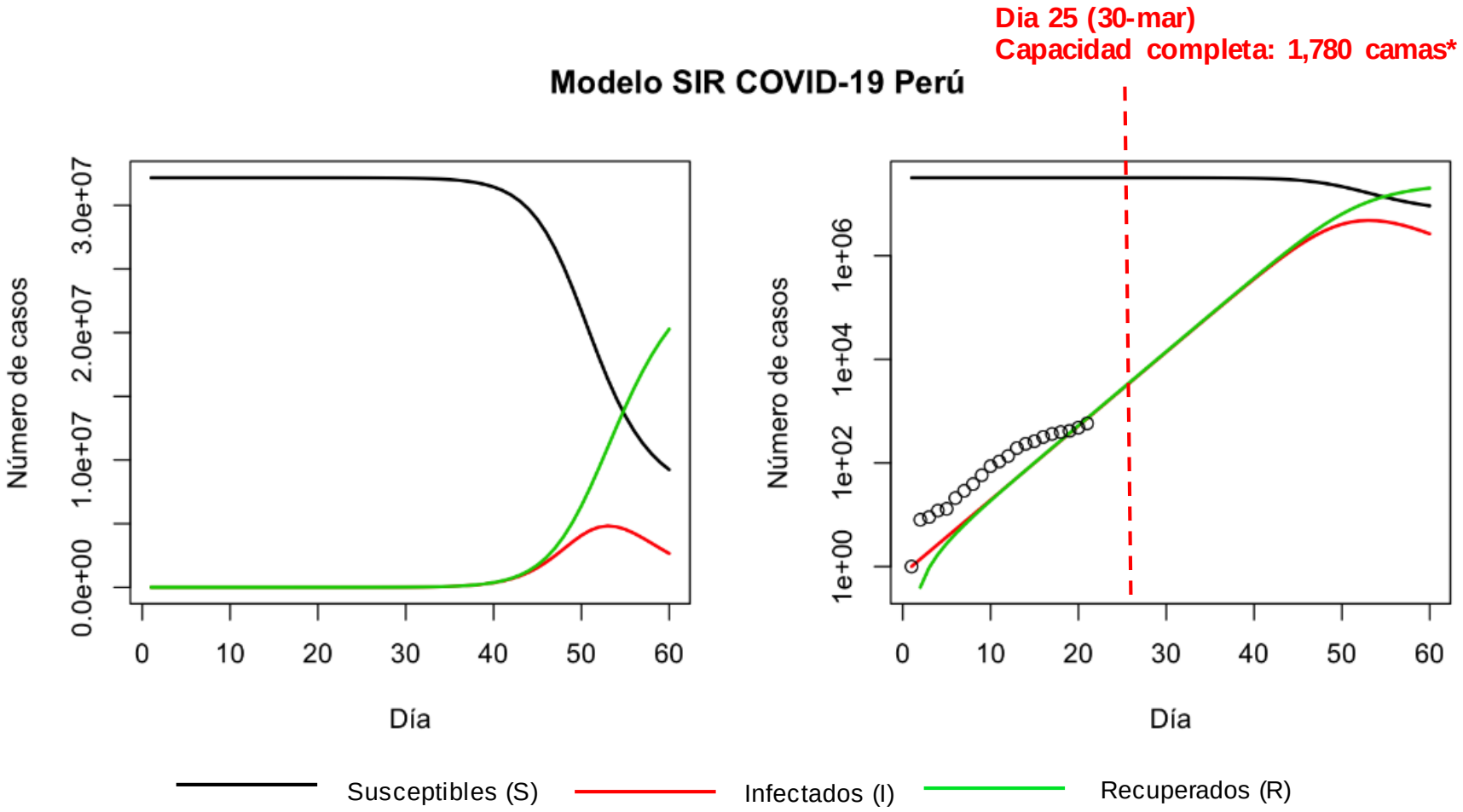
- Parámetros estimados (Ventana de estimación: 60 días)

$$\beta = 0.6644$$
$$\gamma = 0.3356$$
$$R_0 = 1.9796$$

- $R_0 > 1$, cada infección existente causa más de una infección nueva. La enfermedad se propagará entre las personas y puede haber un brote o una epidemia.

