Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Causal Machine Learning para evaluación de efectos causales heterogéneos

Tesis de Licenciatura en Ciencias de Datos

Constanza de Galvagni

Director: Matías López-Rosenfeld

Codirector: Gustavo Landfried

Buenos Aires, 2025

CAUSAL MACHINE LEARNING PARA EVALUACIÓN DE EFECTOS CAUSALES HETEROGÉNEOS

Esta tesis realiza una comparación de modelos de aprendizaje automático para la estimación de efectos causales que varían entre individuos. La motivación central consistió en extender los análisis del influyente artículo de Jennifer Hill *"Bayesian Nonparametric Modeling for Causal Inference"* (2011) a un conjunto de modelos más modernos, que incluyen *Bayesian Additive Regression Trees* (BART), *Causal Forest* (CF) y *Bayesian Causal Forest* (BCF).

La evaluación de modelos se basó en el uso de *datasets* semi-sintéticos, construidos a partir de covariables reales del estudio *"Infant Health and Development Program"* (IHDP). Siguiendo el enfoque de Hill, se simularon los resultados para crear escenarios controlados donde los efectos causales verdaderos eran conocidos. Se diseñaron diferentes tipos de superficies de respuesta, con el objetivo de evaluar efectos causales tanto homogéneos como heterogéneos. El rendimiento de los modelos se midió utilizando métricas como el error absoluto para el efecto promedio (ATE), el error cuadrático medio del efecto causal individual (ITE) y la cobertura de los intervalos de confianza estimados para los efectos causales individuales.

En la estimación del efecto causal promedio (ATE), todos los modelos alcanzaron un desempeño satisfactorio, siendo BART el más destacado en todos los escenarios. En cambio, BCF mostró un rendimiento desigual, con buenos resultados en ciertas zonas del espacio de datos y un desempeño deficiente en otras, mientras que CF presentó un poder predictivo limitado y un error superior al de los demás métodos.

La tesis concluye enfatizando que la evaluación de modelos causales alternativos es un prerrequisito ineludible, tanto para la selección de variables de control previa a la estimación como para la toma de decisiones posterior. Se argumenta que la disciplina de la inferencia causal está actualmente limitada por el alto costo computacional de evaluar estas estructuras causales, y postula que su futuro depende del desarrollo de métodos eficientes que permitan ponderar la probabilidad de diferentes modelos a la luz de la evidencia empírica.

Palabras clave: Inferencia Causal, Efectos Causales Heterogéneos, Machine Learning Causal, Bayesian Additive Regression Trees, Causal Forest, Bayesian Causal Forest, Dataset IHDP.

AGRADECIMIENTOS

A Gustavo, mi codirector. Por darle forma a este proyecto y darme lugar a concretarlo. Por haber sido quien me acercó al mundo de la inferencia causal, sobre el que espero seguir aprendiendo. Por su tiempo, dedicación y esmero puesto en cada instancia de este trabajo. Por el acompañamiento no sólo académico sino también moral. A Matías, mi director, por permitir que esta tesis sea posible.

A Lucía Babino, por sus aportes, correcciones y sugerencias que contribuyeron enormemente a enriquecer y consolidar esta tesis.

A la Universidad Pública Argentina, por brindar formación de excelencia de forma gratuita, y por ser el lugar donde pude aprender, experimentar y crecer tanto humana como profesionalmente.

A mis papás, porque pude estudiar y dedicarme a lo que elegí gracias a su esfuerzo. Por enseñarme desde el ejemplo el valor del trabajo. Por criarme con amor y haberme hecho quien soy. Por esperarme cada noche en la parada cuando volvía de cursar y preguntarme cómo me fue, a pesar del cansancio. A mi hermano, por ser siempre el primero en creer en mí y enorgullecerse de cada paso, y por impulsarme desde el principio a estudiar esta carrera.

A mis abuelos, por inculcar la importancia de la educación a sus hijos y nietos, aunque ellos mismos hayan tenido que postergar sus estudios desde edades tempranísimas para poder subsistir. Por prenderme una velita antes de cada examen, y por regalarme con mucho esfuerzo mi primera computadora propia que me acompañó en cada cursada, exámen y trabajo práctico, incluyendo esta tesis. A mis tíos y padrinos, por el cariño y el apoyo, y por estar siempre orgullosos de mí. A Eliane, mi madrina, por acompañarme en mi crecimiento cumpliendo el rol de compañera, confidente, prima hermana, médica de cabecera y amiga.

A los amigos que me dio la facultad, especialmente a Alan, Anto, Ale, Belu, Max, Naty, Leo, Agus, Batta, Nico, y todos los amigos que me llevo de las cursadas, TPs y grupos de estudio, por hacer más ameno lo difícil de este camino, pero más que nada por todo lo compartido de forma extracurricular.

A los amigos que estuvieron siempre, en especial a Valen y Tute, por haber crecido juntos desde antes de que tengamos memoria y por seguir eligiéndonos a día de hoy. A Jaz, por ser mi compañera incondicional. A Fran y Ro, por compartir todo este proceso conmigo y por ayudarme a distraerme siempre que fue necesario.

A Lucas, a quien tengo la suerte de elegir como mi compañero cada día. Por inspirarme siempre a crecer y ser mejor. Por el amor, la escucha, por tener siempre las palabras justas. Porque no importó el cansancio ni la falta de tiempo de ambos sabiendo que al final del día ibas a estar vos.

Para Cacho y Tere, mis papás

Índice general

1. INTRODUCCIÓN

Todas las ciencias empíricas desarrollan teorías causales para explicar la naturaleza del mundo. Este proceso las conduce a formularse preguntas como: ¿Tiene un tratamiento el efecto esperado para curar una enfermedad? ¿Es efectiva una promoción para aumentar las ventas? Usualmente, estas preguntas dan lugar a otras, directamente relacionadas a la heterogeneidad de la población a estudiar: ¿El tratamiento mejora la condición de todos los pacientes por igual, o varía según ciertas características como edad o género? ¿Cuáles son las características de los clientes que se ven atraídos en mayor medida por la promoción?

Hubo que esperar hasta finales del siglo XX para que comenzara a desarrollarse un lenguaje estadístico específico para responder este tipo de preguntas, distinguiendo correctamente las correlaciones espurias de las relaciones estrictamente causales [1]. En los últimos años se han puesto a prueba una gran cantidad de algoritmos de aprendizaje automático en competencias de datos desarrolladas específicamente para evaluar el desempeño en la estimación de efectos causales heterogéneos, es decir, efectos que varían según las características de los individuos [2, 3].

Para razonar sobre el valor real de las variables que permanecen ocultas en los sistemas naturales, todas las ciencias empíricas, desde la física hasta las ciencias sociales, desarrollan argumentos (o modelos) causales mediante los cuales interpretan los datos observados, permitiéndoles usarlos como indicadores de las variables ocultas. Tanto los posibles valores de las variables ocultas como los argumentos o modelos causales alternativos son hipótesis que deben ser evaluadas en base a la evidencia mediante la aplicación estricta de las reglas de la probabilidad, las cuales componen el sistema de razonamiento para contextos de incertidumbre.

> El **problema fundamental** de todas las *ciencias basadas en datos* es conocer el valor real de las **variables** que permanecen **ocultas** en los sistemas naturales abiertos.

En esta tesis se propone comparar distintos enfoques para la estimación de efectos causales. Comprender en qué circunstancias cada método resulta más adecuado, y bajo qué condiciones sus estimaciones pueden ser confiables, es fundamental para avanzar en el uso riguroso y aplicado de la inferencia causal en dominios empíricos.

1.1. Sistema de razonamiento en contexto de incertidumbre

Las reglas de la probabilidad comenzaron a usarse a finales del siglo XVIII y desde entonces han sido adoptadas por todas las ciencias empíricas para razonar en contextos de incertidumbre. Aunque existen varios sistemas axiomáticos alternativos, desarrollados en el siglo XX, en todos los casos se llega a las mismas dos reglas.

Por un lado, la predicción *a priori* del próximo dato se realiza mediante la distribución de probabilidad marginal dada por la Ecuación 1.1, que calcula la probabilidad de la siguiente posible observación d con la contribución de todas las hipótesis mutuamente

contradictorias: la suma de la creencia previa conjunta entre el valor que se quiere predecir d y cada una de las hipótesis o universos paralelos, h.

$$P(D = d) = \sum_h P(H = h, D = d) \tag{1.1}$$

Por otro lado, la actualización de creencias sobre las hipótesis h dada la evidencia d se realiza mediante la distribución condicional dada por la Ecuación 1.2, a través de la cual se preserva la creencia previa conjunta sobre la hipótesis $H = h$ que sigue siendo compatible con el dato observado d, $P(H = h, D = d)$. El denominador representa la creencia total que se logró preservar, y funciona como normalizador para que la nueva creencia $P(H = h|D = d)$ siga integrando 1.

$$P(H = h|D = d) = \frac{P(H = h, D = d)}{P(D = d)}, \tag{1.2}$$

Por su parte, el renombrado Teorema de Bayes (1.3) no es más que una consecuencia directa de las reglas de la probabilidad ya mencionadas, y es simplemente una forma alternativa de expresar la distribución condicional con la cual se actualizan las creencias. En el Teorema de Bayes la creencia previa conjunta $P(H = d, D = d)$ se descompone como el producto entre la predicción a priori que hace la hipótesis del dato observado y la creencia a priori de la hipótesis, lo que se formaliza en la Ecuación 1.3.

$$\underbrace{P(\text{Hipótesis} \mid \text{Datos})}_{\text{Posterior}} = \frac{\overbrace{P(\text{Datos} \mid \text{Hipótesis})}^{\text{Verosimilitud}}\overbrace{P(\text{Hipótesis})}^{\text{Prior}}}{\underbrace{P(\text{Datos})}_{\text{Evidencia}}}. \tag{1.3}$$

De esta manera, el conocimiento sobre el valor real de las *variables ocultas* se actualiza dinámicamente, sujeto a revisión continua a medida que se acumula nueva información.

1.2. Evaluación de hipótesis al interior de modelos causales

Para actualizar la creencia sobre las variables ocultas en base a la evidencia es fundamental contar con un modelo causal que permita interpretar los datos y razonar sobre los posibles valores de las variables ocultas. Es por esto que, en la práctica, la evaluación de hipótesis H a partir de datos D se realiza siempre mediante un modelo M que vincula las hipótesis ocultas con los datos observados. Aunque en la formulación básica del Teorema de Bayes (1.3) esta dependencia se deja implícita, es importante hacerla explícita pues es fundamental, especialmente para el área de inferencia causal. Una formulación completa del Teorema de Bayes, entonces, debe incorporar explícitamente el modelo en el condicional, tal como se formaliza en la Ecuación 1.4.

$$\underbrace{P(\text{Hipótesis}_i|\,\text{Datos, Modelo})}_{\text{Posterior}} = \frac{\overbrace{P(\text{Datos}\,|\,\text{Hipótesis}_i\text{, Modelo})}^{\text{Verosimilitud}}\overbrace{P(\text{Hipótesis}_i|\,\text{Modelo})}^{\text{Prior}}}{\underbrace{P(\text{Datos} \mid \text{Modelo})}_{\text{Evidencia}}} \tag{1.4}$$

La formulación 1.4 deja en claro que la inferencia depende del modelo. Si bien en aprendizaje automático no se proponen modelos causales, los modelos también inducen sesgo en el resultado justamente porque permiten resolver los problemas inversos $P(\text{Hipótesis} \mid \text{Datos})$ a través de lo que se conoce como *inductive bias* (sesgo inductivo). Incluso los modelos más flexibles, como las redes neuronales profundas, introducen sesgos inductivos relevantes. Por ejemplo, las redes convolucionales fueron exitosas en visión gracias a que introducen una estructura local y una invariancia ante traslaciones que las hace particularmente eficaces en el procesamiento de imágenes [4].

En las ciencias empíricas estos sesgos inductivos suelen adoptar la forma de argumentos causales, es decir, procesos generativos que funcionan como hipótesis de la realidad causal subyacente que genera los datos observados. Estas hipótesis se especifican matemáticamente mediante relaciones causales probabilísticas (distribuciones condicionales entre causas y efectos), que pueden representarse mediante grafos dirigidos acíclicos (DAGs).

Dado que los argumentos causales son hipótesis sobre los procesos generativos ocultos, ¿cómo se pueden evaluar los argumentos causales alternativos?

1.3. Evaluación de modelos causales alternativos

Los modelos o argumentos causales alternativos constituyen hipótesis de alto nivel en las que las variables no observadas permanecen implícitas dentro de su formulación, y por lo tanto pueden ser evaluados en base a la evidencia mediante las reglas de la probabilidad, es decir, a través del ya mencionado sistema de razonamiento en contextos de incertidumbre. Dado que los modelos son hipótesis, es posible usar el Teorema de Bayes para evaluar los argumentos causales alternativos, tal como se explicita en la Ecuación 1.5.

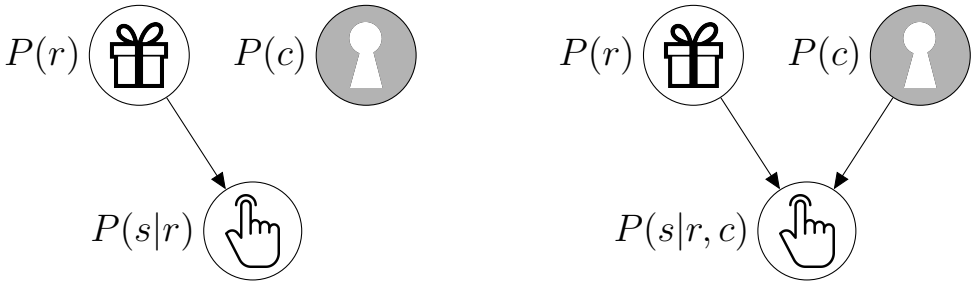
$$\underbrace{P(\text{Modelo}_i \mid \text{Datos})}_{\substack{\text{Posterior}\\ \text{de modelos}}} = \frac{\overbrace{P(\text{Datos} \mid \text{Modelo}_i)}^{\text{Evidencia}} \overbrace{P(\text{Modelo}_i)}^{\substack{\text{Prior de}\\ \text{modelos}}}}{\underbrace{P(\text{Datos})}_{\substack{\text{Predicción con la contribución}\\ \text{de todos los modelos}}}}. \tag{1.5}$$

Es importante notar cómo en el caso de la Ecuación 1.5 ya no se está considerando cuál es la probabilidad de una causa dada una consecuencia, sino qué narrativa causal es más plausible dada la evidencia observada.

Para ejemplificar se revisa un caso sencillo: Un regalo está oculto detrás de una de tres cajas, representada por $r \in \{1, 2, 3\}$. Una persona selecciona una de las cajas, $c \in \{1, 2, 3\}$, y luego otra persona, que conoce la ubicación del regalo, abrirá una de las cajas que **no** lo contiene, $s \in \{1, 2, 3\}$, cumpliendo que $s \neq r$.

Esta información define un modelo causal generativo que se denomina Modelo "Base" (M_0). Si además se impone la restricción adicional de que la caja que se abre no puede coincidir con la caja elegida, es decir, $s \neq c$, se obtiene un modelo causal alternativo, denominado Modelo "Monty Hall" (M_1). Ambos modelos generativos se representan matemáticamente en la Figura 1.1, utilizando la notación de redes bayesianas [5]. Al especificar las distribuciones de probabilidad condicional $P(r)$, $P(c)$ y $P(s \mid r, c)$, se define implícitamente la distribución de probabilidad conjunta, que es suficiente para razonar en contextos de incertidumbre. Las redes bayesianas, que emplean estas distribuciones condicionales para

representar mecanismos entre causas y efectos, poseen propiedades fundamentales para la inferencia causal, como la modularidad y el flujo de asociación, que se analizarán a lo largo de la tesis.



$s \neq r$ | $s \neq r,\ s \neq c$

a Modelo Base (M_0) | *b* Modelo Monty Hall (M_1)

Fig. 1.1: Especificación de modelos causales alternativos (restricciones a priori)

Con el objetivo de simplificar el análisis, se considera que la caja elegida es $c = 1$. A partir del producto de las probabilidades condicionales —definidas bajo el criterio de máxima incertidumbre dado el conocimiento disponible [6]— se obtienen las distribuciones de probabilidad conjunta de la Tabla 1.1.

$P(r, s|c = 1, M_0)$

	r_1	r_2	r_3
s_1	0	1/6	1/6
s_2	1/6	0	1/6
s_3	1/6	1/6	0

$P(r, s|c = 1, M_1)$

	r_1	r_2	r_3
s_1	0	0	0
s_2	1/6	0	1/3
s_3	1/6	1/3	0

Tab. 1.1: Probabilidad conjunta a priori, máxima incertidumbre dada la restricciones de los modelos causales, en el caso de que la caja elegida sea $c = 1$.

Para actualizar creencias a medida que llega nueva información simplemente se preserva la creencia previa que sigue siendo compatible con el nuevo dato. Por ejemplo, si la persona que conoce la posición del regalo abre la caja $s = 2$, simplemente se debe preservar la creencia conjunta a priori que sigue siendo compatible con el dato (el renglón $s = 2$, ver Figura 1.2), y re-normalizarla para que sume 1 (que represente el 100 % del espacio de probabilidad), que es exactamente lo que realiza el Teorema de Bayes.

a $P(r|s = 2, c = 1, M_0)$ | *b* $P(r|s = 2, c = 1, M_1)$

Fig. 1.2: Distribución de probabilidad a posterior dado el dato $s = 2$ y el modelo.

Diferentes modelos conducen a distintas conclusiones, por lo que vale preguntarse: ¿Cuál es la conclusión correcta? ¿Cuál es el modelo causal correcto? Para evaluarlos se computa

la distribución a posteriori de los modelos en un conjunto de datos. Si el conjunto de datos se obtuvo mediante el modelo generativo Monty Hall, el resultado será que el posterior de los modelos alcanza una certidumbre relativa en tan solo una veintena de episodios, como se muestra en la Figura 1.3.

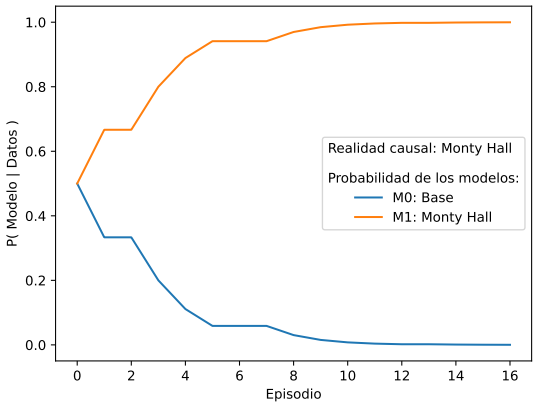


Fig. 1.3: Posterior de los modelos Base y Monty Hall en el tiempo en base a datos generados con el modelo Monty Hall.

En este contexto, ¿Es posible proponer otro modelo que tenga mejor desempeño que el modelo causal que se corresponde con la realidad causal subyacente? La respuesta es no [7]. El elemento que actualiza la creencia sobre los modelos es la predicción a priori que el modelo hace de los datos observados, $P(\text{Datos}|\text{Modelo})$, o evidencia.

La entropía cruzada (1.6) es una métrica que consiste en una transformación monótona de la evidencia: el negativo de su media geométrica medido en escala logarítmica. Maximizar la tasa de predicción (o media geométrica) coincide con minimizar la entropía cruzada, y esto ocurre cuando los modelos causales se corresponden con la realidad causal subyacente.

$$\underbrace{\mathcal{H}(R, M) = \sum_{c,s,r} \underbrace{P(c, s, r|\text{Realidad Causal})}_{\substack{\text{Probabilidad de que} \\ \text{se genere el dato}}} \cdot \underbrace{(-\log P(c, s, r|\text{Modelo Causal}))}_{\substack{\text{Información en} \\ \text{órdenes de magnitud}}}}_{\substack{\text{Entropía cruzada} \\ \text{Tasa de información del sistema de comunicación}}} \tag{1.6}$$

La ventaja de los modelos causales radica justamente en su capacidad predictiva, que se adapta a los diferentes contextos como intervenciones externas que modifican mecanismos causales específicos, aquí representados como distribuciones de probabilidad condicional.

1.4. Objetivos

La motivación central de esta tesis ha sido:

- Estudiar el área de inferencia causal en general;
- Caracterizar el estado actual de la disciplina;
- Implementar modelos del estado del arte;

- Discutir las limitaciones actuales de la inferencia causal.

En particular, el objetivo de implementación propuesto fue extender los análisis del artículo *Bayesian Nonparametric Modeling for Causal Inference* a un grupo de modelos que son considerados actualmente estado del arte para la estimación de efectos causales, poniendo a prueba escenarios desde más simples a más complejos con distintos tipos de efecto causal, comparando el desempeño predictivo en contextos con efectos causales homogéneos y heterogéneos.