- Estimación de magnitud de la epidemia en Perú
  - Infectados: $4{,}821{,}133$
  - Muertes: $210{,}201$ (Supuesto: Tasa de mortalidad igual a 4.36%)

Resultados



*Capacidad instalada de camas para hospitalización: (i) Sector público 1,437: MINSA (830 CI y 250 UCI), EsSalud (502 CI, 152 UCI), (ii) Sector privado 343 UCI: 295 en Lima y 48 en regiones.

Índice

Metodología

- Siguiendo a por Abadie y Gardeazabal (2003) y Abadie et al. (2010, 2015), supongamos que observamos $J+1$ unidades en los periodos $1,2,\ldots,T$. La unidad 1 está expuesta a la intervención de interés durante los periodos $T_0+1,\ldots,T$. Las unidades J restantes son un conjunto controles potenciales que serán utilizados como grupo de comparación.

- Sea $w=\left(w_2,\ldots,w_{J+1}\right)^{'}$ un conjunto de ponderaciones, con $w_j \geq 0$ para $j=2,\ldots,J+1$ y $w_2+\ldots+w_{J+1}=1$. Cada valor de w representa un control sintético potencial. Como se puede apreciar las ponderaciones de los controles sintéticos están restringidas a ser no negativas y sumar 1, éstas se generan para minimizar la diferencia entre los resultados pre-intervención de la unidad tratada y el control sintético de la unidad

Metodología

- Sea X_1 un vector $(k \times 1)$ de características pre-intervención para la unidad tratada. De manera similar, dejemos que X_0 sea una matriz $(k \times J)$ que contiene las mismas variables para las unidades de comparación.

- El vector $w^* = \left(w_2^*, \ldots, w_{J+1}^*\right)'$ se elige para minimizar $\|X_1 - X_0 w\|$, sujeto a las restricciones ponderadas. El estimador de control sintético del efecto del tratamiento para la unidad tratada en un periodo post intervención $t \geq T_0$ es:

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$$

Metodología

- Una norma euclidiana ponderada se emplea comúnmente para medir la discrepancia entre características de la unidad tratada y del control sintético:

$$\|X_1 - X_0 w\| = \sqrt{(X_1 - X_0 w)' V (X_1 - X_0 w)}$$

donde V es una matriz diagonal con elementos no negativos en la diagonal principal que controlan la importancia relativa de obtener una buena coincidencia entre cada valor en X_1 y el correspondiente valor en $X_0 w^*$. Abadie et al. (2010) proponen un método inferencial para controles sintéticos basado en inferencia de aleatorización de Fisher.

Estrategia de identificación

- Se explotará información disponible de países donde se han registrado infectados de COVID-19. En particular, se tomarán países donde el virus tiene por lo menos 50 días.

- Los países seleccionados para la construcción del control sintético son: Australia, Canadá, China, Finlandia, Alemania, Italia, Japón, Corea del Sur, Malasia, España, Suecia, Tailandia, Estados Unidos y Reino Unido.

- Por otro lado, se utilizaron dos grupos de covariables: (i) variables relacionadas con el tamaño de la económica, características de la población y el sector salud, (ii) variables de política adoptadas para reducir la propagación del COVID-19.

Covariables utilizadas

Variables relacionadas con el tamaño de la económica, características de la población y el sector salud

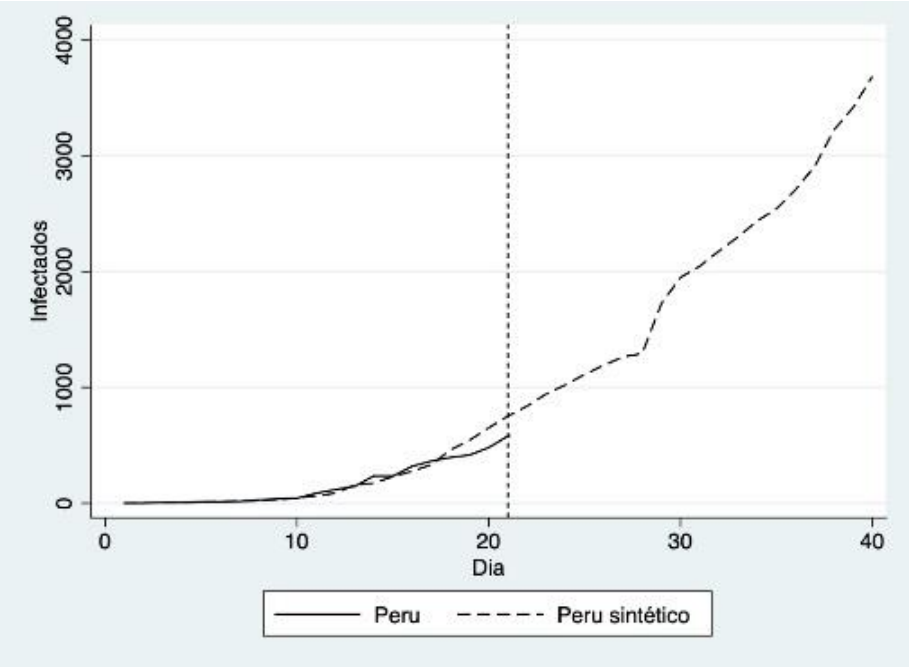
- PBI per cápita (ajustado por PPP)
- Deuda del gobierno central (% PBI)
- Inversión Extranjera Directa (%PBI)
- Población total
- Flujo neto de migración
- Esperanza de vida al nacer
- Número de muertes por enfermedades infecciosas (% del total)
- Número de camas de hospital (por cada 1,000 habitantes)
- Tasa de mortalidad infantil
- Población mayor de 65 años (% del total)

Variables relacionadas con el tamaño de la económica, características de la población y el sector salud

- Dummy temporal sobre política de aislamiento social
- Dummy temporal sobre política de declaratoria de estado de emergencia
- Dummy temporal sobre política de cierre de fronteras