2. INFERENCIA CAUSAL

2.1. La paradoja de Simpson y su solución

Para estimar el efecto causal de datos observados sin intervenciones es necesario interpretar los datos mediante un modelo causal. Se ejemplifica con un caso donde se tienen 3 variables:

- Estado inicial: $E_0 \in \{\text{Leve (0), Severo (1)}\}$
- Tratamiento: $Z \in \{\text{Sin tratamiento (0), Tratado (1)}\}$
- Estado final: $E_1 \in \{\text{Leve (0), Severo (1)}\}$

A partir de los datos recolectados se pretende responder si el tratamiento es efectivo para mejorar el estado del paciente. Para ello, se propone estimar las probabilidades presentes en la Tabla 2.1. En la última fila se almacena la probabilidad de terminar en estado severo E_1 dado el estado inicial E_0, diferenciando entre grupo de tratamiento y de control. En las primeras dos columnas se condiciona por el estado inicial del paciente, mientras que en la última columna se condiciona únicamente por grupo de tratamiento o de control.

Tab. 2.1: Probabilidades conjuntas estimadas para el experimento

	$E_0 = 0$	$E_0 = 1$	
$Z = 0$	$P(E_1 = 1 \mid Z = 0, E_0 = 0)$	$P(E_1 = 1 \mid Z = 0, E_0 = 1)$	$P(E_1 = 1 \mid Z = 0)$
$Z = 1$	$P(E_1 = 1 \mid Z = 1, E_0 = 0)$	$P(E_1 = 1 \mid Z = 1, E_0 = 1)$	$P(E_1 = 1 \mid Z = 1)$
	$P(E_1{=}1 \mid \boldsymbol{Z{=}1}, E_0{=}0)$ $-P(E_1{=}1 \mid \boldsymbol{Z{=}0}, E_0{=}0)$	$P(E_1{=}1 \mid \boldsymbol{Z{=}1}, E_0{=}1)$ $-P(E_1{=}1 \mid \boldsymbol{Z{=}0}, E_0{=}1)$	$P(E_1{=}1 \mid \boldsymbol{Z{=}1})$ $-P(E_1{=}1 \mid \boldsymbol{Z{=}0})$

En la Tabla 2.2 se exhiben los datos recolectados para este experimento.

Tab. 2.2: Simulación de datos para el experimento de ejemplo

	$E_0 = 0$	$E_0 = 1$	
$Z = 0$	15 % 210/1400	30 % 30/100	16 % 240/1500
$Z = 1$	10 % 5/50	20 % 100/500	19 % 105/550
	−5 %	−10 %	+4 %

Notar que al condicionar por el estado inicial, tanto para estado inicial leve como severo, la probabilidad estimada de terminar en estado severo se reduce si se aplica el tratamiento, tal como se calcula en las Ecuaciones 2.1 y 2.2. Este resultado sugiere que en ambos casos

el tratamiento es efectivo para mejorar el estado de los pacientes.

$$\hat{P}(E_1 = 1|\boldsymbol{Z} = \mathbf{1}, E_0 = 0) - \hat{P}(E_1 = 1|\boldsymbol{Z} = \mathbf{0}, E_0 = 0) = -0{,}05 \tag{2.1}$$

$$\hat{P}(E_1 = 1|\boldsymbol{Z} = \mathbf{1}, E_0 = 1) - \hat{P}(E_1 = 1|\boldsymbol{Z} = \mathbf{0}, E_0 = 1) = -0{,}10 \tag{2.2}$$

Sin embargo, al evaluar la probabilidad del estado final sin condicionar sobre el estado inicial (2.3), se observa que el tratamiento aumenta la proporción de personas en estado final severo.

$$\hat{P}(E_1 = 1|\boldsymbol{Z} = \mathbf{1}) - \hat{P}(E_1 = 1|\boldsymbol{Z} = \mathbf{0}) = 0{,}04 \tag{2.3}$$

Esta inversión del efecto se conoce como paradoja de Simpson. ¿Cuál es el efecto causal del tratamiento? ¿El tratamiento es o no es efectivo?

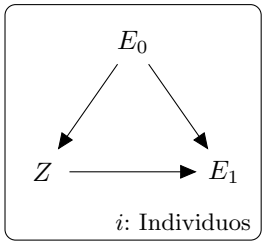
La respuesta emerge naturalmente cuando se propone un modelo causal que permita interpretar los datos observados. En este caso se propone una estructura causal en la que el tratamiento depende del estado inicial y el estado final depende del estado inicial y el tratamiento, lo que se representa en el grafo de la Figura 2.1. Con esta estructura se pueden reconstruir las distribuciones de probabilidad condicional. La población total es de 2050 personas, de las cuales 1450 tienen estado inicial leve y 600 tienen estado inicial severo. De las 1450 personas con estado inicial leve, solo a 50 se les aplica el tratamiento, mientras que de las 600 personas con estado inicial severo a 500 se les aplica el tratamiento. Con esta información, los porcentajes que se muestran en las tablas a la derecha de la Figura 2.1 representan la estimación de la probabilidad de los posibles *outcomes* para cada grupo.

$\hat{P}(E_0)$

$E_0 = 0$	$E_0 = 1$
1450/2050	600/2050

$\hat{P}(Z|E_0)$

	$Z = 0$	$Z = 1$
$E_0 = 0$	1400/1450	50/1450
$E_0 = 1$	100/600	500/600



$\hat{P}(E_1|Z, E_0)$

$(E_0 = 0)$	$E_1 = 0$	$E_1 = 1$
$Z = 0$	0,85	**0,15**
$Z = 1$	0,90	**0,10**
$(E_0 = 1)$	$E_1 = 0$	$E_1 = 1$
$Z = 0$	0,70	**0,30**
$Z = 1$	0,80	**0,20**

Fig. 2.1: Estructura causal subyacente propuesta para interpretar los datos observados, junto con las estimaciones de probabilidades condicionales inducidas por la estructura causal propuesta. La placa indica que esa misma estructura causal aplica para cada individuo de forma independiente.

Teniendo en cuenta los resultados de las tablas en la derecha de la Figura 2.1 y la estructura causal propuesta, se puede concluir que el tratamiento es efectivo para tratar la enfermedad: en el caso en que el estado inicial del individuo era leve (arriba) la probabilidad de desarrollar un estado final severo es un 5 % menor para el grupo tratado, mientras que para los individuos con estado inicial severo (abajo) la probabilidad de desarrollar estado final severo es un 10 % menor.

Proponer un modelo causal es fundamental para interpretar los efectos causales.

2.2. Intervenciones y efectos causales

Las intervenciones causales modifican la realidad causal subyacente, pues al asignar un valor determinado al tratamiento modifican el mecanismo causal con el que se genera naturalmente esa variable. En términos matemáticos, las intervenciones se expresan mediante el *do-operator* (2.4).

$$P(E_1|\text{do}(Z = z)) \tag{2.4}$$

El operador $\text{do}(Z = z)$ representa una intervención en el mecanismo causal generativo de la variable Z, es decir, en su distribución de probabilidad condicional. Una intervención que asigna a la variable Z el valor z, $\text{do}(Z = z)$, no es más que una modificación de la distribución de probabilidad condicional de esa variable por una función indicadora que asigna probabilidad 1 cuando $Z = z$ y 0 en caso contrario. Esto produce un modelo generativo intervenido, denominado M_z, en el cual la distribución de probabilidad condicional de la variable Z pasa a estar dado por la Ecuación 2.5.

$$P_{M_Z}(z) = \mathbb{I}(Z = z) \tag{2.5}$$

El resto de las distribuciones condicionales del modelo intervenido M_z permanecen iguales al modelo original, como se visualiza en la Figura 2.2. Esta propiedad de las redes causales se conoce como *modularidad*.

$$P_{M_Z}(Z) = \mathbb{I}(Z = z)$$
$$P_{M_Z}(E_0) = P(E_0)$$
$$P_{M_Z}(E_1|Z, E_0) = P(E_1|Z, E_0)$$

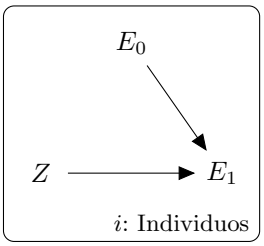


Fig. 2.2: Estructura causal intervenida para el ejemplo de la Tabla 2.1

Los experimentos aleatorizados son el método considerado más seguro para estimar efectos causales y son el estándar utilizado en la actualidad para la aprobación de nuevos medicamentos de la industria farmacéutica. En un experimento aleatorizado se remplazaría la distribución de probabilidad condicional original de Z por una distribución Bernoulli. Tanto en los experimentos aleatorizados como en las intervenciones deterministas como la descrita, resulta fundamental que la variable tratamiento pierda la dependencia con sus causas. Por lo tanto, la definición de efecto causal queda definida en la Ecuación 2.6, donde M_z representa el modelo causal modificado.

$$P(E_1|\text{do}(Z = z)) = P_{M_z}(E_1|Z = z) \tag{2.6}$$

En particular, el valor del efecto causal se computa como en la Ecuación 2.7.

$$P(E_1 = e_1|\text{do}(Z = z)) = \sum_{e_0} P_{M_z}(e_0, z, e_1) = \sum_{e_0} P(e_1|e_0, z)P(e_0). \tag{2.7}$$

2.3. Flujo de asociación en estructuras causales

Para estimar un efecto causal $P(E_1|\text{do}(Z))$ usando datos observados sin intervenciones es necesario eliminar las correlaciones espurias entre el tratamiento y la variable objetivo. Para saber cómo cortar la correlación espuria en general, se revisa primero el flujo de asociación en las tres estructuras elementales: el *fork*, el *pipe* y el *collider*, las cuales se definen a continuación.

2.3.1. *Fork*

La estructura básica del *fork* y su distribución de probabilidad conjunta se visualizan en la Figura 2.3, donde x, y y w son variables cualquiera.

$$x \longleftarrow w \longrightarrow y$$

$$P(x, w, y) = P(w)P(x|w)P(y|w)$$

Fig. 2.3: Fork

En el *fork*, las variables ubicadas en los extremos, x e y, se vuelven independientes al condicionar por la causa común. La expresión matemática de este comportamiento se encuentra en la Ecuación 2.8, donde se observa que w corta el flujo de asociación entre x e y.

$$P(x, y|w) = \frac{\cancel{P(w)}P(x|w)P(y|w)}{\cancel{P(w)}} = P(x|w)P(y|w) \tag{2.8}$$

Cuando no se condiciona por w, las variables no son independientes. Para demostrarlo es suficiente mostrar un contraejemplo: en la Figura 2.4 se definen las distribuciones de probabilidad condicional para cada variable y se generan datos aleatorios: w es una Bernoulli y las variables de los extremos son normales que están centradas en $4w$.

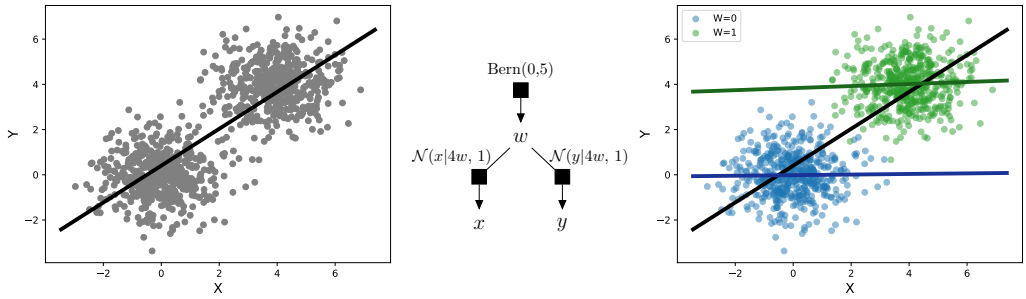


Fig. 2.4: Efecto causal de la estructura *Fork*: Al no condicionar por w existe una relación entre x e y, pero al hacerlo se vuelven independientes.

2.3.2. *Pipe*

La estructura del *pipe* y su distribución de probabilidad conjunta se visualizan en la Figura 2.5, donde x, y y w son variables cualquiera.

En los *pipe* los elementos de los extremos, x e y, también son independientes al condicionar por el mediador w. Esto se representa matemáticamente en la Ecuación 2.9, donde se

$$x \longrightarrow w \longrightarrow y$$
$$P(x, w, y) = P(x)P(w|x)P(y|w)$$

Fig. 2.5: Pipe

observa que el mediador w corta el flujo de asociación entre x e y.

$$P(x, y|w) = \frac{P(x)P(w|x)P(y|w)}{P(w)} = \frac{P(x)P(w|x)}{P(w)}P(y|w) = P(x|w)P(y|w) \tag{2.9}$$

Cuando no se condiciona por w, las variables no son independientes. Para demostrarlo se exhibe un contraejemplo: en la Figura 2.6 se definen las distribuciones de probabilidad condicional para cada variable y se generan datos: x es una Gaussiana, w una indicadora que se activa cuando $x > 0$ e y es una normal centrada en w.

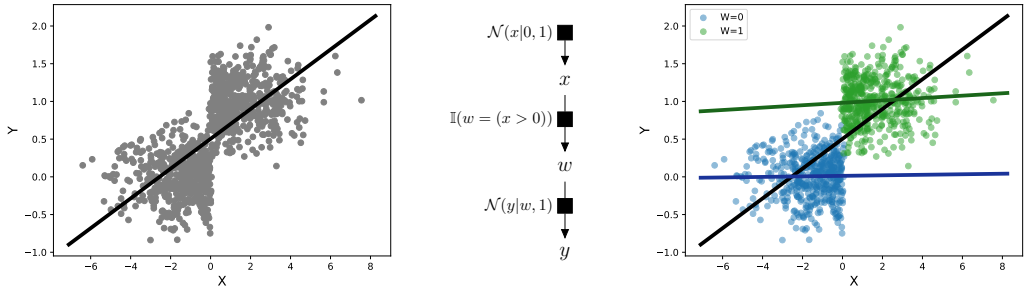


Fig. 2.6: Efecto causal de la estructura *pipe*: cuando no se condiciona por w existe una relación entre x e y, mientras que al hacerlo las variables de los extremos se vuelven independientes.

2.3.3. *Collider*

La estructura del *collider* y su distribución de probabilidad conjunta se visualizan en la Figura 2.7, donde x, y y w son variables cualquiera.

$$x \longrightarrow w \longleftarrow y$$
$$P(x, w, y) = P(x)P(y)P(w|x, y)$$

Fig. 2.7: Collider

A diferencia de los *fork* y los *pipe*, los elementos de los extremos en una estructura *collider* donde x e y son causas simultáneas de una variable w son independientes entre sí cuando **no** se condiciona por la consecuencia común. Esto se formaliza en la Ecuación 2.10, donde se muestra que al no contar con la información de la consecuencia común w no existe flujo de asociación entre x e y.

$$P(x, y) = \sum_w P(x)P(y)P(w|x, y) = P(x)P(y)\cancel{\sum_w P(w|x, y)} = P(x)P(y) \tag{2.10}$$

Sin embargo, al condicionar por ella se abre el flujo de asociación entre las causas comunes, haciéndolas no independientes. Para demostrarlo se exhibe un contraejemplo: en la

Figura 2.8 se definen las distribuciones de probabilidad condicional para cada variable y se generan datos: x e y son Gaussianas independientes y la consecuencia común w es una indicadora que se activa cuando $x + y > 0$.

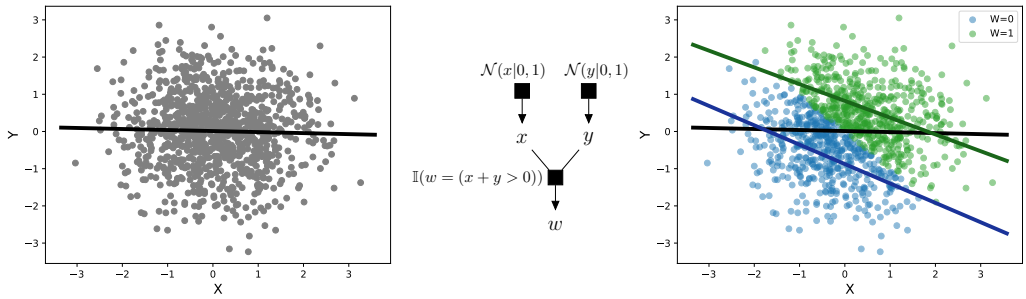


Fig. 2.8: A diferencia de lo que ocurre con los *fork* y pipe, en los *collider* cuando no se condiciona por w no existe una relación entre x e y, mientras que al hacerlo las variables de los extremos dejan de ser independientes. Ocurre lo mismo si se condiciona sobre alguna consecuencia de w.

2.3.4. *d-separation*

En la Tabla 2.3 se puede ver un resumen de las estructuras básicas en el flujo de asociación, donde el símbolo $\perp\!\!\!\perp$ denota independencia entre variables.

Tab. 2.3: Resumen de las estructuras básicas del flujo de asociación

		Intermedio no observable	Intermedio observable
Fork:	$x \leftarrow w \rightarrow y$	$x \not\perp\!\!\!\perp y$	$x \perp\!\!\!\perp y \mid w$
Pipe:	$x \rightarrow w \rightarrow y$	$x \not\perp\!\!\!\perp y$	$x \perp\!\!\!\perp y \mid w$
Collider:	$x \rightarrow w \leftarrow y$	$x \perp\!\!\!\perp y$	$x \not\perp\!\!\!\perp y \mid w$

El comportamiento de estas tres estructuras permite determinar cuándo los extremos de una cadena están asociados.

Definición 2.3.1 (***d-separation***)**.** *Hay asociación entre los extremos de una cadena (camino d-conectado) si se cumplen simultáneamente las siguientes condiciones.*

- *Todas las consecuencias comunes o* colliders *(o sus descendientes) son observables*
- *Ninguna otra variable es observable*

Y no hay asociación (camino "d-separado") cuando no se cumple alguna de las condiciones, es decir, cuando al menos un collider no es observado o al menos otra variable es observada.

2.3.5. Regla general

Finalmente, el criterio de d-separación permite determinar cuándo está abierto o cerrado el flujo de asociación entre dos variables en cualquier tipo de estructura causal. En la Figura 2.9 las variables Z e Y están conectados por varios caminos, los cuales pueden verse en detalle en la Figura 2.10.

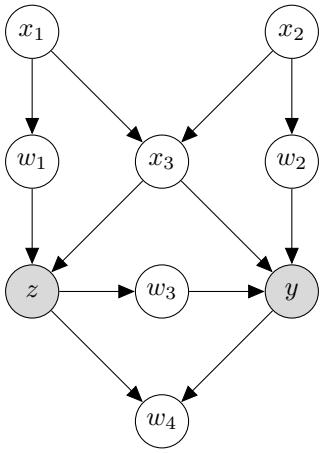


Fig. 2.9: Ejemplo de estructura causal compleja entre Z e Y

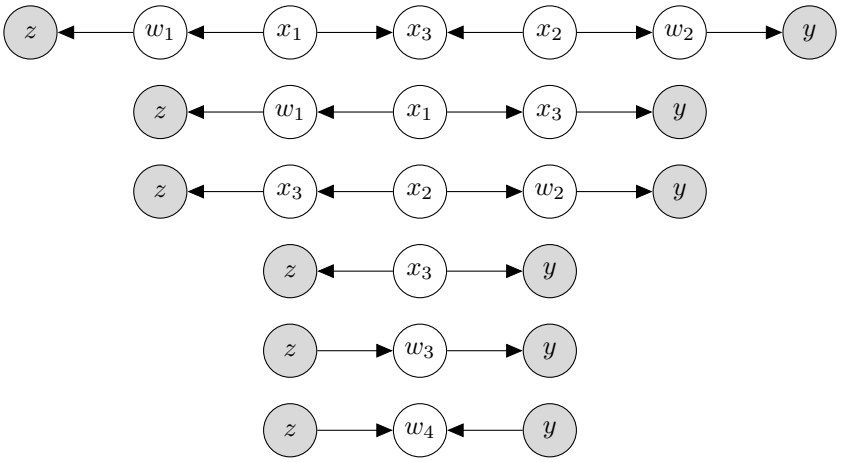


Fig. 2.10: Caminos que conectan al tratamiento z y al *outcome* y en el ejemplo de la Figura 2.9

Luego, Z e Y son independientes dado un conjunto de variables $\boldsymbol{Q}$ si y solo si $\boldsymbol{Q}$ d-separa a Z e Y en todos los caminos que conectan a Z e Y. Por ejemplo, el conjunto $\boldsymbol{Q} = \{X_3, W_2, W_3\}$ cierra todos los caminos entre Z e Y. Esto se formaliza en la Ecuación 2.11.

$$Y \perp\!\!\!\perp Z | X_3, W_2, W_3 \tag{2.11}$$

2.4. Identificación de efectos causales: *backdoor* y *adjustment formula*

Uno de los objetivos principales del área de inferencia causal es estimar efectos causales en datos observados en los que no se ejecutó ningún tipo de experimento aleatorizado. Intuitivamente, las asociaciones espurias entre las variables tratamiento y objetivo, Z e Y, circulan a través de todos los caminos ascendentes (no causales) que van de Z a Y. Por otro lado, la asociación no espuria entre Z e Y fluye por los caminos causales (de estructura pipe, descendente en todos los saltos) que conectan Z e Y. Esto motiva la definición del criterio *backdoor* para definir variables de control, de acuerdo a la definición dada por Judea Pearl en su libro *Causality* (2009) [5].