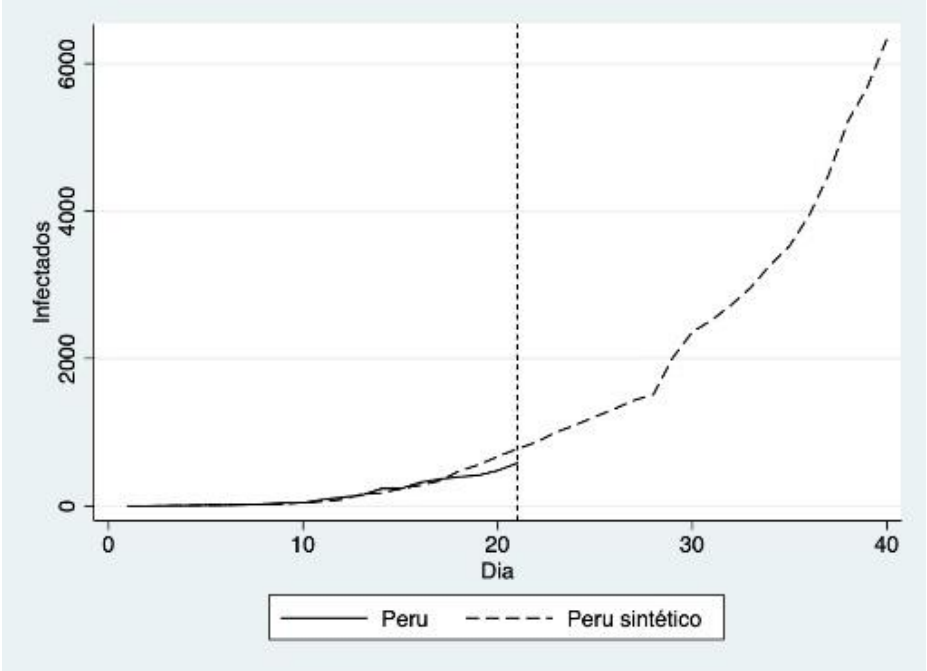
Resultados

Ventana de estimación: 40 días



- Infectados: 3,875
- Muertes: 169

Ventana de estimación: 50 días



- Infectados: 6,132
- Muertes: 267

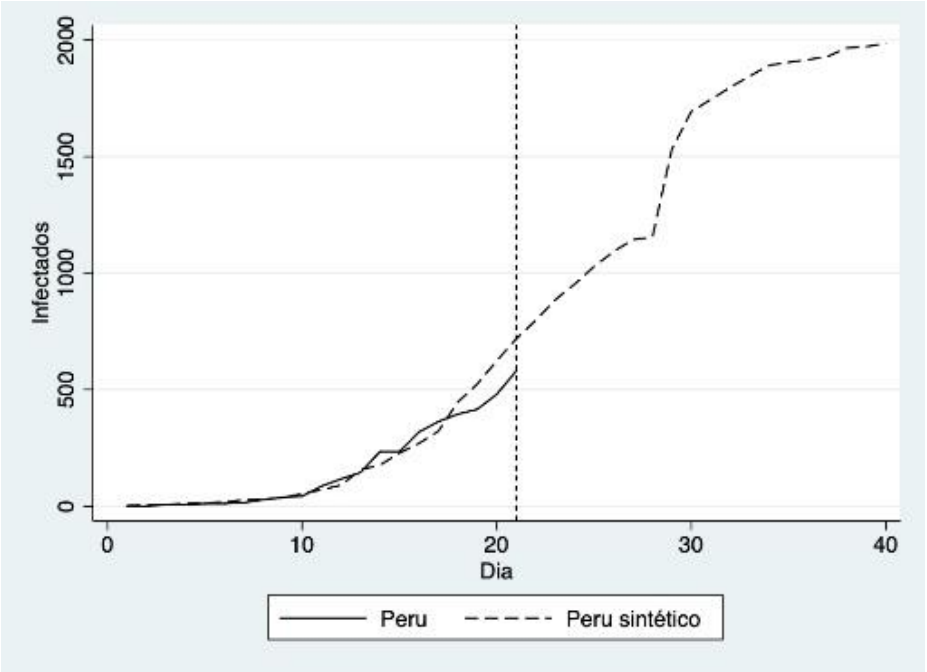
Segmentación de controles

- Sin embargo, el impacto de la adopción de medidas de política para controlar la propagación del COVID-19 depende directamente del espacio temporal en la cual se implementa.

- Por ello, se construirán dos sub-grupos de la siguiente forma:
    - **Grupo 1:** Países que adoptaron alguna de estas medidas por lo menos en los primeros 15 días posteriores a la detección del primer caso (Australia, Bélgica, Francia, Finlandia, Alemania, Japón, Corea del Sur, Suecia, Tailandia).
    - **Grupo 2:** Países que no adoptaron ninguna de estas medidas en los primeros 15 días posteriores a la detección del primer caso (China, Italia, España, Reino Unido, Estados Unidos).