Definición 2.4.1 (***Backdoor criterion***)**.** *Un conjunto de variables* $\boldsymbol{Q}$ *satisface el backdoor criterion relativo a las variables* Y *y* Z *en un DAG si:*

- $\boldsymbol{Q}$ *cierra todos los caminos ascendentes de* Z *e* Y *(caminos no causales).*
- $\boldsymbol{Q}$ *no contiene ninguna variable descendente de* Z *a* Y *(caminos causales).*

Cuando se tiene un conjunto de variables de control que cumple con el criterio *backdoor* vale el Teorema 2.4.1, nuevamente tal como lo definió Pearl [5].

Teorema 2.4.1 (***Adjustment Formula***)**.** *Si un conjunto de variables* $\boldsymbol{Q}$ *cumple con el criterio backdoor entre* Z *e* Y*, entonces el efecto causal de* Z *en* Y *es identificable a través de la Fórmula 2.12.*

$$\begin{aligned} P(Y=y|do(Z=z)) &= P_{M_z}(Y=y|Z=z) \\ &= \sum_q P_{M_z}(Y=y, Q=q|Z=z) \\ &= \sum_q P_{M_z}(Q=q|Z=z) P_{M_z}(Y=y|Z=z, Q=q) \\ &\overset{*}{=} \sum_q \underbrace{P(Q=q)}_{\substack{\text{Peso del efecto} \\ \text{causal específico } q}} \underbrace{P(Y=y|Z=z, Q=q)}_{\substack{\text{Efecto causal} \\ \text{específico a } q}} \end{aligned} \tag{2.12}$$

La *adjustment formula* establece que el efecto causal de una intervención puede identificarse usando exclusivamente distribuciones de probabilidad sin intervenir. Dado que el conjunto de de control Q corta las asociaciones espurias, la distribución de probabilidad $P(Y = y|Z = z, Q = q)$ representa fielmente el efecto causal específico a Q, $P_{M_z}(Y = y|Z = z, Q = q)$. Para obtener el efecto causal marginal, $P(Y = y|\text{do}(Z = z))$ deben integrarse cada uno de esos efectos causales específicos a Q por la probabilidad de que ocurra Q.

La demostración está basada exclusivamente en criterios gráficos y reglas del *do-calculus* que se detallarán en el transcurso de la misma. No requiere mencionar variables contrafactuales como ocurre bajo el paradigma de *potential outcomes*, que se revisará más adelante.

Demostración 2.4.1. *Las primeras dos igualdades valen por definición. Únicamente es necesario demostrar la igualdad marcada con un asterisco (*$\overset{*}{=}$*). La demostración se puede descomponer en dos partes.*

1. $P_{M_z}(Q = q|Z = z) = P(Q = q)$.

   En el lado izquierdo la intervención elimina todas las flechas entrantes a Z *haciendo que la variable* Z *se convierta en un nodo raíz. La clave está en que esta la variable* Z *sólo afecta a sus descendientes pero no a su ancestros, porque todos los caminos que conectan a* Z *con* Q *en el grafo intervenido necesariamente tienen un collider no observado que la hacen independiente de* Q*. Esta independencia permite eliminar a la variable* Z *del condicional sin que se produzca ningún cambio en la probabilidad de* Q *en el grafo intervenido. Luego,* $P_{M_z}(Q = q|Z = z) = P_{M_z}(Q = q)$

   Por otro lado, es necesario probar que las marginales de las variables de control en el modelo intervenido y sin intervenir son iguales, $P_{M_z}(Q = q) = P(Q = q)$*. Por*

definición, ambas distribuciones pueden ser obtenidas marginalizando de sus respectivas distribuciones conjuntas. Además, la distribución conjunta de cualquier modelo siempre es el producto de todas sus distribuciones condicionales, y dado que las intervención sólo modifica las distribución condicional de la variable Z (el resto de las distribuciones de probabilidad condicional, para el resto de variables, son exactamente iguales en ambos modelos), allí radica la única diferencia entre las distribuciones conjuntas de los modelos. Sean Y, Z, Q, X todas las variable del modelo, donde X representa el resto de variables que no son Y, Z y Q:

$$\begin{aligned} P(Q) &= \sum_{Y,Z,X} P(Q,X,Z,Y) \\ &= \sum_{Y,Z,X} P(Q)P(X,Z,Y|Q) \\ &= P(Q) \sum_{Y,Z,X} P(X,Z,Y|Q) \\ &= P(Q) \sum_{Y,Z,X} P(X,Y|Q)P(Z|Q,X,Y) \\ &= P(Q) \sum_{Y,X} P(X,Y|Q) \underbrace{\sum_{Z} P(Z|Q,X,Y)}_{1} \end{aligned} \tag{2.13}$$

Cada uno de los pasos vale por definición. En el último paso se observa que la distribución de probabilidad condicional de Z no cumple ningún rol en la marginal de $P(Q)$ pues la integral sobre Z con todos las variables del condicional fijas siempre vale 1, tanto en el modelo intervenido como en el no intervenido.

Luego, queda probado el primer caso. El razonamiento aquí expresado es un caso particular de la regla 3 del do-calculus.

2. $P_{M_z}(Y = y|Z = z, Q = q) = P(Y = y|Z = z, Q = q)$

   En el lado izquierdo de la igualdad se está realizando una intervención que corta todas las flechas entrantes a Z. Esto elimina forzosamente cualquier posible flujo de información a través de causas comunes de Z e Y. Toda la información que fluye desde Z hacia Y lo hace exclusivamente por los caminos causales directos que conectan a Z con Y.

   En el lado derecho de la igualdad no se está interviniendo, y por lo tanto la variable Z recibe todas las flechas que la vinculan con sus causas naturales. A pesar de que ahora existan caminos backdoor, el flujo de asoaciación está bloqueado en todos ellos debido a que las variables de control Q cumplen el criterio backdoor (cierran flujo trasero).

   En conclusión, el flujo de asociación es el mismo en ambos lados de la igualdad, y por lo tanto es irrelevante si se interviene(lado izquierdo) o si no se interviene (lado derecho). Ambos lados son equivalentes. El razonamiento aquí expresado es un caso particular de la regla 2 del do-calculus.

Luego, vale el teorema 2.12.

Existe una expresión alternativa del *adjustment formula* 2.12 basada en lo que se conoce como *propensity score*, $P(z|q)$.

$$P(y \mid \text{do}(z)) = \sum_q \underbrace{\frac{P(z \mid q)}{P(z \mid q)}}_{1} \underbrace{P(y \mid z, q)P(q)}_{\text{Adjustment formula}} = \underbrace{\sum_q \frac{P(y, z, q)}{P(z \mid q)}}_{IPW} \tag{2.14}$$

El último elemento se conoce como *Inverse Probability Weighting* (IPW).

2.5. Enfoque basado en resultados potenciales contrafactuales

Hasta mediados de la década de 1990 no existía un criterio preciso que permitiera seleccionar correctamente las variables de control. La condición que se requería antes de la existencia del *backdoor criterion* estaba basada en una analogía directa con los experimentos aleatorizados. Por definición, en los experimentos la elección del tratamiento se realiza aleatoriamente, independientemente del resultado potencial que tendría en el individuo (Ecuación 2.15). Los resultados potenciales contrafactuales, para el tratamiento y para no tratamiento, se notan como Y_1 e Y_0, y z representa los posibles tratamientos contrafactuales.

$$Y_z \perp\!\!\!\perp Z \quad \forall z \in 0, 1 \tag{2.15}$$

Antes de la existencia del criterio *backdoor* se decía que un conjunto de variables de control Q era válido si y solo si ellas hacían que el tratamiento observado fuera independiente de los resultados potenciales contrafactuales (Ecuación 2.16).

$$Yz \perp\!\!\!\perp Z \mid Q \quad \forall z \in \{0, 1\} \tag{2.16}$$

Para verificar el cumplimiento de este criterio es suficiente mostrar que la distribución conjunta entre el tratamiento y los resultados potenciales contrafactuales pueden descomponerse como el producto de sus marginales, como se indica en la Ecuación 2.17.

$$P(Y_z, Z \mid Q) = P(Y_z \mid Q)\, P(Z \mid Q) \quad \forall z \in \{0, 1\} \tag{2.17}$$

Se supone, por el momento, que el criterio de *ignorability* resulta adecuado para la estimación de efectos causales; lo que será revisado más adelante. Uno de los problemas para encontrar un estimador del contrafactual $\mathbb{E}[Y_z]$ está relacionado con el hecho de que las variables contrafactuales Y_z son, por definición, no observables. Este problema tiene solución bajo el supuesto de "consistencia", según el cual se afirma que el valor de la variable observada es igual al valor del resultado potencial contrafacual que se corresponde con el tratamiento observado (2.18).

$$Z = z \implies Y = (1 - z)\, Y_0 + z\, Y_1 \tag{2.18}$$

De esta forma es posible suponer que la observación realizada es igual a su contrafactual. Estos dos supuestos, el de ignorabilidad y el de consistencia, permiten definir el estimando del contrafactual de la Expresión 2.19.

$$\begin{aligned} \mathbb{E}[Y_z] &= \mathbb{E}_Q[\mathbb{E}[Y_z|Q]] \\ &\overset{2.16}{=} \mathbb{E}_Q[\mathbb{E}[Y_z|Q, Z = z]] \\ &\overset{2.18}{=} \mathbb{E}[\mathbb{E}_Q[Y|Q, Z = z]] \end{aligned} \tag{2.19}$$

El estimando que se ha derivado en la Ecuación 2.19 no es más que el *adjustment formula* que se había derivado de forma general en la sección anterior (2.12).

Para cerrar la sección, se mencionan algunos de los estimandos más utilizados en la literatura de inferencia causal. Se define el ***Average Treatment Effect (ATE)***, definido como la esperanza de las diferencia entre los contrafactuales, en la Ecuación 2.20. Por linearidad de las esperanzas es posible descomponerlo como la diferencia de las esperanzas de los contrafactuales, que tienen el *adjustment formula* (2.12) como estimando.

$$\mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] \tag{2.20}$$

Bajo ignorabilidad condicional (Ecuación 2.16) el ATE equivale a la Ecuación 2.21, que será el funcional con el que se trabajará a la hora de estimar el ATE en este trabajo.

$$\text{ATE} = \mathbb{E}\left[\,\mathbb{E}[Y \mid Z = 1, Q] - \mathbb{E}[Y \mid Z = 0, Q]\,\right] \tag{2.21}$$

Por otro lado, se define el efecto individual del tratamiento como ***Individual Treatment Effect (ITE)*** de acuerdo a la Ecuación 2.22, representando la diferencia entre los posibles *outcomes* para un mismo individuo. La notación i indica que la métrica se observa sobre el individuo i.

El ITE es quizás la de mayor interés en el área, pero en escenarios reales es inobservable al involucrar contrafactuales.

$$ITE(i) = Y_1(i) - Y_0(i) \tag{2.22}$$

Otro de los estimandos ampliamente utilizados en el área se conoce como ***Conditional Average Treatment Effect (CATE)*** (2.23) y representa el efecto promedio esperado condicionado a un valor particular de las covariables, $Q = q$

$$E[Y_1 - Y_0 \mid Q = q] \tag{2.23}$$

Bajo ignorabilidad condicional (Ecuación 2.16) el CATE equivale a la Ecuación 2.24. En la experimentación realizada en esta tesis se utilizará esta reescritura como estimador del ITE, condicionando a través del conjunto Q de covariables disponibles bajo el supuesto de que cumplen ignorabilidad condicional 2.16.

$$\text{CATE} = \mathbb{E}[Y \mid Z = 1, Q] - \mathbb{E}[Y \mid Z = 0, Q] \tag{2.24}$$

2.6. Equivalencia entre enfoques: *Twin Networks* y *Structural Causal Models*

Previo al desarrollo del criterio *backdoor*, el enfoque basado en resultados potenciales contrafactuales no había desarrollado un criterio que permitiera determinar cuáles son los conjuntos de variables de control correctos para la estimación de efectos causales. Verificar el cumplimiento de la ignorabilidad condicional entre el tratamiento factual Z y

el resultado potencial contrafactual Y_z requiere mostrar que su distribución conjunta puede descomponerse como el producto de sus marginales, como se indicó en la Ecuación 2.17.

Recién en el año 1994 Judea Pearl mostró cómo definir la distribución de probabilidad conjunta entre un tratamiento Z y sus contrafactuales [8, 9]. Los modelos que incluyen contrafactuales son conocidos como *Twin Networks* [5], y se desprenden naturalmente de la lógica generativa de los modelos causales. Esto permite verificar si un conjunto de variables cumple con *ignorability*. Gracias a ello, hoy es posible verificar si *ignorability* es un criterio correcto para determinar variables de control revisando si se cumple la Equivalencia 2.25.

$$Y_z \perp\!\!\!\perp Z \mid Q \iff Q \text{ cumple el criterio } \textit{backdoor} \tag{2.25}$$

Se revisa este punto con un ejemplo: un tratamiento Z ejerce un efecto causal sobre la variable objetivo Y a través de un mediador M. Una red bayesiana posible para este escenario se representa en la Figura 2.11. Se asume que toda la aleatoriedad presente en estas variables está determinada por ruidos exógenos mutuamente independientes.

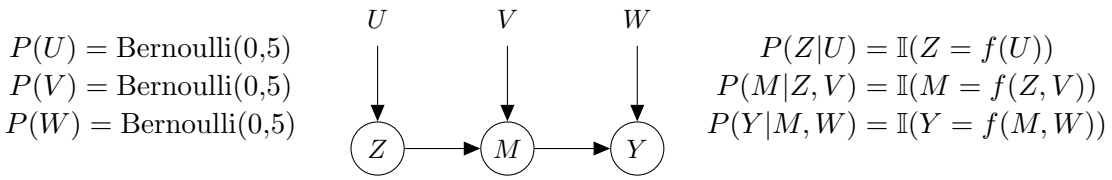


Fig. 2.11: Modelo generativo deterministico con variables exógenas

Si la realidad causal se comportara siguiendo el modelo de la Figura 2.11, se podría generar un conjunto de datos obteniendo muestras de los ruidos U, V y W, y luego generar los valores deterministas de las variables Z, M e Y. Hasta aquí nada nuevo.

Dado que la Figura 2.11 representa un modelo generativo, no resulta difícil imaginar qué pasaría si se hubiera aplicado un tratamiento contrafactual $Z' = z$, que se denota con el nombre Z_z. En términos generativos se obtendría un contrafactual del mediador, denominado M_z, el cual depende del tratamiento contrafactual Z_z y de su ruido original V, $M_z = f(Z_z, V)$. Además, se obtendría también un contrafactual de la variable objetivo, denotado con el nombre Y_z, el cual depende del mediador contrafactual M_z y de su ruido original W, $Y_z = f(M_z, W)$.

Este modelo generativo, que incluye contrafactuales, se puede especificar matemáticamente mediante la *Twin Network* representada en la Figura 2.12. Este tipo de redes bayesianas son la especificación matemática de la distribución conjunta entre las variables factuales Z, M, Y, U, V y W y las variables contrafactuales Z_z, M_z e Y_z.

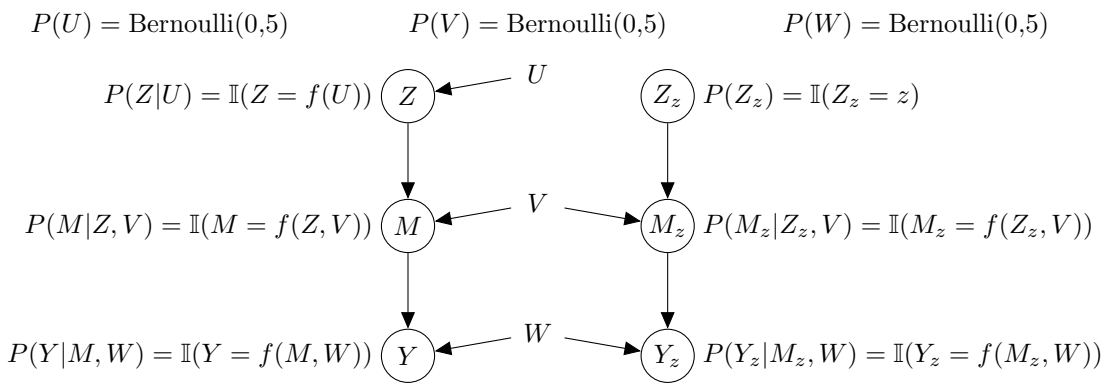


Fig. 2.12: Twin Network del modelo generativo determinístico con variables exógenas

En la Figura 2.12 la variable Y_z representa el valor contrafactual, que en caso de que el tratamiento Z sea binario, podría ser Y_0 o Y_1 según cuál sea el valor de z. Las *Twin Networks* son la definición de la distribución conjunta entre el tratamiento observado Z y los resultados potenciales contrafactuales Y_z, por lo que ahora es posible verificar efectivamente el cumplimiento o no del criterio *ignorability* expresado en la ecuación 2.16.

Revisando la Figura 2.11 se puede ver que en este caso no es necesario un conjunto de control, pues no hay caminos traseros entre Z e Y. Más aún, el único conjunto de control válido es el conjunto vacío, pues el criterio *backdoor* impide que se incluya el mediador M. Por otro lado, revisando la figura 2.12 se puede ver que todos los caminos entre Z e Y_z tienen un collider, por lo que el conjunto de control vacío cumple con hacer independientes el tratamiento factual Z del resultado potencial Y_z. Controlar por M abriría el flujo entre Z y V, haciendo que Z deje de ser independiente de Y_z. En este caso se cumple la equivalencia 2.25 entre *ignorability* y *backdoor*.

Además, notar que todas las variables endógenas fueron definidas como determinísticas, a través de distribuciones de probabilidad indicadoras que valen 1 si se cumple la condición al interior de la función indicadora y 0 en caso contrario. Modelar las variables endógenas como variables determinísticas no es casual. Si se hubiera definido al mecanismo causal de la variable M mediante una distribución de probabildiad no determinística, como por ejemplo $P(M|Z,V) = \text{Bernoulli}(M|f(Z,V))$, entonces el valor observado de la variable objetivo Y no necesariamente sería consistente con el valor de su contrafactual, Y_z (2.26).

$$Z = z \not\Longrightarrow Y = (1-z)\,Y_0 + z\,Y_1 = Y_z \tag{2.26}$$

Es decir, el supuesto de consistencia no vale en cualquier tipo de modelo generativo, sino sólo en aquellos en los que todas las variables endógenas sean deterministicas. **El enfoque de *potential outcomes* basado en el supuesto de consistencia no es válido en ningún modelo generativo para el cual algunas de las variables endógenas sea determinística**.

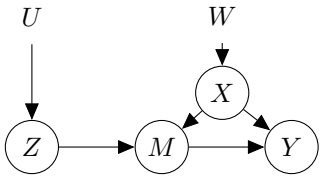
Pero incluso si se restringieran los modelos causales generativos para los cuales todas las variables endógenas son deterministicas, existen casos en los que la equivalencia 2.25 entre *ignorability* y *backdoor* no se cumple. En la Figura 2.13 se exhibe un modelo generativo sutilmente más complejo donde se representa un sistema M que para funcionar requiere exclusivamente que haya electricidad Z y que esté prendido el interruptor X. La presencia

de electricidad y la posición del interruptor depende de variables exógenas independientes entre sí. Por último, existe una alarma que se prende cuando el sistema no funciona a pesar de que el interruptor esté prendido, indicando así la falta de energía eléctrica.

Z: Electricidad (0: No hay, 1: Hay)
X: Interruptor (0: Apagado, 1: Prendido)
M: Sistema (0: No funciona, 1: Funciona)
Y: Alarma (0: Apagada, 1: Prendida)

$P(U) = \text{Bernoulli}(0{,}5)$

$P(W) = \text{Bernoulli}(0{,}5)$



$$P(Z|U) = \mathbb{I}(Z = U)$$
$$P(X|W) = \mathbb{I}(X = W)$$
$$P(M|Z, X) = \mathbb{I}(M = Z \cdot X)$$
$$P(Y|M, X) = \mathbb{I}(Y = M + X)$$

Fig. 2.13: Modelo generativo del funcionamiento de una sistema y su alarma

Para construir la *Twin Network* donde se interviene Z en el modelo 2.13 hay que seguir el razonamiento generativo. Si el estado de la energía eléctrica hubiera sido $Z_z = z$, luego habría un valor alternativo para el mediador M_z que depende de X y Z_z, y habría un valor alternativo de la variable objetivo Y_z que depende de X y M_z. El resultado se visualiza en la Figura 2.14. A diferencia del modelo generativo de la Figura 2.11, en la Figura 2.14 el funcionamiento del sistema M no tiene una variable exógena asociada, su comportamiento depende determinísticamente de variables que ya son endógenas, Z y X.

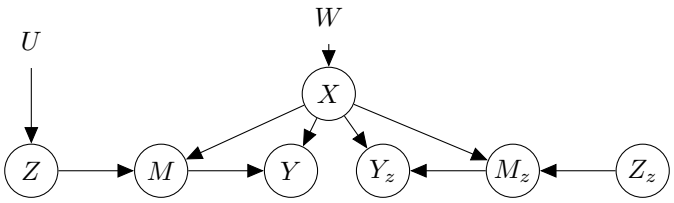


Fig. 2.14: Twin Network del modelo generativo del funcionamiento de una sistema y su alarma

En este caso tampoco hay caminos traseros que vayan de Z a Y, por lo que no es necesario controlar para calcular el efecto causal, $P(Y|\text{do}(Z)) = P(Y|Z)$. Sin embargo, existe un conjunto de control válido no vacío, que incluye únicamente a la variable X. Controlar por X permite evaluar efectos causales heterogéneos específicos al contexto X, $P(Y|\text{do}(Z), X) = P(Y|Z, X)$. En este caso, sin embargo, el conjunto de variables $\{X, M\}$ hace independiente la variable Z de contrafactual Y_z a pesar de que ese conjunto no cumpla el criterio backdoor (dado que controla por un mediador).

$$Y_z \perp\!\!\!\perp Z \mid \{X, M\} \not\Longrightarrow \{X, M\} \text{ cumple el criterio } \textit{backdoor} \tag{2.27}$$

Este es un ejemplo en el cual *ignorability* no implica *backdoor*. Para que se cumpla la equivalencia entre el criterio *backdoor* e *ignorability* es necesario que los modelos generativos tengan una variable exógena por cada una de las variables endógenas, que sean independientes entre sí y que todas las variables endógenas sean deterministas.

Para conciliar ambos mundos, Judea Pearl discute los problemas de inferencia causal sobre un subconjunto específico de modelos causales generativos, que él llama *Structural Equation Model* o *Structural Causal Model* [5]. Este subconjunto de modelos generativos deben cumplir dos condiciones: para que se cumpla el supuesto de consistencia 2.18 que requiere el enfoque de *potential outcomes* es necesario que todas las variables endógenas sean deterministicas, como ya se ha mencionado en esta sección. Pero además, cada variable endógena tiene que tener una variable exógena (y todas las variables exógenas deben ser independientes entre sí) para que se cumpla la equivalencia entre el criterio *backdoor* e *ignorability*.

2.7. Controles

Sólo para exhibir algunos ejemplos, se revisarán brevemente distintas estructuras causales y se discutirá si ciertos conjuntos de variables son conjuntos de control buenos, malos o neutrales [10].

Los controles buenos se caracterizan por bloquear únicamente los caminos *backdoor* entre Z e Y, sin bloquear los caminos delanteros entre Z e Y. En todos los casos de la Figura 2.15 X bloquea todos los caminos *backdoor* entre Z e Y, sin bloquear los caminos *frontdoor*, lo que es esperable para un buen conjunto de control.

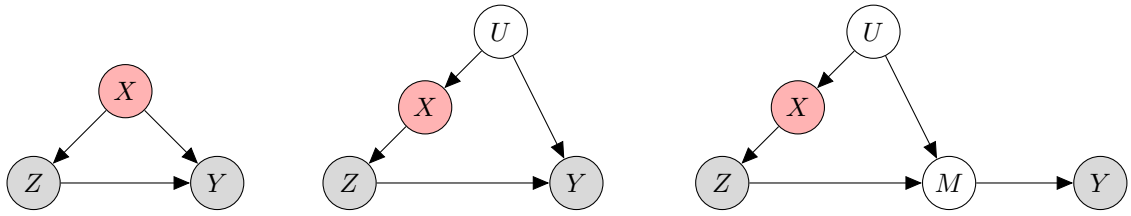


Fig. 2.15: Ejemplos de buenos controles

Por su parte, los malos controles se caracterizan por ser variables que son **afectadas causalmente por el tratamiento**, o por no cerrar los flujos de asociación *backdoor*.

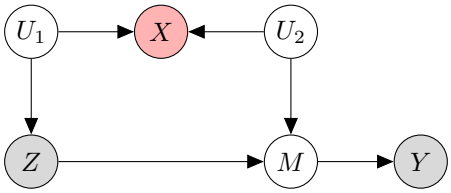


Fig. 2.16: Control malo que abre *backdoor path*

En el caso de la Figura 2.16 no hacía falta controlar, pues X es un *collider* que al ser oculto hace que sus extremos sean independientes. Al incluir X como variable de control se abre el flujo *backdoor* $Z \leftarrow U_1 \rightarrow X \leftarrow U_2 \rightarrow Y$, lo que genera correlaciones espurias entre Z e Y. Este caso es conocido como *M-bias*, y es relevante porque muestra que no siempre es buena idea agregar variables que están en el camino *backdoor*.

Finalmente hay una serie de conjuntos de control llamados neutrales, que al ser incluidas en un modelo no introducen sesgos en la estimación del efecto causal de interés, pero tampoco

son necesarias para lograr identificabilidad del efecto causal. En general son variables que no forman parte de ningún camino causal entre el tratamiento y el resultado, y tampoco bloquean *backdoor paths*. Sin embargo, incluir este tipo de variables al conjunto de control puede mejorar la precisión de la estimación, al añadir granularidad en el modelado de la relación entre las covariables y el *outcome*.

Fig. 2.17: Ejemplos de controles neutrales

Se puede ver en la Figura 2.17 que en el modelo de la izquierda controlar por X no abre ni bloquea *backdoor paths* entre Z e Y, pero es útil para un modelo de aprendizaje automático para reducir la varianza en la estimación de Y ya que es una variable causante del resultado. Algo similar ocurre con el modelo de la derecha: controlar por X es útil para aumentar la granularidad en la estimación de Y a través del intermediario M. Este tipo de controles neutrales que ayudan a la estimación de la variable objetivo son utilizados en la literatura de inferencia causal para estimar efectos causales heterogéneos, es decir, efectos que cambian según el contexto.

2.8. Estado del arte

A pesar de que las ciencias empíricas necesitan evaluar teorías causales alternativas como un todo, el elevado costo computacional asociado a calcular la distribución de probabilidad posterior de los modelos $P(\text{Modelo}|\text{Datos})$ ha hecho que esta tarea sea casi imposible. En consecuencia, la inferencia causal en su estado actual está limitada a la estimación de efectos causales entre pares de variables.

Cuando la estimación de efectos causales se realiza sobre datos observados sin intervenciones, conocer la estructura causal subyacente es un requisito ineludible para determinar cuál es el conjunto de variables de control que elimina correctamente las correlaciones espurias entre pares de variables. Por lo tanto, a pesar de que los modelos causales alternativos no pueden ser evaluados correctamente, la literatura de inferencia causal supone siempre una estructura causal subyacente para estimar efectos causales de datos observados.

Existe una creencia muy extendida de que la inferencia causal es algo fundamentalmente distinto al aprendizaje automático. Matheus Facure es un científico de datos brasilero, con varios años de experiencia en la industria, autor del libro *Causal Inference for the Brave and True* [11]. Este libro se ha convertido en una referencia en la industria y Facure es consultado frecuentemente por grandes empresas para recibir consejos relativos a inferencia causal. En la introducción de su libro, caracteriza la diferencia entre el aprendizaje automático y la inferencia causal afirmando que, si bien el primero es una gran herramienta para las tareas predictivas, su desempeño no es satisfactorio a la hora de predecir preguntas del tipo *what-if*, ya que no puede observar simultáneamente los efectos de tratar y no tratar a una misma unidad. Esta visión es compartida por Brady Neal, autor del libro

Introduction to Causal Inference from a Machine Learning Perspective [12], el cual se ha convertido en una referencia por la calidad de sus definiciones conceptuales y matemáticas. Este libro contiene una sección introductoria titulada *The fundamental problem of causal inference*. Allí Neal afirma que la inferencia causal enfrenta el problema fundamental de no poder observar ambos resultados potenciales para una misma unidad, lo que impide conocer directamente su efecto causal y la distingue del aprendizaje automático, que sólo busca predecir el resultado observado.

En conclusión, dos de los libros más influyentes actualmente en el área caracterizan al aprendizaje automático como una actividad predictiva, en contraste a la inferencia causal que estaría enfocada en estimar variables ocultas.

En primer lugar, es importante tener en cuenta que toda predicción por definición precede a la observación efectiva de las variables, y por lo tanto también son inicialmente estimaciones sobre variables ocultas. En segundo lugar, las estimaciones sobre los efectos causales son a su vez la predicción respecto del impacto que una intervención tiene sobre una variable objetivo. Mientras que el aprendizaje automático tradicional asume que los datos de entrenamiento y los datos futuros provienen de la misma distribución, las predicciones causales consideran los cambios que se producen en la distribución de probabilidad conjunta para mantener la validez predictiva cuando se producen intervenciones.

La estimación de efectos causales en datos observados sin intervenciones es en definitiva un problema de predicción *out-of-sample*: se entrena sobre datos que provienen de la estructura causal sin intervenir pero se quiere predecir el comportamiento para un contexto diferente, en el que la estructura causal está intervenida. Este problema fue resuelto a través del criterio *backdoor*, gracias a que las variables de control cortan las correlaciones espurias, permitiendo usar datos sin intervenciones para predecir el impacto que tendrían intervenciones no observadas. Esta solución abrió la puerta para que durante el primer cuarto de este siglo se desarrollara toda una línea de investigación dedicada a proponer y evaluar modelos de *Machine Learning* para inferencia causal.

Los modelos de *Causal Machine Learning* se construyen usando variables de control válidas que permitan la estimación de la probabilidad condicional de interés (Ecuación 2.28, izquierda), o en general simplemente su media (Ecuación 2.28, derecha).

$$P(Y|Z,\boldsymbol{Q}) \qquad \text{ó} \qquad \mathbb{E}[Y|Z,\boldsymbol{Q}] \tag{2.28}$$

Notar que la probabilidad condicional de interés $P(Y|Z,\boldsymbol{Q})$ es la predicción causal al interior de cada subgrupo $\boldsymbol{Q}=q$. Si se quisiera calcular el efecto causal general $P(Y|\text{do}(Z))$ se debería aplicar la *adjustment formula* (2.12), integrando cada uno de los efectos causales específicos al subgrupo $\boldsymbol{Q}$ por la probabilidad de que ocurra $\boldsymbol{Q}$. Si bien esto permite resumir la información del efecto causal, pierde al mismo tiempo información valiosa, relativa a la heterogeneidad de los efectos causales. Por ese motivo, en *Causal Machine Learning* se suele reportar la predicción granular, $P(Y|Z,\boldsymbol{Q})$, que es en sí misma una estimación de efectos causales heterogéneos, donde el impacto causal varía según el contexto $\boldsymbol{Q}$. Cuando el conjunto de variables $\boldsymbol{Q}$ es grande y la relación causal es no lineal, los modelos simples como la diferencia de medias o las regresiones lineales no alcanzan para calcular la superficie de los valores potenciales para el tratamiento y el control.

Los algoritmos más relevantes para la inferencia causal se basan en técnicas de ensamble de árboles de decisión tales como el *Boosting* [13,14], *Bagging* [15] y *Random Forest* [16], como

por ejemplo los *Bayesian Additive Regression Trees* (BART). Cada uno aplica diferentes técnicas para ajustar una combinación lineal de árboles.

En el año 2011 Jennifer Hill publicó su artículo *Bayesian Nonparametric Modeling for Causal Inference* [17]. Uno de sus argumentos a favor de usar modelos no paramétricos para inferencia causal consiste en que BART permite incluir múltiples covariables, incluso irrelevantes, lo que facilita la eliminación de las asociaciones espurias mediante variables de control válidas que cumplen el criterio *backdoor*. Este trabajo ha adquirido una enorme relevancia en la literatura de inferencia causal por haber introducido un enfoque más flexible, robusto y automatizado para el análisis de efectos causales. Se destaca además por el tipo de *benchmark* propuesto para evaluar el desempeño de modelos alternativos para inferencia causal: simulaciones basadas en conjuntos de datos reales. En otra sección del artículo, la autora señala que las simulaciones se basan en covariables reales y sólo simulan los resultados, asegurando *ignorability* al depender únicamente de variables observadas.

Esta metodología de evaluación es adoptada por el *Atlantic Causal Inference Conference* (ACIC) para las competencias de inferencia causal que comienzan en el año 2016 [3] y continuan los años 2017 [18], 2018 [19], 2022 [2], entre las competencias que se encuentran mejor documentadas. Todavía en el año 2022 (últimas competencias a las que se pudo acceder) el modelo de Hill continúa en las primeras posiciones del *ranking*.

Estas competencias aceleraron el desarrollo de modelos de *Causal Machine Learning*. En el año 2016 se presentaron equipos de empresas y universidades como IBM, Universidad de Harvard, Universidad de California, entre otras. En esa primera competencia, los modelos basados en BART obtuvieron los primeros lugares. En el año 2018 Wager y Athey presentaron un modelo basado en un *Random Forest* [20], que luego pasa a llamarse *Causal Forest*. En el año 2020 se presenta una mejora del BART, que al igual que el *Causal Forest* modela por separado el efecto base del efecto causal, además de la inclusión del *propensity score* como una covariable adicional [21]. Este modelo se conoce hoy en día como *Bayesian Causal Forest* y fue utilizado como *baseline* oficial en la competencia del año 2022. Sólo 4 tipos de modelos logran superar este *benchmark*. Uno de ellos es una extensión del *BART* presentado por la misma Jennifer Hill. Otro, el *diConfounder*, está basado en el *Causal Forest*. Es por estos *benchmarks* que, como se explicará en el capítulo siguiente, los modelos elegidos para evaluar efectos causales en esta tesis fueron BART, BCF y Causal Forest.

3. DOCUMENTACIÓN DE MODELOS

A continuación se presentarán los modelos utilizados en este trabajo. Para cada uno se describirán sus características, su funcionamiento y los motivos por los cuales su aplicación resulta pertinente para la estimación del efecto causal heterogéneo. Por otro lado, se brindará una breve explicación de las implementaciones utilizadas, junto con los hiperparámetros que fueron considerados para su ajuste.

3.1. *Boosting*, *Bagging* y *Random Forest*

Antes de adentrarse en el funcionamiento de *Bayesian Additive Regression Trees* (BART), *Bayesian Causal Forest* (BCF) y *Causal Forest* (CF), es importante repasar tres técnicas de aprendizaje automático fundamentales en las que se asientan sus bases: ***Bagging***, ***Boosting*** y ***Random Forest***. Estos enfoques consisten, principalmente, en el ensamble de múltiples modelos simples (en este caso, árboles de decisión) para formar un predictor más robusto, estable y preciso. Si bien en principio no fueron desarrollados específicamente para inferencia causal, su lógica de ensamble y sus ventajas en términos de sesgo-varianza sentaron un precedente para el diseño de los modelos de *Causal Machine Learning* actuales.

Bagging (acrónimo de *Bootstrap Aggregating*) fue introducido por Breiman en 1996 [15]. Se basa en entrenar múltiples árboles de decisión independientes sobre subconjuntos *bootstrap* del conjunto original de datos. El concepto de *bootstrap* [22] se refiere a una técnica de remuestreo con reposición: a partir de los datos dispoinibles, se generan múltiples subconjuntos aleatorios con el tamaño del conjunto de datos original, lo que permite estimar la variabilidad de un estimador de manera no paramétrica y aproximar su distribución sin la necesidad de contar con nuevos datos ni definir supuestos fuertes.

En el contexto de *Bagging*, cada árbol se entrena sobre un conjunto *bootstrap* diferente. Luego, sus predicciones se combinan por promedio en problemas de regresión, o por voto mayoritario en clasificación. El objetivo es reducir la varianza del modelo sin aumentar el sesgo, ya que el promedio de múltiples modelos ruidosos pero no sesgados tiende a estabilizarse cerca de la verdad subyacente.

Boosting [13, 14], por su parte, construye los árboles de manera secuencial: cada nuevo árbol se entrena sobre los errores residuales del conjunto anterior, enfocándose en las observaciones mal estimadas. Esto tiende a reducir el sesgo del modelo global.

Finalmente, ***Random Forest***, propuesto también por Breiman en 2001 [16], mejora aún más el desempeño del *Bagging* al introducir aleatoriedad adicional: en cada división del árbol, en lugar de considerar todas las variables disponibles, se selecciona aleatoriamente un subconjunto. Esta técnica reduce la correlación entre los árboles del ensamble, lo que mejora aún más la reducción de la varianza y el poder predictivo del modelo combinado.

3.2. BART: *Bayesian Additive Regression Trees*

El primer modelo utilizado en este trabajo fue introducido en 2010 como un modelo de regresión bayesiana no paramétrica basado en un ensamble de árboles de decisión. [23]

BART propone modelar el *output* Y a través de un modelo dado por las Ecuaciones 3.1 y 3.2, siendo $g(x; T_j, M_j)$ la predicción obtenida para el vector de covariables x, considerando el árbol de regresión T_j con el conjunto de hojas M_j. De esta manera, la esperanza de las estimaciones de Y coincide con la suma de los árboles.

$$Y = \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon, \tag{3.1}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{3.2}$$

Para controlar las predicciones generadas por BART y acotarlas a un rango de valores acorde al Y de interés, los autores proponen las siguientes *priors* [23]:

- ***Prior* sobre la estructura de los árboles T_j:** Se favorecen árboles poco profundos disminuyendo la probabilidad de divisiones con la profundidad, con el objetivo de regularizar la predicción final sobre todos los árboles y evitar el sobreajuste.
- ***Prior* sobre los valores de las hojas M_j:** Para que la suma total de árboles tenga una magnitud controlada, se proponen distribuciones normales centradas en cero para las hojas de los árboles, con varianza ajustada según la cantidad total de árboles.
- ***Prior* sobre el ruido σ^2:** Se modela el ruido con una distribución inversa χ^2, el cual se calibra empíricamente para reflejar de forma razonable la variabilidad esperada en los datos, y permitiendo también cuantificar la incertidumbre en las predicciones.

Tras definir estas variables, la posterior es estimada utilizando cadenas Markov Chain Monte Carlo (MCMC) en un estilo de *Boosting*. De esta manera, en cada iteración se recorren los árboles y se calculan los residuos, tal como se indica en la Ecuación 3.3, entrenando luego el árbol T_j para explicar este residuo.

$$R_j \equiv y - \sum_{k \neq j} g(x; T_k, M_k) \tag{3.3}$$

3.2.1. BART para inferencia causal

En su trabajo de 2011, Jennifer Hill utilizó BART para estimar efectos causales heterogéneos [17]. Su metodología consistió en ajustar el modelo como $f(z, x)$, siendo z la asignación del tratamiento y x la matriz de atributos (válidos como variables de control). Además de elegir este modelo por su gran *performance* predictiva, su poder para capturar relaciones no lineales y su estabilidad, Hill lo utilizó porque el sampleo a través de cadenas MCMC lo hacen un modelo ideal para estimar efectos causales.

Cuando se utiliza BART para estimar efectos causales se obtienen sucesivas estimaciones del efecto causal, $c(x, f) \equiv f(1, x) - f(0, x)$ durante las iteraciones de las cade-

nas MCMC, lo que constituye una estimación de la distribución conjunta de $C(f) = (c(x_1, f), \ldots, c(x_k, f))$ con $x_1 \ldots x_k$ el valor de las covariables de los k individuos.

Este sampleo, como se verá más adelante, facilita el cálculo de métricas de interés como la predicción del ITE a través de la estimación del CATE, además de permitir calcular intervalos de predicción para el ITE.

3.2.2. Paquete `dbarts`

Para este trabajo se utilizó la implementación de BART del paquete `dbarts` [24] de R, principalmente, por las siguientes razones:

- **Implementación eficiente:** Al estar implementado internamente en C++, presenta un mejor rendimiento computacional que otras implementaciones, como por ejemplo `bayesTree`, utilizada en el artículo original de Hill [17].
- **Soporte para predicciones posteriores:** A diferencia de otras implementaciones utilizadas en el ámbito como `bayesTree` o `bartMachine`, `dbarts` permite acceder fácilmente a la distribución posterior de las predicciones para cada observación, lo que es fundamental para el cálculo de contrafactuales, necesarios para estimar métricas como el ITE y el ATE.
- **Compatibilidad con análisis de convergencia:** El paquete facilita la extracción de muestras MCMC completas, lo que permite ejecutar métricas sobre las cadenas para evaluar convergencia y eficiencia del muestreo bayesiano para medir la calidad del modelo.

En particular se utilizó la función `bart2`, que es una implementación más reciente y eficiente que la función `bart` del mismo paquete. Los hiperparámetros que fueron tenidos en cuenta en este trabajo fueron los siguientes:

- `n.chains`: Especifica cuántas cadenas MCMC independientes serán ejecutadas.
- `n.burn`: Número de iteraciones iniciales que se descartan para evitar el sesgo inicial de la cadena.
- `n.samples`: Número de sampleos de la distribución posterior tras descartar las iteraciones de *burn-in* del ítem anterior.
- `n.thin`: Indica, para los sampleos de la posterior, cada cuántas iteraciones se guardará una muestra. El objetivo de agregar este hiperparámetro es reducir la autocorrelación entre muestras consecutivas.