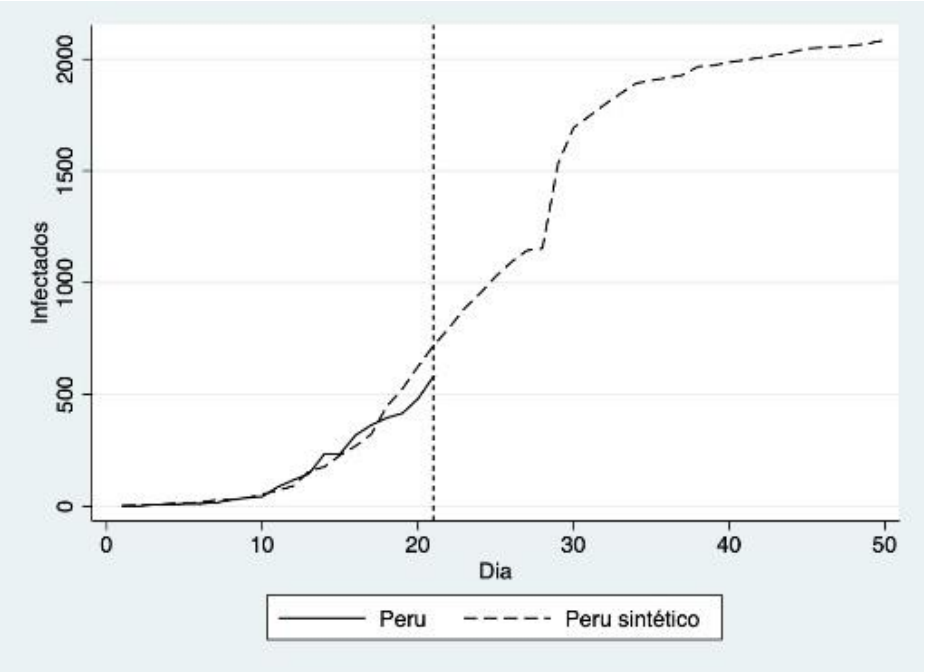
Resultados del Grupo 1

Ventana de estimación: 40 días



- Infectados: 1,989
- Muertes: 87

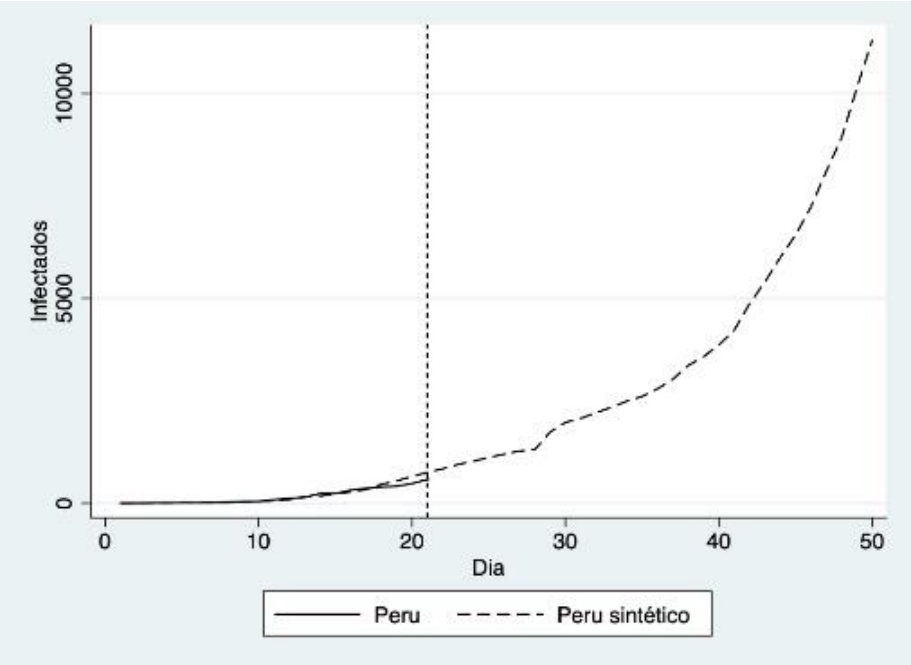
Ventana de estimación: 50 días



- Infectados: 2,076
- Muertes: 91

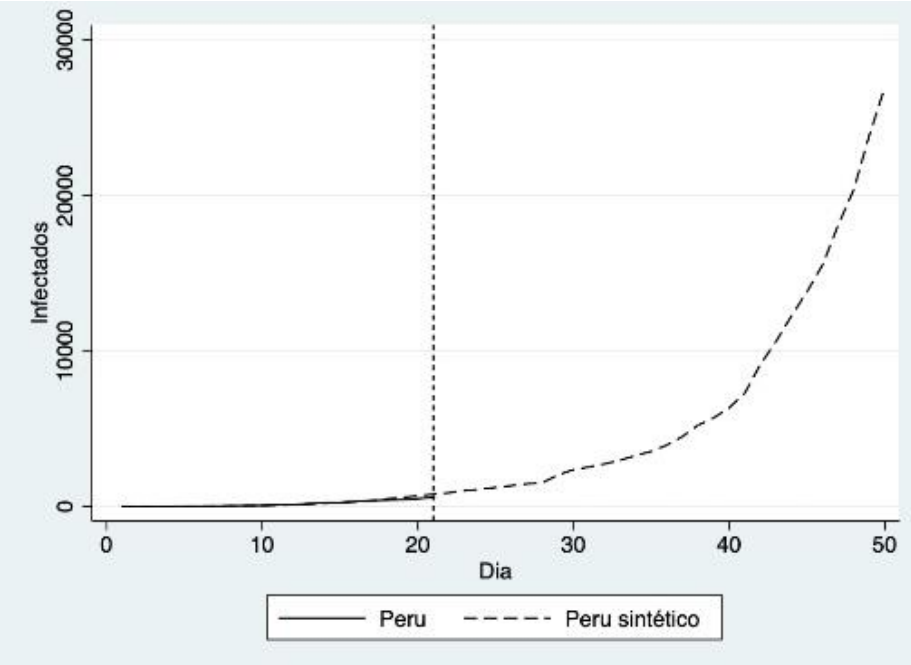
Resultados del Grupo 2

Ventana de estimación: 40 días



- Infectados: 11,728
- Muertes: 511

Ventana de estimación: 50 días



- Infectados: 27,381
- Muertes: 1,194