3.2.3. Criterios de convergencia para las cadenas MCMC

Como se mencionó anteriormente, BART genera cadenas MCMC para muestrear la distribución *a posteriori* de la variable objetivo. La convergencia de las mismas es fundamental para garantizar validez en las estimaciones generadas y la eliminación del sesgo inicial en las muestras *burn-in*.

Para evaluar la convergencia de las cadenas generadas por los ajustes de BART en este trabajo se recurrió a las siguientes métricas:

- **Diagnóstico de convergencia de Gelman-Rubin:** Es un estadístico para evaluar la convergencia de cadenas MCMC paralelas [25]. Para esto compara la varianza inter-cadenas con la varianza intra-cadena. Conceptualmente, si todas las cadenas han convergido al mismo equilibrio, ambas varianzas deberían ser similares. El estadístico utilizado se define como en la Ecuación 3.4, donde W es la varianza promedio dentro de las cadenas y $\hat{V}$ es una estimación de la varianza total esperada, que combina la información dentro y entre cadenas.

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}} \tag{3.4}$$

  Un valor de $\hat{R}$ cercano a 1 indica que las cadenas convergieron satisfactoriamente, mientras que valores mayores sugieren que aún puede haber dependencia de las condiciones iniciales. La implementación utilizada fue a través de la función `gelman.diag()` del paquete `coda` [26].

- ***Effective Sample Size*:** Es una métrica que estima el número de muestras verdaderamente independientes generadas por el algoritmo MCMC [27]. Debido a la autocorrelación entre iteraciones sucesivas, no todas las muestras aportan información nueva. Esta métrica ajusta el tamaño total de la muestra para reflejar sólo la información efectiva. La estimación se basa en la Ecuación 3.5, donde N es el número total de iteraciones de la cadena y ρ_k es la autocorrelación en el *lag* k, es decir, el número de pasos entre dos observaciones de la secuencia que se consideran.

$$n_{\text{eff}} \simeq \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho_k} \tag{3.5}$$

  En la práctica, la suma se detiene cuando las autocorrelaciones dejan de ser significativas. Un valor bajo de *effective size* sugiere que la cadena tiene alta autocorrelación, por lo que idealmente se buscan valores lo más cercanos posible al tamaño total de la cadena.

  La implementación que se utilizó fue mediante la función `effectiveSize` del paquete `coda` [26].

3.3. BCF: Bayesian Causal Forest

Bayesian Causal Forest se trata de un modelo bayesiano no paramétrico desarrollado para la estimación de efectos causales heterogéneos, buscando obtener una estimación no sesgada del ITE en situaciones donde BART, al estar orientado a la predicción y no diseñado para inferencia causal, no presenta buena *performance* para la estimación causal [28].

En contextos observacionales es común que el tratamiento esté asignado en función de covariables que también influyen en el *outcome*, generando *confounding*. Este tipo de selección se conoce como *targeted selection*. Por ejemplo, se suele ver que en casos médicos el tratamiento es asignado a pacientes con peores antecedentes, con la esperanza de que efectivamente sea capaz de mejorar su pronóstico.

Por otro lado, los modelos predictivos pueden inducir un sesgo conocido como *Regularization Induced Confounding* (RIC), en el cual la regularización del modelo lleva a confundir el efecto del tratamiento con el efecto pronóstico (*outcome* que tendría una persona si no recibe el tratamiento dadas sus covariables, es decir $E(Y \mid Z = 0, X)$), ya que la regularización favorece explicaciones más simples y puede privilegiar estructuras asociadas a covariables altamente predictivas del *outcome*, desfavoreciendo la predicción a través de efectos del tratamiento.

En este contexto, BCF propone reparametrizar el modelo de acuerdo a la Ecuación 3.6, donde:

- $\pi(X) = P(Z = 1 \mid X)$ es el *propensity score*
- $\mu(X, \pi(X))$ representa el resultado esperado en ausencia de tratamiento (efecto pronóstico)
- $\tau(X)$ representa el CATE

$$\mathbb{E}[Y \mid X, Z] = \mu(X, \pi(X)) + \tau(X) \cdot Z \tag{3.6}$$

Con la reparametrización 3.6 se proponen nuevos priors para μ (efecto pronóstico) y τ (CATE), de manera de captar mejor su comportamiento esperado:

- Para $\mu(X, \pi(X))$ se asigna un prior más flexible, de forma que pueda capturar mejor el posible *confounding*
- Para $\tau(X)$ se asigna un prior más restrictivo, a partir de la suposición de que la heterogeneidad en los efectos del tratamiento es limitada

En cuanto a μ y τ, el modelo asignado a sus ajustes consiste en un BART independiente para cada uno, con valores específicos para sus hiperparámetros que se detallan en profundidad en el artículo original [28] y exceden el alcance de esta tesis, mientras que la estimación de π es requerida como un parámetro de entrada del modelo y no la estima por sí mismo.

3.3.1. Paquete `stochtree`

La implementación de BCF utilizada en este trabajo es la del paquete `stochtree` [29], principalmente por las siguientes razones:

- **Control de hiperparámetros:** Al estar implementado en R, si bien no está optimizado para producción o *big data*, ofrece mucho control sobre los hiperparámetros y el análisis bayesiano de la cadena MCMC generada.
- **Rapidez de ejecución:** En *datasets* de la magnitud que se manejó en este trabajo, el paquete ofrece mayor velocidad que otras alternativas más modernas como el paquete `bcf`.

Los hiperparámetros tenidos en cuenta en este caso fueron los siguientes:

- `num_grf`: Número de iteraciones *warm-start* previas a usar el algoritmo *grow-from-root* [30]. Se trata de iteraciones que ayudan a establecer una configuración inicial

estable para los árboles, lo que mejora la convergencia del algoritmo MCMC posterior.

- `num_burnin`: Número de iteraciones iniciales de la cadena MCMC a descartar para evitar el sesgo inicial.
- `num_mcmc`: Muestras a guardar de la cadena MCMC tras descartar las iteraciones de *burn-in*.
- `prognostic_forest_params$num_trees`: Número de árboles a utilizar para la estimación del efecto pronóstico μ.
- `treatment_effect_forest_params$num_trees`: Número de árboles a utilizar para la estimación del CATE τ.

Si bien el paquete más aceptado y utilizado en la literatura es `bcf`, que fue introducido en el artículo original del modelo, se optó por esta alternativa debido a que el tiempo de ejecución del paquete original resultó prohibitivo para su uso en este trabajo.

Las principales desventajas de la implementación del paquete `stochtree` radican en la **pérdida de reproducibilidad**: aunque el modelo incluye un parámetro para fijar la semilla, esta no se emplea durante el ajuste. También se pierde la posibilidad de manejar múltiples cadenas MCMC, que es una característica muy importante para el modelo tal como fue planteado originalmente.

Además, **el paquete `stochtree` realiza una estimación interna del propensity score** π, mientras que `bcf` tal como fue planteado originalmente requiere a π como parámetro de entrada, tal como se había explicado anteriormente. Esto demuestra, por un lado, una mayor flexibilidad respecto a la implementación original, aunque también evidencia la dificultad para justificar los resultados obtenidos con `stochtree`, cuyo soporte y documentación son limitados, sin dejar en claro de qué manera se estima el parámetro π internamente.

3.3.2. Criterios de convergencia para las cadenas MCMC

Como se mencionó en la sección anterior, a pesar de que en el artículo original del modelo [28] se menciona que el mismo genera múltiples cadenas MCMC paralelas, el paquete `stochtree` sólo genera una. Es por este motivo que para BCF no se pudo calcular el Diagnóstico de Gelman Rubin como se hizo para BART, ya que esta métrica evalúa la convergencia entre cadenas paralelas.

Considerando esta limitación se evaluó la convergencia de la cadena generada mediante el *effective size* (3.2.3), y mediante la **autocorrelación**.

La autocorrelación mide el grado de independencia entre los valores sucesivos de una cadena MCMC por *lag*, de manera muy similar al *effective size*. Una alta correlación deja en evidencia que las muestras generadas por la cadena están muy relacionadas entre sí y, por lo tanto, aportan menos información nueva. Matemáticamente, la autocorrelación en el *lag* k se define en la Ecuación 3.7.

$$\rho_k = \frac{\mathrm{Cov}(X_t, X_{t+k})}{\mathrm{Var}(X)} \tag{3.7}$$

Se espera que la autocorrelación disminuya rápidamente con el *lag*.

3.4. Causal Forest

Los *Causal Forest* son *Random Forest* con una función de costo particular enfocada a la estimación de efectos heterogéneos [31, 32].

La estimación del CATE (τ) se obtiene minimizando la función de costo de la Ecuación 3.8, donde Y_i es la variable objetivo, Z_i el tratamiento, X_i los controles, $\Lambda_n(\tau(\cdot))$ el regularizador, $\hat{m}^{-i}(x) = \hat{E}[Y_i|X_i = x]$, $\hat{\pi}^{-i}(x) = \hat{E}[Z_i|X_i = x]$ con el supra-índice $-i$ indica que las estimaciones fueron realizadas sin el i-ésimo elemento.

$$\hat{\tau}(\cdot) = \underset{\tau}{\operatorname{argmin}} \sum_i \Big(\big(Y_i - \hat{m}^{-i}(X_i)\big) - \tau(X_i)\big(Z_i - \hat{\pi}^{-i}(X_i)\big) \Big)^2 + \Lambda_n(\tau(\cdot)) \tag{3.8}$$

A continuación se explica esta función de costo 3.8 de los *Causal Forest*, conocida como *R-Learner* por Robinson (1988) o también como *Residual-Learner* [33].

Sean X_1 y X_2 dos conjuntos de variables, se pretende estimar los parámetros asociados a la regresión de la Ecuación 3.9.

$$\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \tag{3.9}$$

Existe una forma de obtener el parámetro $\hat{\beta}_1$ que requiere los siguientes pasos:

1. Realizar una regresión de Y solo con X_2, $\hat{Y}^* = \hat{\gamma}_1 X_2$.
2. Realizar una regresión de X_1 solo con X_2, $\hat{X}_1 = \hat{\gamma}_2 X_2$
3. Obtener los residuos $\tilde{X}_1 = X_1 - \hat{X}_1$ y $\tilde{Y} = Y - \hat{Y}^*$
4. Realizar una regresión entre los residuos, $\tilde{Y} = \hat{\beta}_1 \tilde{X}_1$

La estimación $\hat{\beta}_1$ que se obtiene en la regresión entre residuos resulta ser exactamente igual a la estimación $\hat{\beta}_1$ que se obtiene de una regresión clásica (3.9). Este resultado puede usarse para estimar los efectos causales que el tratamiento tiene sobre la variable objetivo a través de los residuos, tal como se indica en la Ecuación 3.10, donde $\sim$ representa el operador de la regresión.

$$\big(Y - (Y \sim X)\big) \sim \big(Z - (Z \sim X)\big) \tag{3.10}$$

En otras palabras, el parámetro del efecto causal τ queda determinado por la Ecuación 3.11.

$$Y_i - E[Y_i|X_i] = \tau \cdot (Z_i - E[Z_i|X_i]) + \varepsilon \tag{3.11}$$

Esta formulación ahorra mostrar la tabla con parámetros asociados a los controles que no son de interés, y en particular permiten evitar modelar explícitamente los controles si se utilizan modelos de aprendizaje automático para estimar las esperanzas. Esto se formaliza en la Ecuación 3.12, donde $\hat{M}_y(X_i)$ y $\hat{M}_z(X_i)$ son estimaciones basadas en modelos de aprendizaje automático flexibles capaces de capturar interacciones y relaciones no lineales. Es decir, se generaliza el método descripto que utiliza regresiones a modelos más generales.

$$Y_i - \hat{M}_y(X_i) = \tau \cdot (Z_i - \hat{M}_z(X_i)) + \varepsilon \tag{3.12}$$

El punto importante aquí es que no es necesario hacer ningún supuesto paramétrico entre las covariables X y el objetivo Y ni entre las covariables X y el tratamiento Z, de manera tal que los pasos a seguir quedan determinados por:

1. Estimar el objetivo Y en base a las covariables X mediante un modelo de aprendizaje automático flexible, M_y
2. Estimar el objetivo Z en base a las covariables X mediante un modelo de aprendizaje automático flexible, M_z
3. Obtener los residuos $\tilde{Z} = Z - \hat{M}_z$ y $\tilde{Y} = Y - \hat{M}_y^*$
4. Obtener una estimación de τ a través de una regresión entre los residuos, $\tilde{Y} = \tau\tilde{Z} + \epsilon$

Los modelos de aprendizaje automático proveen de flexibilidad, pero se debe tener precaución para no sobreajustar. En particular, si los modelos M_y sobreajustan, los residuos $\tilde{Y}$ van a ser menores de lo que deberían ser, por lo que la regresión entre residuos va a estar sesgada hacia cero. Por otro lado, si los modelos M_z sobreajustan, la varianza de los residuos $\tilde{Z}$ va a ser menor a la que debería, por lo que la varianza de la regresión entre residuos va a ser mayor a lo esperado.

Para evitar estos problemas, las estimaciones puntuales se obtienen por validación cruzada, sin usar el i-ésimo elemento para predecirlo ($M_y^{-i}(X_i)$). Los *Random Forest* que reportan este tipo de estimaciones sobre validación cruzada son denominados como *Honest Random Forest*. Se dice que los residuos sobre los tratamientos están des-sesgados (*debiased*) en tanto que al ser ortogonales a los controles, no pueden ser explicados por ellos. Los residuos sobre los objetivos se dice que están libres de ruido (*denoised*) en tanto que al ser ortogonales a las covariables, no pueden ser explicados por ellas.

Para estimar efectos causales heterogéneos (CATE) se considera que el parámetro del efecto casual τ es una función de las covariables, tal como se formaliza en la Ecuación 3.13, **estimándolo finalmente a través de una agregación de las hojas de múltiples árboles individuales**. Finalmente,

$$Y_i - \hat{M}_y(X_i) = \tau(X_i) \cdot (Z_i - \hat{M}_z(X_i)) + \varepsilon \tag{3.13}$$

3.4.1. Paquete `grf`

La implementación de *Causal Forest* utilizada en este trabajo fue la del paquete `grf` [34], principalmente por las siguientes razones:

- **Implementación original:** Es la implementación más fiel al artículo original de Athey, Tibshirani y Wager [31].
- **Eficiencia:** Al igual que `dbarts` está implementado en C++, lo que aumenta considerablemente su eficiencia temporal.
- ***Honest splitting*:** Permite utilizar el enfoque de árboles que, como se vio anteriormente, evita el sobreajuste y permite estimar los efectos causales de manera más robusta al no utilizar los mismos datos que se utilizan para los *splits* para la estimación de los efectos causales de las hojas.
- **Intervalos de confianza:** Devuelve una estimación de la varianza de las estimaciones del τ como la varianza obtenida entre las hojas de los árboles. Esto es fun-

damental para construir intervalos de confianza para las predicciones del ITE, tal como se hizo con los otros modelos.

En este caso se ajustaron los siguientes hiperparámetros de la función `causal_forest` del paquete:

- `num.trees`: Número de árboles en el bosque.
- `sample.fraction`: Fracción de la muestra a utilizar para entrenar cada árbol, que luego además se vuelve a dividir para el *honest splitting*.
- `mtry`: Número de variables a considerar en cada división de los árboles.
- `min.node.size`: Número mínimo de observaciones en cada hoja de los árboles.
- `alpha`: Parámetro de regularización que controla el máximo imbalance de los *splits*.
- `honesty.fraction`: Fracción de la muestra a utilizar para el *honest splitting*.

Una característica importante a tener en cuenta de este paquete es que no devuelve estimaciones del outcome Y, por lo tanto, para los casos en los que fue necesaria esta estimación se calculó a través de las estimaciones de π, μ y τ, como se muestra en la Ecuación 3.13.

4. *DATASETS* Y *BENCHMARK*

¿Cómo hacer para evaluar el desempeño que tienen diversos modelos de *Machine Learning* para la estimación de efectos causales? Este es un problema que no es de fácil solución cuando se estiman efectos causales en datos observados sin intervenciones, pues la estimación de efectos causales es una predicción *out-of-sample* y por lo tanto no se cuenta con la posibilidad de evaluar el modelo sobre un conjunto de datos de testeo, ya que tal conjunto no existe.

En este sentido, encontrar un *benchmark* para *Causal Machine Learning* ha sido esencial para el desarrollo de la disciplina. En esta sección se introducirá uno de los *benchmark* con mayor aceptación en el ámbito, basado en la metodología usada por Jennifer Hill en su artículo fundamental [17]. La propuesta de Hill consiste en evaluar el desempeño de los modelos en conjuntos de datos simulados. Sin embargo, las simulaciones a menudo son demasiado simples respecto de los análisis de datos del "mundo real". Por ese motivo, Hill propone utilizar datos reales simulando únicamente las variables contrafactuales Y_0 e Y_1, logrando así simular conjuntos de datos suficientemente complejos, a los cuales se le conoce los verdaderos efectos causales para cada uno de los individuos.

Las simulaciones basadas en datos reales fueron adoptadas luego como *benchmark* para las competencias de datos organizadas por el *Atlantic Causal Inference Conference*.

En este trabajo se seleccionaron tres procesos generativos que dan lugar a distintos tipos de *datasets*, cada uno con características particulares que justifican su inclusión. A continuación se describen sus principales propiedades, las covariables disponibles y, cuando corresponde, los procesos generativos utilizados para su construcción.

4.1. Simulación simple

En primera instancia se seleccionó un conjunto de datos pequeño con observaciones simuladas. El mismo fue introducido por Jennifer Hill en su trabajo de 2011 [17], y contiene 120 puntos generados a partir de distribuciones conocidas. Se utilizó para visualizar y analizar las propiedades más visibles de los modelos considerados en este trabajo.

El conjunto de datos consiste en tres atributos:

- Z, la variable tratamiento
- X, una variable confundidora
- Y, la variable objetivo

Las ecuaciones a partir de las que se generó el dataset se encuentran en las Ecuaciones 4.1.

$$
\begin{aligned}
P(x) &= \frac{1}{2}\mathcal{N}(20,\ 10^2) + \frac{1}{2}\mathcal{N}(40,\ 10^2) \\
P(z|x) &= \text{Bernoulli}\Big(z\big|\sigma\big(\frac{x}{5} - 6\big)\Big) \text{ con } \sigma \text{ la función sigmoidea} \\
P(y|x,z) &= \mathcal{N}\Big(y\big|(72 + 3\sqrt{|x|})^{1-z}(90 + \exp(0{,}06x))^z,\ 1\Big)
\end{aligned}
\tag{4.1}
$$

Estas distribuciones de probabilidad condicional representan la misma distribución conjunta $P(x, y, z)$ con la que generó los datos Hill [17]. En el artículo original, la autora genera los datos ddefiniendo distribuciones condicionales en las que x aparece como variable mediadora entre z e y, $P(z)P(x \mid z)$, y luego estimaba el efecto causal condicionando sobre x, lo cual como se explicó en la sección 2 no es correcto en este tipo de contextos.

Para evitar este problema, asegurar una correcta estimación del efecto causal al condicionar por x, y poder comparar los resultados a los de Hill, en este trabajo se modificó la estructura causal del proceso generativo, redefiniendo x como un *fork*. De este modo, condicionar sobre x es válido ya que cierra el flujo de asociación trasero. Bajo las Ecuaciones 4.1, la relación causal entre las covariables queda determinada por el grafo de la Figura 4.1.

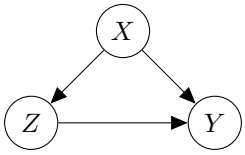


Fig. 4.1: Modelo causal de la simulación simple

4.2. Simulación basada en el *Infant Health and Development Program*

El *dataset* IHDP (*Infant Health and Development Program*) proviene de un estudio clínico aleatorizado iniciado en 1985 en EE.UU., diseñado para evaluar intervenciones sobre el desarrollo cognitivo de bebés nacidos con bajo peso o de forma prematura. El grupo tratado recibió visitas domiciliarias, cuidado de alta calidad y participación en un centro de desarrollo infantil. El objetivo del estudio era mitigar los efectos adversos de un nacimiento con riesgo.

Fue introducido por Jennifer Hill en su trabajo de 2011 [17], quien digitalizó los datos recolectados en 1985 con el objetivo de construir un conjunto de datos basado en covariables reales que permitiera generar *outcomes* en entornos de simulación controlados. Si bien no se cuenta con una variable de *outcome* en el dataset original, es un conjunto de datos ampliamente utilizado en el ámbito para la generación de escenarios sintéticos.

4.2.1. Estructura del *dataset* y generación de *outcomes*

El *dataset* original recolectado en 1985 consta de aproximadamente 80 variables pretratamiento, pero en los estudios realizados por Hill y posteriores se seleccionó un subconjunto de 28 covariables, sin incluir la columna de tratamiento. Las mismas se eligieron para generar los *outcomes* controlados con un subconjunto que incluye variables continuas

y binarias relacionadas con condiciones biomédicas del niño, características demográficas de la madre y contexto familiar.

Las covariables seleccionadas se describen en la Tabla 4.1. [35]

#	Atributo	Descripción	Tipo
1	`treat`	Asignación al tratamiento	Binaria
2	`bw`	Peso al nacer en gramos	Continua
3	`b.head`	Circunferencia de la cabeza del bebé al nacer (cm)	Continua
4	`preterm`	Semanas de gestación al nacer (prematuro)	Continua
5	`birth.o`	Orden de nacimiento del bebé	Continua
6	`nnhealth`	Índice de salud neonatal	Continua
7	`momage`	Edad de la madre al momento del nacimiento	Continua
8	`sex`	Género del bebé	Binaria
9	`twin`	Indica si el bebé es parte de un parto múltiple	Binaria
10	`b.marr`	Indica si la madre estaba casada al momento del parto	Binaria
11	`mom.lths`	Madre con educación inferior a secundaria	Binaria
12	`mom.hs`	Madre con educación secundaria completa	Binaria
13	`mom.scoll`	Madre con algo de educación universitaria	Binaria
14	`cig`	Indica si la madre fumó durante el embarazo	Binaria
15	`first`	Indica si el bebé es el primer hijo	Binaria
16	`booze`	Indica si la madre consumió alcohol durante embarazo	Binaria
17	`drugs`	Indica si la madre consumió drogas durante embarazo	Binaria
18	`work.dur`	Indica si la madre trabajó durante el embarazo	Binaria
19	`prenatal`	Indica si la madre recibió cuidado prenatal	Binaria
20-26	`site1-site7`	Sitio del programa (1 a 7)	Binaria
27-29	`momwhite, black, hisp`	Indica si la madre es blanca, negra o hispana	Binaria

Tab. 4.1: Atributos a utilizar del dataset IHDP

La autora afirma que se introdujo artificialmente un **sesgo de selección** al eliminar del *dataset* a los niños tratados cuyas madres no fueran blancas, simulando artificialmente un sesgo común en estudios observacionales. Sin embargo, para que esto suceda la raza debería ser un collider en un camino que conecte el tratamiento y el *outcome*, pero no queda claro que la estructura causal subyacente tenga esa forma. De todas formas, se replicó el procedimiento seguido por Hill.

De esta manera, el *dataset* utilizado quedó con las 25 covariables restantes de la Tabla 4.1.

Para los *outcomes*, se definen dos funciones generativas —denominadas en la literatura como "*Response Surface A*" y "*Response Surface B*"— [17] que dan lugar a diferentes niveles de complejidad estructural.

Para la generación de las superficies se utilizaron todas las covariables X, estandarizando previamente a media cero y varianza unitaria a las covariables continuas, y se mantuvo a la variable de tratamiento Z con el valor original observado.

4.2.1.1. Superficies de tipo A: efecto homogéneo

Con esta superficie se simula un efecto de tratamiento homogéneo para todas las unidades. Los *outcomes* potenciales se generan de acuerdo a las Ecuaciones 4.2, donde el vector de coeficientes β_A se construye seleccionando cada entrada de forma independiente a partir del conjunto $\{0, 0.1, 0.2, 0.3, 0.4\}$ con probabilidades respectivas $(0.5, 0.2, 0.15, 0.1, 0.05)$.

$$\begin{aligned} Y_0 &= X\beta_A + \varepsilon \\ Y_1 &= X\beta_A + 4 + \varepsilon \\ \varepsilon &\sim \mathcal{N}(0, 1) \end{aligned} \tag{4.2}$$

En el artículo original, la autora hace el sampleo a partir del conjunto $\{0, 1, 2, 3, 4\}$ [17], pero al emplear este vector en el presente trabajo se observó que inducía un ruido tan grande que opacaba el efecto homogéneo que se pretendía simular: la única forma de que los modelos tuvieran un desempeño aceptable en la estimación de los efectos individuales era a través del sobreajuste, lo cual se evidenciaba al evaluar los modelos en el conjunto de entrenamiento.

Por otro lado, en el código del artículo original el vector β_A varía en cada iteración del proceso generativo de las superficies, causando que el efecto causal simulado fuera distinto entre los individuos. Esto también fue modificado para este trabajo, donde el vector β_A no varía en las distintas generaciones de *outcomes*.

Para visualizar con mayor claridad de qué manera se relacionan las medias de estas superficies, se realizó un gráfico de medias en función de una grilla con valores factibles de la multiplicación entre X y β_A, que se visualiza en la Figura 4.2. Se ve claramente que las superficies de respuesta para individuos tratados y no tratados tienen una media paralela, simulando un tratamiento que produce el mismo efecto para todos.

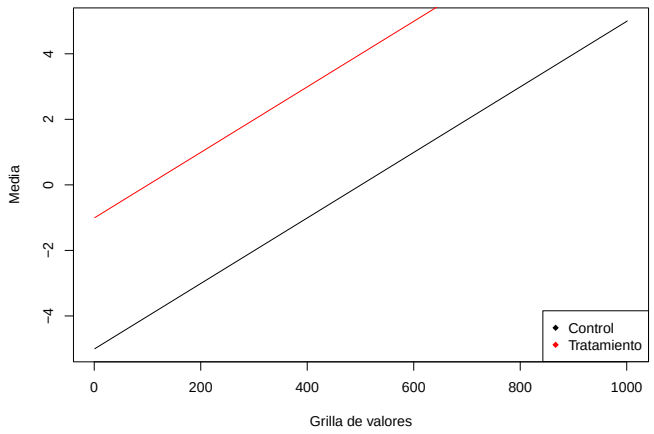


Fig. 4.2: Medias de la superficie de respuesta A. La línea roja representa la media de la distribución de Y_1 condicional a X, es decir $X\beta_A + 4$, mientras que la línea negra representa la media de la distribución de Y_0 condicional a X, es decir $X\beta_A$.

4.2.1.2. Superficies de tipo B: efecto heterogéneo y no lineal

En este caso, los *outcomes* fueron generados de acuerdo a las Ecuaciones 4.4, donde:

- W es una matriz de igual dimensión que X con todos los elementos iguales a 0.5;
- β_B se genera seleccionando cada componente aleatoriamente de {0, 0.1, 0.2, 0.3, 0.4} con probabilidades (0.6, 0.1, 0.1, 0.1, 0.1);
- el término ω_B^s es una constante calculada numéricamente en cada simulación para que el ATE sea igual a 4, de acuerdo a la Ecuación 4.3.

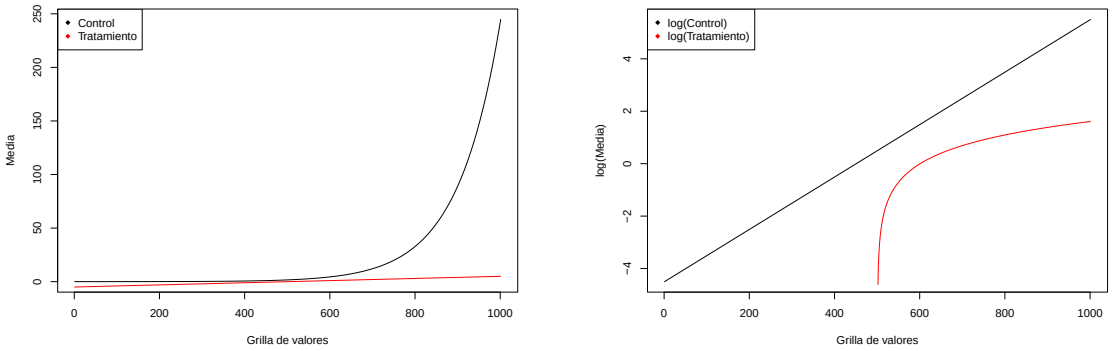
$$E[Y_1 - Y_0] = 4 \tag{4.3}$$

$$\begin{aligned} Y_0 &= \exp((X + W)\beta_B) + \varepsilon \\ Y_1 &= X\beta_B - \omega_B^s + \varepsilon \\ \varepsilon &\sim \mathcal{N}(0, 1) \end{aligned} \tag{4.4}$$

En el artículo original [17], la autora señala que el término ω_B^s se calcula de manera tal que el CATT (efecto causal en los individuos tratados) sea igual a 4. Sin embargo, a la hora de evaluar los modelos, compara sus estimaciones del ATE utilizando como valor teórico también 4. Esto resulta inconsistente, ya que los *datasets* fueron generados para que el CATT, y no el ATE, tome dicho valor. Por este motivo, en el presente trabajo se modificó el proceso generativo para que el ATE sea igual a 4, en línea con lo que se presume fue la intención original de la autora.

Tal como ocurrió en las superficies de tipo A, el vector β_B generado una única vez por cada dataset, de forma tal que el efecto causal sea aleatorio entre variables y generando una función distinta para el CATE en cada dataset.

Las Ecuaciones 4.4 fueron propuestas por la autora como un proceso generativo que permita simular simular **efectos heterogéneos no lineales** del tratamiento [17]. Para visualizar la relación entre las medias de los *outcomes* potenciales condicionales a X se realizó nuevamente un gráfico de medias en función de una grilla de valores factibles para la transformación de X. La Figura 4.3 muestra que las medias de las superficies de respuesta para individuos tratados y no tratados no son paralelas ni lineales.



(a) Media de la superficie B

(b) Logaritmo de la media de la superficie B

Fig. 4.3: Representación de la media para la superficie B. La línea roja representa la media de la distribución del outcome para los individuos tratados, es decir, $X\beta_B - \omega_B^s$, mientras que la línea negra representa la media de la distribución del outcome para el grupo de control, es decir $\exp((X+W)\beta_B) + \epsilon$.

5. RESULTADOS

En esta sección se implementa la evaluación de los modelos de Causal Machine Learning escogidos a través de las simulaciones basadas en *datasets* reales. Los *scripts* que generan los resultados y las Figuras están en el siguiente repositorio de GitHub: https://github.com/constanzadegalvagni/codigo_tesis. A lo largo de esta sección se hará referencia a archivos dentro de este repositorio.

5.1. Selección de hiperparámetros

Los modelos estudiados en esta sección tienen asociados distintos hiperparámetros, explicados en detalle en el capítulo 3. A través de *Randomized Search* se identificaron combinaciones óptimas para este caso de estudio.

Para elegir hiperparámetros con buena *performance* sobre superficies con distintos niveles de complejidad se seleccionaron diez superficies aleatorias generadas bajo las ecuaciones de las superficies de tipo A (4.2) y otras diez superficies aleatorias generadas bajo las ecuaciones de las superficies de tipo B (4.4). De esta manera, se evaluó el desempeño de 7 combinaciones aleatorias de hiperparámetros para cada modelo sobre estas 20 superficies promediando los resultados obtenidos por las métricas de interés para cada modelo, intentando comparar la *performance* en una cantidad justa de superficies y contemplando también la falta de reproducibilidad para el paquete que implementa BCF.

El código utilizado para este procedimiento se encuentra en el archivo `opt_hypers.Rmd`, aunque también se puede visualizar el *knit* desde `opt_hypers.html`, ya que su ejecución es notoriamente más lenta que la de los demás archivos. Para cada modelo se exploró un espectro amplio de combinaciones de hiperparámetros, desde opciones livianas y de bajo costo computacional hasta configuraciones más pesadas y demandantes, priorizando combinaciones con tiempo de ejecución para el ajuste de los datos menor a diez segundos.

Las combinaciones seleccionadas para cada modelo y las métricas que se utilizaron para evaluar estas decisiones se detallan a continuación.

Combinación seleccionada para BART

Para seleccionar la combinación de hiperparámetros con mejor *performance* para BART se evaluaron los criterios de convergencia introducidos en la Sección 3.2.3.

La combinación de hiperparámetros con mejor *performance* sobre estas métricas resultó ser la siguiente:

- `n.chains = 5`
- `n.burn = 50`
- `n.samples = 2000`
- `n.thin = 2`

Si bien al analizar los resultados se halló otra combinación con un mayor *effective size*, se encontró que la métrica de Gelman-Rubin fue casi igual para ambas combinaciones, y con un mejor desempeño temporal para la elegida. De todas formas, en este caso la métrica del promedio del *effective size* no es la mejor para comparar las combinaciones, porque depende del largo de las cadenas: en el caso de la combinación con mayor *effective size* se consideran 10 cadenas de 3000 iteraciones, lo que genera un modelo más lento con un desempeño que no supera al de la combinación ganadora. Es por esto que la métrica que motivó la decisión tomada es el diagnóstico de Gelman-Rubin.

Combinación seleccionada para BCF

Se seleccionó la combinación de hiperparámetros con mejor desempeño para los criterios de convergencia explicados en la Sección 3.3.2, considerando también un desempeño temporal que no resultara limitante.

Considerando estas métricas, la combinación seleccionada tiene los siguientes valores para los hiperparámetros:

- `num_gfr = 5`
- `num_burnin = 400`
- `num_mcmc = 400`
- `prognostic_forest_params$num_trees = 20`
- `treatment_effect_forest_params$num_trees = 5`

Resulta llamativo que la combinación seleccionada es una cadena con muy pocas iteraciones. El rango que se tuvo en cuenta estuvo dado por el vector `num_burnin <- c(100, 200, 400, 500, 1000, 2000)`. Tras un sampleo aleatorio para este valor, se elegía otro valor aleatorio del vector `proporciones_mcmc <- c(1,2,3,4)`, indicando qué largo tendría la cadena respecto al `num_burnin`: el largo de la cadena quedaba determinado por `num_burnin` $\times$ `proporciones_mcmc`. Habiendo incluso probado valores "largos" para las cadenas el valor ganador fue el de una cadena "corta".

Un análisis en mayor profundidad debería revisar, de todos modos, si efectivamente los resultados finales para BCF no se vieron afectados por la decisión de utilizar este tipo de cadena.

Combinación seleccionada para CF

Los modelos no bayesianos, si bien no requieren ejecutar métodos de Monte Carlo para realizar la inferencia, suelen contener una gran cantidad de hiperparámetros propios del modelo, que deben ser optimizados en términos de su desempeño. Por ese motivo, el *Causal Forest* fue evaluado por su error absoluto entre el efecto causal τ estimado y el τ verdadero (un punto que no se puede realizar en datos observables reales). Si bien todas las combinaciones presentaron valores de error muy similares, se priorizó la de mejor balance entre error y desempeño temporal. Los hiperparámetros que resultaron óptimos fueron:

- `num_trees = 5000`
- `sample_fraction = 0.3`
- `honesty.fraction = 0.5`

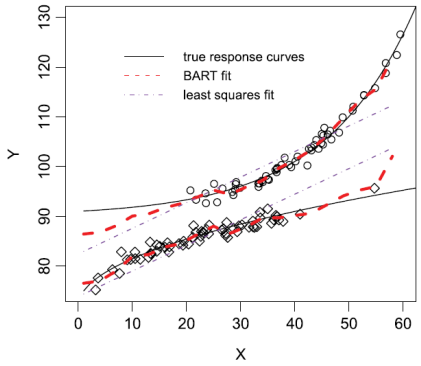
- `mtry = 12`
- `min.node.size = 1`
- `alpha = 0.005`

5.2. Resultados en simulación simple

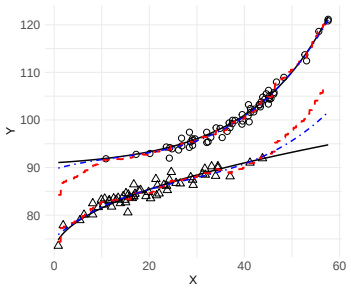
En la experimentación inicial con el *dataset* simple (4.1) el objetivo propuesto fue replicar los resultados obtenidos en el artículo original [17] para BART, comparando su desempeño frente a BCF y CF. El código utilizado para la experimentación y generación de figuras se encuentra en el archivo `experimentacion_inicial.Rmd`.

En el artículo original [17], la autora opta por mostrar el ajuste de BART en comparación con un polinomio de grado 1. Este ajuste resulta muy pobre para los datos propuestos, por lo que se decidió compararlo con un ajuste de un polinomio de mayor grado para visualizar una comparación más justa.

Los resultados se visualizan en la Figura 5.1. En esta visualización, si bien se evidencia la capacidad de BART para ajustarse al *outcome* Y de los tratados, también se observa que en el ajuste generado en este trabajo existe una tendencia del modelo a sobreestimar el resultado en la población de control, especialmente en la zona que la autora denomina *non-overlap* [17]: el área del eje X donde los grupos de tratamiento y control presentan menor intersección. En esa región, la predicción del *outcome* para los individuos no tratados se aproxima a la curva teórica de los tratados de manera muy similar al comportamiento del ajuste del polinomio de tercer grado.



Resultados obtenidos por Hill



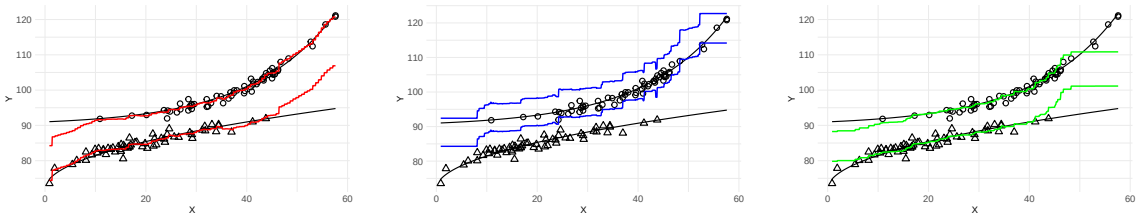
Ajustes generados en este trabajo

Fig. 5.1: Comparación de ajustes BART y polinomial. Los círculos representan a los individuos tratados, mientras que los rombos o triángulos representan al grupo de control. La línea roja representa el ajuste obtenido por BART, y la línea de puntos y rayas representa los ajustes polinomiales.

o Esta tendencia se debe a la ausencia de datos del grupo de control en regiones con X superior a 40, generando que el ajuste se acerque a la región del grupo tratado. Es interesante notar que, si bien ambos *datasets* provienen de las mismas ecuaciones generadoras, la aleatoriedad de la generación de datos produjo que en el *dataset* de Hill se genere un outlier en el grupo de individuos no tratados, quedando muy alejado en el eje X respecto a los demás. Por el contrario, en el *dataset* simulado para este trabajo el punto más extremo

del grupo de control se ubica a casi 10 unidades de distancia del *outlier* generado por Hill, mucho más cercano a los demás individuos del grupo de control. Esta disparidad en la distribución de los datos podría explicar el hecho de que en el ajuste BART del artículo original [17] la predicción para el grupo de control se acerca más a la curva teórica: el modelo se ajusta al punto aislado, generando un efecto visual de mejor ajuste para el grupo de control. En el caso de este trabajo, en cambio, al no contar con individuos de control a partir de cierto valor de X, se visualiza un peor ajuste a la curva teórica, aún cuando el modelo está comportándose correctamente dadas las características de los datos disponibles.

La comparación de los ajustes para BART, BCF y CF se visualiza en la Figura 5.2.



Ajuste BART Ajuste BCF Ajuste CF

Fig. 5.2: Comparación de ajustes BART, BCF y CF.

En el casos de CF y BCF, en contraposición con BART, se observa un aplanamiento del modelo en las zonas con poca densidad de datos. Este comportamiento puede interpretarse como una señal positiva en términos de sobreajuste, dado que BART tiende a hacerlo en ese tipo de regiones. No obstante, estos modelos tampoco logran replicar adecuadamente la forma de la curva teórica.

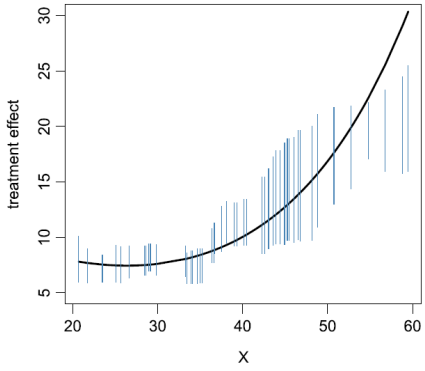
De todas formas, al tratarse de áreas donde hay pocos datos observados, es esperable que el modelo no se ajuste con precisión a la realidad. El aplanamiento puede explicarse por la naturaleza de los modelos basados en bosques: los puntos extremos son seleccionados con menor frecuencia en las particiones, lo que reduce la sensibilidad del modelo en esos rangos.

El sesgo observado en las curvas contrafactuales generadas por BCF podría deberse a características internas del paquete elegido. Los *outcomes* que se observan en la Figura 5.2 son los que devuelve el modelo, contrario al paquete utilizado para Causal Forest que permite reconstruír las curvas contrafactuales a través de la reparametrización 3.13. Esto no es posible con el modelo bayesiano, ya que al no devolver la estimación interna del *propensity score* π que realiza no es posible reconstruir los resultados a través de la reparametrización 3.6: la estimación de π que hace el modelo puede ser muy mala, lo que estaría perjudicando la estimación de los resultados.

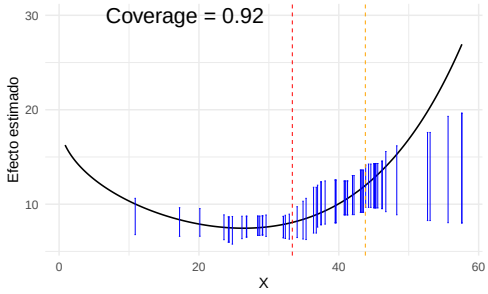
De todas formas, la diferencia entre curvas contrafactuales muestra una estimación del efecto causal similar a la obtenida para los demás modelos, lo que podría explicar que posteriormente se obtuvieron buenos resultados en la estimación del CATE para este modelo, a pesar de evidenciarse su mal desempeño para predecir los *outcomes*.

Por otro lado, se calcularon los intervalos de confianza de nivel 0.95 para las estimaciones de τ para cada uno de los modelos estudiados a partir de las estimaciones del desvío estándar

devueltas por cada uno de ellos. En la Figura 5.3 se visualizan los resultados obtenidos en comparación con los del artículo original [17] para BART. Si bien al tratarse de un *dataset* pequeño con regiones de bajo *overlap* el cubrimiento de los intervalos varía para cada situación, se observa un patrón similar para ambos ajustes. El tamaño de los intervalos de confianza aumenta considerablemente en las zonas de menor *overlap* entre grupos de tratamiento y de control, y más allá del punto con mayor valor de X el cubrimiento de los intervalos (*coverage*) comienza a decaer. Se observa también la diferencia en la distribución de los individuos tratados sobre el eje X entre los distintos *datasets*. La presencia del punto alejado sobre el eje X en los individuos no tratados que presenta el *dataset* de Hill conlleva a un leve sobreajuste de su modelo, mostrando en consecuencia un mejor *coverage*. Sólo se visualizan 4 intervalos que no cubren la curva teórica frente a los aproximadamente 60 que contiene el *dataset*.



Resultados obtenidos por Hill



Resultados del BART de este trabajo

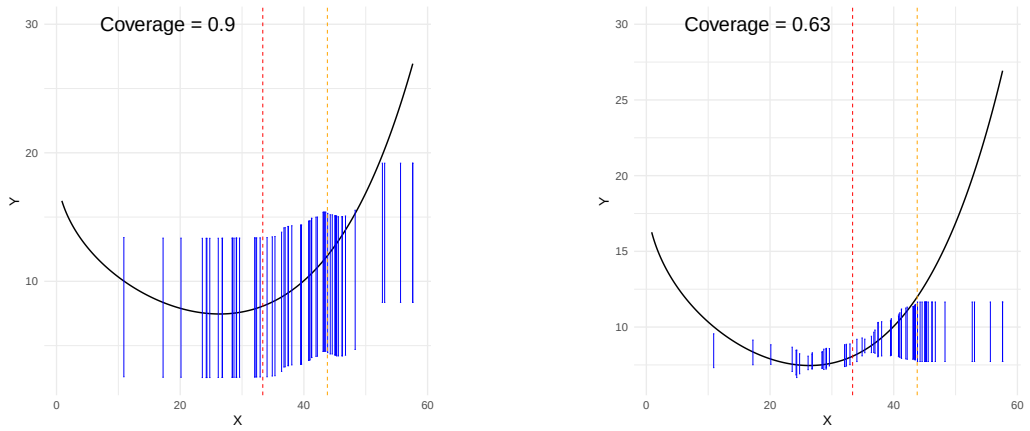
Fig. 5.3: Comparación de intervalos de confianza del ajuste BART obtenido por Hill y el de este trabajo. En negro se visualiza la curva teórica del τ, que resulta de restar las curvas teóricas para *outcome* bajo tratamiento y no-tratamiento. En rojo se visualiza el cuantil 0.95 en el eje X del grupo de control, y en naranja el punto más extremo de dicho conjunto.

Los resultados de los intervalos de confianza del τ para los otros modelos se visualizan en la Figura 5.4.

Dado que para BCF no se cuenta con reproducibilidad de los ajustes, para cada corrida del código se obtenía un resultado diferente de intervalos de confianza y *coverage*. Para visualizar un número que fuera más representativo del *coverage* real del modelo sobre estos datos, se decidió ajustar 10 BCFs diferentes sobre los datos y calcular, para todos ellos, el promedio de los intervalos de confianza y del cubrimiento final. Es por esto que en el gráfico no se visualizan intervalos de confianza, sino aproximaciones de los mismos.

El patrón observado para BART se mantiene en estos ajustes, con un peor *coverage* en la zona de menor *overlap*. Se observa una diferencia considerable entre el promedio de los tamaños para los intervalos de confianza generados por BCF y el de los intervalos de confianza generados por CF, con un mayor *coverage* en promedio para BCF con intervalos mucho más grandes a lo largo de todo el soporte, pero intervalos mucho más pequeños con menor coverage para CF.

Las predicciones para el valor de τ son más acertadas en el caso de CF en la región donde presenta mayor soporte de los individuos tratados, y para la zona con menor *overlap* el



Resultados para BCF Resultados para CF

Fig. 5.4: Comparación del promedio de intervalos de confianza para BCF e intervalos de confianza para CF. En negro se visualiza la curva teórica del τ, que resulta de la resta entre la curva teórica de tratados y no-tratados. En rojo se visualiza el cuantil 0.95 en el eje X del grupo de control, y en naranja el punto más extremo de dicho conjunto.

coverage de BCF es mayor. De todas formas, este *coverage* también conlleva a la generación de intervalos de confianza de mayor tamaño, por lo que es apresurado decir que tendría mejores predicciones para τ que CF.

En términos de tamaño de intervalos de confianza y *coverage*, se evidencia para el *dataset* simple y para la semilla utilizada que el mejor ajuste es BART. La incertidumbre de los intervalos de confianza no se dispara considerablemente, y el coverage del τ en las regiones con buen *overlap* es ampliamente satisfactorio.

5.3. Resultados en simulación IHDP

En este caso, se propone comparar la *performance* de los modelos elegidos para las superficies basadas en IHDP (4.2). Para esto fueron generadas 500 superficies de cada tipo a través de la variación de distintas semillas. En cada iteración se seleccionó un 70 % de la muestra como conjunto de entrenamiento y un 30 % como conjunto de testeo, de manera de poder evaluar el desempeño predictivo de los modelos, y con el objetivo de hallar, si existieran, diferencias en cuanto al desempeño en la estimación del CATE para individuos *out-of-sample*.

Notar que, cuando el interés detrás del uso de los modelos es estimar efectos causales, identificar *confounding* u otros sesgos estructurales, la separación entre conjunto de entrenamiento y testeo es poco relevante. Sin embargo, en este diseño, con datos simulados bajo ignorabilidad condicional y con estimandos conocidos, el *split* sí es informativo, ya que permite evaluar el desempeño de los métodos al contrastar sus estimaciones con los datos verdaderos [36].

El código utilizado en esta sección se encuentra en el archivo `experimentacion_ihdp.Rmd`, y el código que genera las Figuras se encuentra en `analisis_ihdp.Rmd`.

A continuación se detallan las métricas tenidas en cuenta. Para cada una de ellas, la nota-

ción m, d representa a cada modelo $m \in \{\text{BART, BCF, CF}\}$ entrenado con el subconjunto de datos de entrenamiento d_{train} seleccionado al azar sobre el *dataset* d.

- **$\widehat{\text{ATE}}_{m,d}$:** Se estimó el ATE para cada modelo y *dataset* promediando sobre las estimaciones individuales del CATE (5.1), donde la expresión x_i denota las covariables observadas para el individuo i y N representa el tamaño de la muestra total.

  La estimación del ATE se formaliza en la Ecuación 5.2, deducida a partir de la esperanza de la *adjustment formula* (2.12).

$$\widehat{\text{CATE}}_{m,d}(x_i) \overset{(2.23)}{=} \hat{Y}_{m,d}(Z=1, x_i) - \hat{Y}_{m,d}(Z=0, x_i) \tag{5.1}$$

$$\underbrace{\mathbb{E}_{m,d}[Y_1 - Y_0]}_{\text{ATE}} \overset{2,12}{\approx} \frac{1}{N} \sum_{i=1}^{N} \underbrace{\hat{Y}_{m,d}(Z=1, x_i) - \hat{Y}_{m,d}(Z=0, x_i)}_{\widehat{\text{CATE}}_{m,d}(x_i)} \tag{5.2}$$

- **Error($\widehat{\text{ATE}}_{m,d}$):** Se computó el módulo del error del ATE estimado en cada *dataset* d y cada modelo m respecto al ATE real, que tal como se mencionó en las Ecuaciones 4.2 y 4.4 equivale a 4 para todas las superficies simuladas.

$$\text{Error}(\widehat{\text{ATE}}_{m,d}) = |\widehat{\text{ATE}}_{m,d} - 4| \tag{5.3}$$

- **RMSE($\widehat{\text{ITE}}$):** Como se explicó en la Sección 2.5, el CATE será utilizado como un estimador del ITE dado el supuesto de ignorabilidad condicional (2.24). Dado que cada modelo devuelve múltiples estimaciones individuales del efecto causal, ya sea a través de las cadenas de Markov paralelas en el caso de BART o de las estimaciones individuales de los distintos árboles para BCF o CF, el promedio de estas estimaciones para cada individuo constituye un predictor del ITE.

  Para cada *dataset* d y modelo m se calculó la raíz cuadrada del error cuadrático medio (RMSE) obtenido para las predicciones del ITE de los individuos, comparando cada predicción con el ITE real al cual se tiene acceso por tratarse de datos simulados. El resultado se expresa en la Ecuación 5.4, donde x_i representa a las covariables observadas para el individuo i, y donde $d_c \in \{$ train, test $\}$ dado un dataset d.

$$\text{RMSE}(\widehat{\text{ITE}}_{m,d}(i \in d_c)) = \sqrt{\frac{1}{N_{d_c}} \sum_{i \in d_c} (\widehat{\text{CATE}}_{m,d}(x_i) - \text{ITE}(i))^2} \tag{5.4}$$

- **Coverage del ITE:** Como se explicó en el capítulo de documentación de modelos (3), cada una de las implementaciones utilizadas devuelve estimaciones del desvío estándar para las estimaciones individuales del CATE. Teniendo en cuenta también al CATE como estimador del ITE es posible calcular intervalos de confianza para el ITE, los cuales en este caso fueron de nivel 0.95.

  El *coverage* fue calculado como la proporción de intervalos que contienen el valor real del ITE, como se expresa en la Ecuación 5.5, donde $I(\widehat{\text{CATE}}_{m,d}(x_i))$ representa

el intervalo de predicción de nivel 0.95 para el ITE con las covariables observadas x_i bajo el modelo m y el *dataset* d.

Se tuvo en cuenta esta métrica para las estimaciones del CATE en los individuos del conjunto de entrenamiento y para el conjunto de testeo de cada *dataset*, lo que se representa con el subíndice d_c de la Ecuación 5.5.

$$\text{Coverage}_{m,d}(\text{ITE}(i \in d_c)) = \frac{1}{N_{d_c}} \sum_{i \in d_c} \mathbb{I}[\text{ITE}(i) \in I_{m,d}(\widehat{\text{CATE}}(x_i))] \tag{5.5}$$

El ITE es una variable aleatoria, por lo que para predecirlo deberían usarse intervalos de predicción para el ITE. Sin embargo, por la forma en que fueron generados los contrafactuales en este trabajo, los errores se cancelan y el ITE de cada individuo coincide con su CATE, por lo que sería válida la utilización de intervalos de confianza del CATE como intervalos de predicción del ITE.

Por otro lado, a lo largo del trabajo se hace un abuso de notación para los modelos bayesianos, ya que estos en lugar de contar con intervalos de confianza cuentan con **intervalos de credibilidad**. Es importante tener esto en cuenta cuando se mencionen intervalos de confianza para este tipo de modelos.

- **Tamaño medio de los intervalos de confianza:** Para los intervalos de confianza del ITE generados se calculó la media de la diferencia entre el límite superior y el límite inferior, tal como se formaliza en la Ecuación 5.6. Esta métrica fue calculada tanto para el conjunto de entrenamiento como para el de testeo para cada *dataset*, lo que se denota con el subíndice d_c.

$$\overline{I_{m,d}}(\text{ITE}_{m,d}(i \in d_c)) = \frac{1}{N_{d_c}} \sum_{i \in d_c} \hat{I}^{\text{sup}}_{m,d}(\text{ITE}(i)) - \hat{I}^{\text{inf}}_{m,d}(\text{ITE}(i)) \tag{5.6}$$

- **Tiempo:** Se midió el tiempo de entrenamiento de cada modelo en segundos.

Adicionalmente se computaron métricas inherentes al desempeño de las cadenas MCMC, en los casos que correspondiera:

- Tanto para BART como para BCF se calculó el *effective size* de la/s cadenas generadas, para evaluar la convergencia. Se evaluó tanto el número absoluto como relativo al largo de la/s cadenas generadas.
- Para BART se calculó el diagnóstico de Gelman-Rubin, como métrica adicional de convergencia.
- Para BCF se calculó la autocorrelación de la cadena generada por el ajuste.

5.3.1. Estimación del ATE

En cada superficie se representó la distribución del error absoluto en la estimación del ATE, tal como se define en la Ecuación 5.3. Los resultados para las superficies de tipo A y B se muestran en las Figuras 5.5 y 5.6, respectivamente. Cada punto de los gráficos representa la métrica calculada para un dataset, por lo que las zonas más "anchas" del eje

X representan mayor frecuencia para el respectivo valor en el eje Y, tal como sucede en los histogramas o gráficos de densidad.

No se observaron diferencias relevantes para esta métrica entre los conjuntos de entrenamiento y de testeo de cada *dataset*, por lo que los gráficos representan el resultado de esta métrica considerando todos los puntos, sin discriminación entre entrenamiento y testeo.

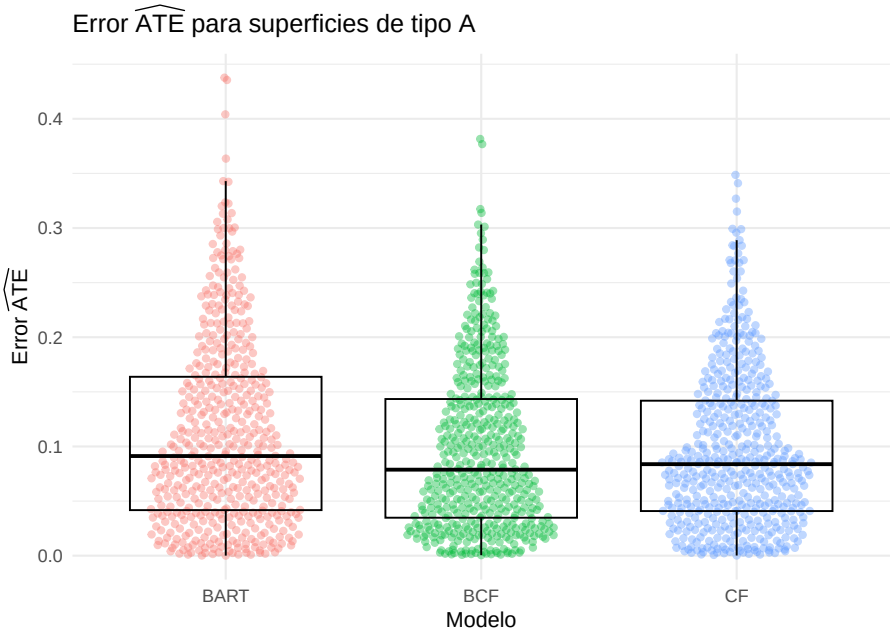


Fig. 5.5: Distribución del error en la estimación del ATE para las superficies de tipo A

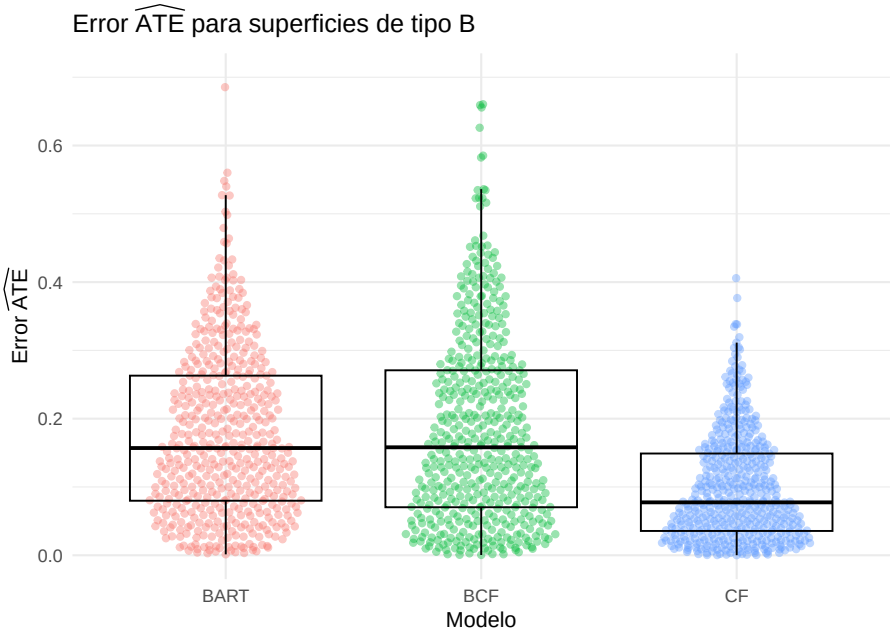


Fig. 5.6: Distribución del error en la estimación del ATE para las superficies de tipo B.

En el caso de las superficies de tipo A, el comportamiento del error es muy similar entre los tres modelos evaluados. Para las superficies de tipo B, en cambio, se aprecia que Causal Forest alcanza errores de estimación del ATE notablemente menores que los modelos bayesianos, los cuales presentan un desempeño muy parecido entre sí.

Cabe destacar que, al considerar el cambio de escala en el eje Y entre la Figura 5.5 y la Figura 5.6, se observa que la mediana del error en los modelos bayesianos prácticamente se duplica en las superficies de tipo B respecto a las de tipo A, mientras que el error en Causal Forest se mantiene estable entre ambos tipos de superficies.

5.3.2. Predicción del ITE

Se calculó el RMSE del $\widehat{\text{ITE}}$ (5.4) para los conjuntos de entrenamiento y de testeo de cada *dataset*.

Los resultados para las superficies de tipo A se visualizan en la Figura 5.7.

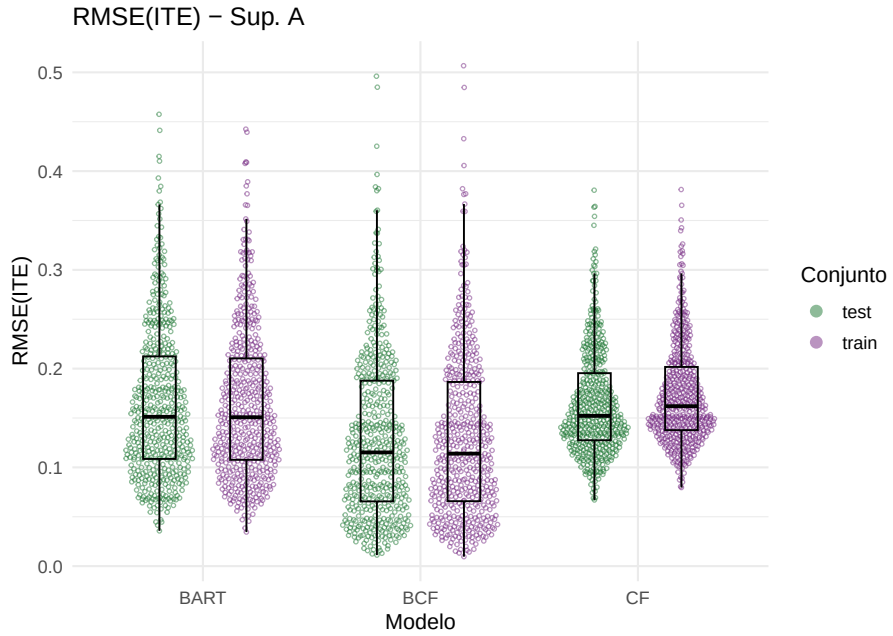


Fig. 5.7: Distribución del RMSE de la estimación del ITE para las superficies A.

Las distribuciones del RMSE no presentan diferencias considerables entre el conjunto de entrenamiento y testeo para los tres modelos, por lo que no se observa evidencia de sobreajuste para este tipo de superficies.

Al comparar entre modelos, BART y CF presentan medianas muy similares con cajas levemente más pequeñas para CF, el cual también presenta menor dispersión de los puntos. El RMSE del ITE en Causal Forest presenta una variabilidad levemente menor a la observada para los demás modelos.

BCF presenta una mediana ligeramente inferior que los otros dos modelos y se visualizan datasets cuyo RMSE(ITE) es menor a 0.05, valores que no fueron alcanzados por los otros modelos. Sin embargo, la longitud de las colas superiores y el solapamiento entre las cajas

para todos los modelos indica que la ventaja del desempeño de BCF es reducida para las superficies con efectos causales homogéneos.

Los resultados del RMSE de la predicción del ITE para las superficies de tipo B se visualizan en la Figura 5.8.

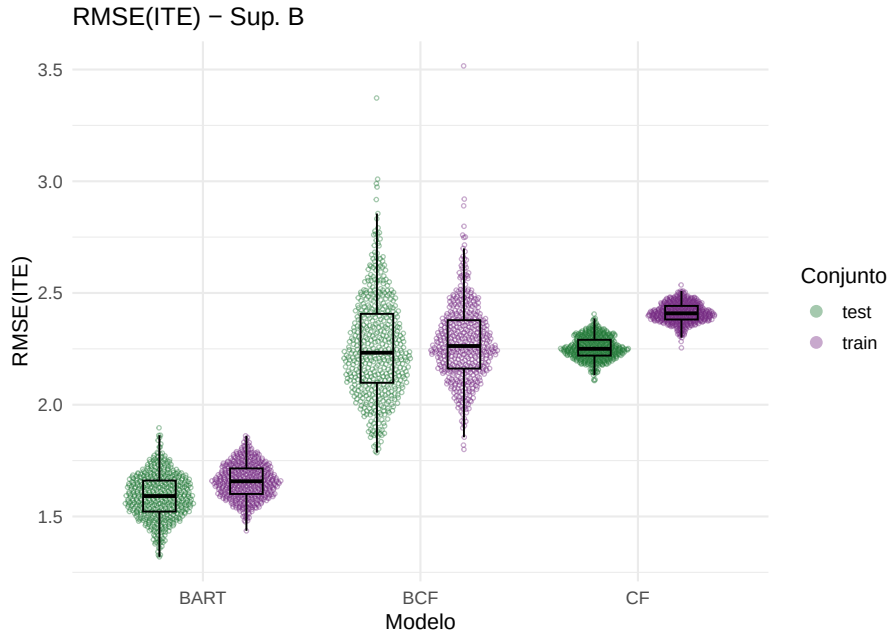


Fig. 5.8: Distribución del RMSE de la estimación del ITE para las superficies B

Para este tipo de superficies es notorio el cambio de escala en el eje Y, evidenciando que el escenario de efectos causales heterogéneos es más complejo para todos los modelos en cuanto a la predicción del ITE. En cuanto a la comparación entre modelos, BART muestra la menor mediana y una dispersión moderada, lo que indica mejor desempeño. BCF y CF tienen errores mayores que BART. En el caso de BCF, su mediana es cercana a CF pero muestra una varianza bastante mayor, con colas superiores largas (*outliers* altos), lo que evidencia un mal desempeño en cuanto a sesgo y varianza. Por su parte, CF tiene una distribución más concentrada pero centrada en valores más altos que BART, indicando menor varianza pero mayor sesgo que BART.

En los tres modelos se observa que las diferencias entre los conjuntos de entrenamiento y testeo son prácticamente nulas. No se evidencian señales de sobreajuste, aunque resulta llamativo que para CF el error del conjunto de testeo sea casi 0.5 unidades menor que el del conjunto de entrenamiento.

Notar que, si bien al calcular el error de estimación del ATE las medianas se mantuvieron por debajo del 0.2, el error de estimación del ITE es mucho mayor para las superficies de tipo B. Esto pareciera ser contraintuitivo, ya que el ATE se calcula como un promedio de las predicciones del ITE (5.2). En la Sección 5.3.5 se propone una explicación para este comportamiento.

5.3.3. Análisis del tamaño de intervalos de confianza

Se realizó un análisis de la distribución del tamaño de los intervalos de confianza para las predicciones del ITE (5.6).

Para las superficies de tipo A los resultados se visualizan en la Figura 5.9.

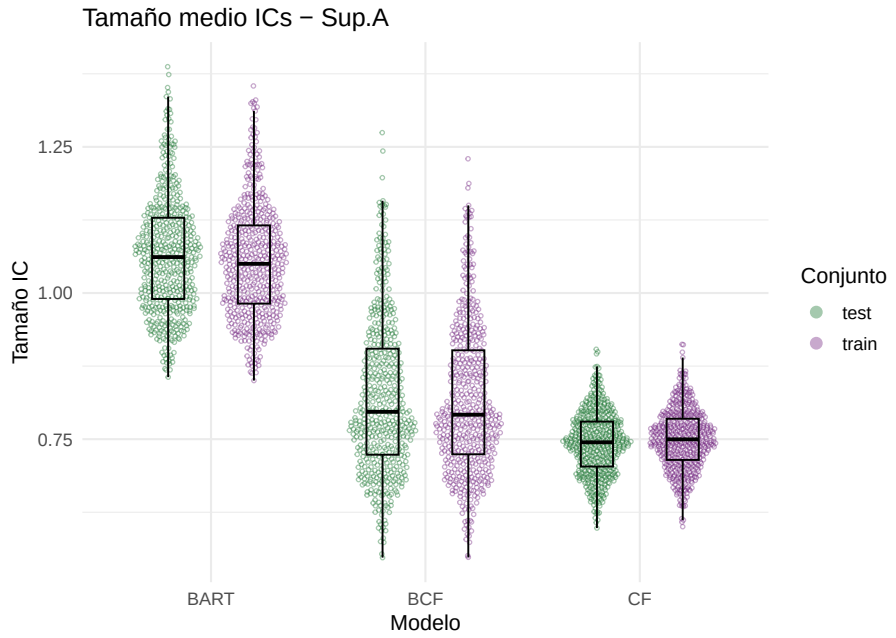


Fig. 5.9: Distribución del tamaño de los intervalos de confianza para el CATE generados por los distintos modelos para las superficies de tipo A, discriminados según conjunto.

En todos los modelos se observan diferencias pequeñas entre los conjuntos de entrenamiento y testeo, indicando una cuantificación de la incertidumbre estable para los distintos tipos de conjuntos.

En cuanto a comparaciones entre modelos, el hecho de que BART muestre los intervalos más largos para este tipo de superficie apunta a una mayor incertidumbre en la estimación de los efectos individuales, lo cual también debería corresponder a un mayor cubrimiento del valor real.

BCF y CF generan intervalos más cortos que BART, pero BCF presenta mayor variabilidad en el tamaño de los intervalos. Esto sugiere que en la mayoría de los puntos el modelo acota la incertidumbre, pero en ciertas regiones pareciera generar intervalos más grandes en un intento de obtener un mejor cubrimiento.

En la sección siguiente se realizará un análisis del cubrimiento de los intervalos del ITE, donde se verificará si efectivamente los modelos bayesianos presentan un mayor cubrimiento del ITE que Causal Forest, que generó los intervalos más cortos entre los tres modelos.

Se visualiza la dispersión del tamaño de los intervalos de confianza del ITE para las superficies de tipo B en la Figura 5.10.

Se observa un considerable cambio de escala en el eje Y respecto a las superficies de tipo A, en especial para los modelos bayesianos. BART presenta una dispersión más moderada que

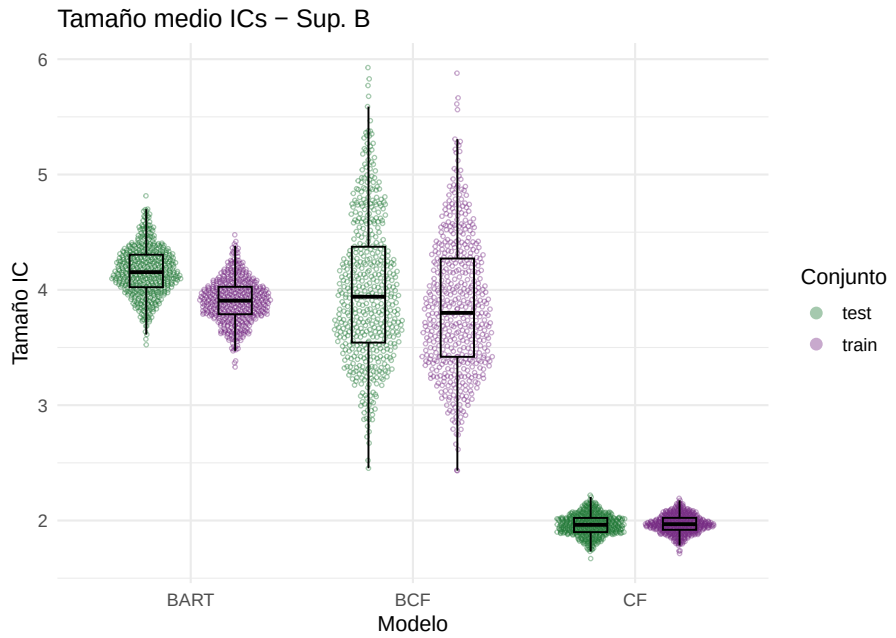


Fig. 5.10: Distribución del tamaño de los intervalos de confianza para el ITE generados por los distintos modelos para las superficies de tipo B.

BCF, representando una cuantificación más conservadora y estable de la incertidumbre que BCF. Este último, si bien presenta intervalos ligeramente más angostos que BART en mediana, presenta mayor variabilidad y colas más largas, indicando regiones donde el modelo percibe poca información y ensancha fuertemente los ICs y otras donde estima el valor con mayor certeza.

En cuanto a CF, sus intervalos de confianza son más cortos que para los modelos bayesianos. Sin embargo, como se observó en la Figura 5.8, el modelo presenta un RMSE alto para la estimación del ITE, lo que indica que el modelo está subestimando la incertidumbre de las predicciones con intervalos cortos que probablemente no cubran el valor real del ITE.

5.3.4. Análisis del coverage del ITE

Los resultados para las superficies de tipo A se visualizan en la Figura 5.11.

.

Para los modelos bayesianos se visualiza una cobertura muy cercana a 1 para casi todos los *datasets*, representando una sobrecobertura sistemática, ya que se esperaría ver valores más cercanos a 0.95 por tratarse de intervalos de confianza de ese nivel. Sin embargo, como se vio en las dos secciones anteriores, estos modelos presentan intervalos de confianza más amplios que CF, por lo que es esperable que se cubran la mayor parte de las estimaciones a través de intervalos grandes.

Los intervalos de Causal Forest presentan un cubrimiento más cercano a 0.95, pero mayor dispersión y múltiples casos de cobertura menor a los demás modelos. Esta menor cobertura coincide con los intervalos más cortos generados por este modelo (Figura 5.9).

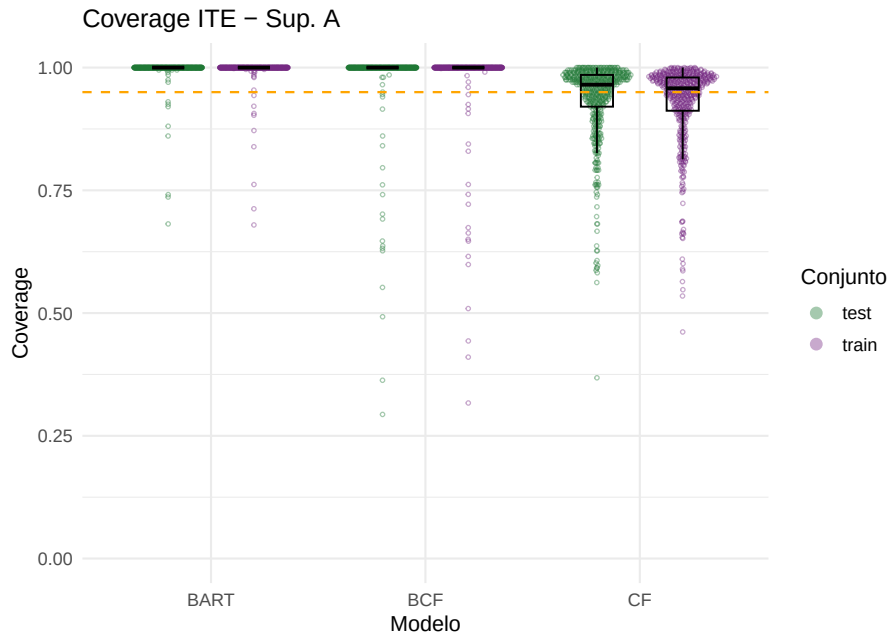


Fig. 5.11: Distribución del *coverage* del ITE para las superficies de tipo A. La línea punteada naranja indica el valor 0.95, correspondiente al nivel de los intervalos de confianza calculados.

Por su parte, los resultados para las superficies de tipo B se visualizan en la Figura 5.12.

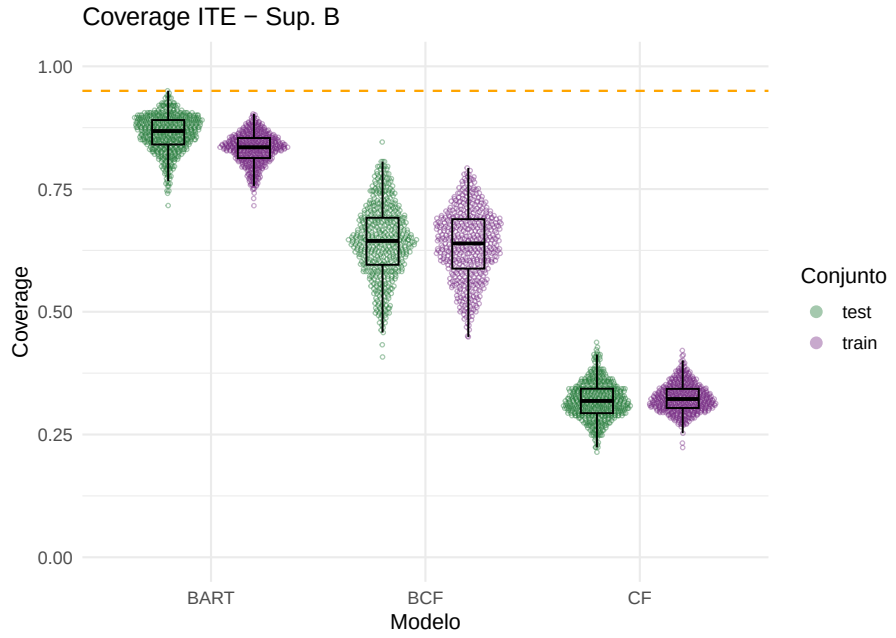


Fig. 5.12: Distribución del coverage del ITE para las superficies de tipo B. La línea punteada naranja indica el valor 0.95, correspondiente al nivel de los intervalos de confianza calculados.

Es notorio que ningún modelo alcanza el 0.95 de cobertura, en contraposición con los datos observados para las superficies de tipo A. El *coverage* de todos los modelos se posiciona

por debajo de la línea del 0.95, indicando una sub-cobertura de los intervalos de confianza.

BART ofrece la mayor cobertura relativa, con la mayoría de puntos para los conjuntos de entrenamiento y testeo ubicados por encima del 0.8, mucho más cerca del nivel deseado que los demás modelos, además de que exhibe distribuciones menos dispersas del cubrimiento que los demás. Por su parte, BCF muestra una cobertura intermedia, con mayor dispersión. Esta variabilidad también es consistente con lo mencionado en secciones anteriores: las estimaciones individuales del modelo parecieran depender fuertemente de las características de las regiones del espacio de covariables, presentando regiones con mucho mejor desempeño en todas las métricas que otras. Causal Forest, por último, muestra la menor cobertura, a la vez que muestra una menor dispersión, indicando que la subcobertura es sistemática. Para este modelo la estimación de la incertidumbre no se adapta a la dificultad de este tipo de superficies.

5.3.5. Análisis entre *coverage* y predicciones del ITE

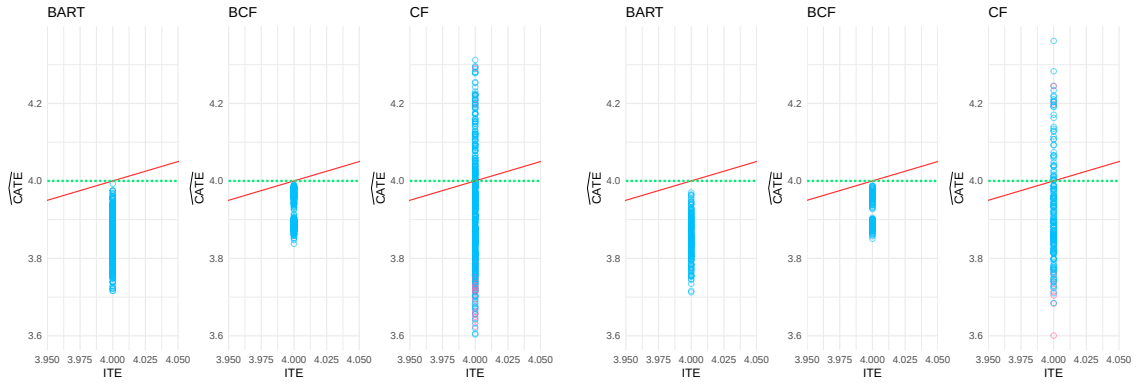
Con el fin de evaluar la calidad de las predicciones del efecto individual del tratamiento (ITE), se construyeron gráficos que comparan los valores verdaderos con sus predicciones mediante el CATE en una superficie específica generada a partir de una semilla aleatoria. La lógica detrás de este análisis es que el conjunto de covariables se mantiene constante para la generación de *outcomes*, mientras que las semillas se modifican únicamente para muestrear distintos resultados a partir de la misma distribución de probabilidad. De este modo, se espera que el comportamiento de las predicciones individuales sea consistente entre semillas.

El gráfico relaciona estas variables en un *scatterplot*, de forma tal que las predicciones ideales deberían estar sobre la diagonal, y los colores de los puntos varían según el coverage alcanzado para la estimación individual del efecto causal: el celeste indica puntos cuyo intervalo de confianza cubre el valor real del ITE, y en caso contrario se colorean en fucsia.

El resultado de este gráfico para las superficies de tipo A se visualiza en la Figura 5.13. El ITE teórico siempre equivale a 4 por tratarse de superficies tal que el efecto causal es homogéneo para todos los individuos, mientras que para los distintos modelos se visualizan distintos niveles de dispersión.

Tal como se vio en las Figuras 5.7, BCF resulta ser el modelo que alcanza menor RMSE para este tipo de superficies, seguido por BART y luego por CF. El resultado también es consistente con la Figura 5.11: no se visualizan puntos fucsias para los modelos bayesianos, pero sí para CF.

La dispersión en el eje Y de la nube de puntos sugiere qué valor tendrá la estimación del ATE, ya que con este promedio se realiza la estimación. Si bien la nube de puntos de los modelos bayesianos está por debajo del valor del ATE teórico y Causal Forest está centrado en este valor, en la Figura 5.5 se vio que el error de estimación para el ATE era muy similar en todos los modelos. Esto ocurre porque las estimaciones de CF tienen mayor varianza que los modelos bayesianos, compensando a través del promedio el sesgo que presentan estos modelos. El hecho de que todas las predicciones realizadas por los modelos bayesianos se encuentren por debajo del valor real muestra el sesgo que presentan este tipo de modelos, patrón que también se observa sistemáticamente en las competencias ACIC [3] [18].

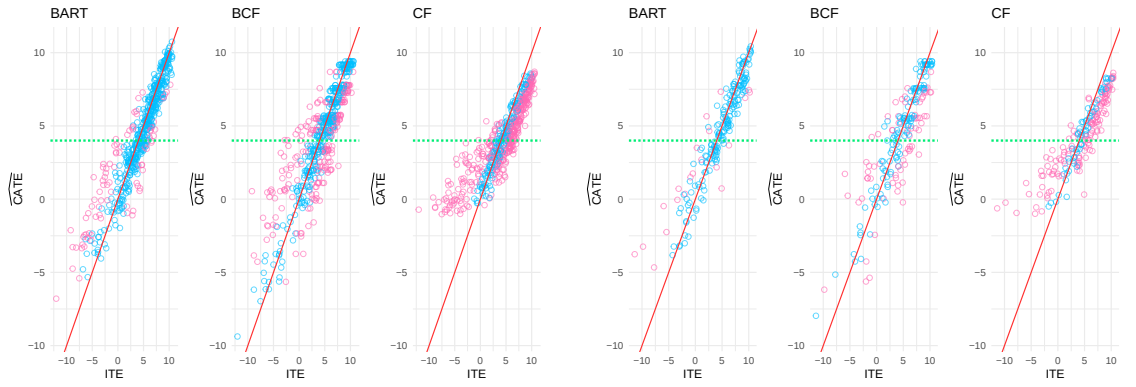


Conjunto de entrenamiento Conjunto de testeo

Fig. 5.13: Comparación entre ITE verdadero y su estimación a través del CATE para las superficies A. La diagonal roja representa la relación ideal entre el valor y su estimación, la línea punteada verde representa el ATE teórico = 4. Los puntos azules son las predicciones cuyo IC contiene el ITE real y los puntos fucsias son los puntos que no alcanzan a cubrir el ITE.

Por último, no se observan diferencias significativas entre los conjuntos de entrenamiento y de testeo, coincidiendo con el patrón observado para todas las métricas analizadas. Esto es una buena señal en términos de que no se aprecia sobreajuste para ningún modelo.

El gráfico análogo para las superficies de tipo B se visualiza en la Figura 5.14. En este caso, el panorama es diferente: el ITE no toma siempre el mismo valor, ya que este caso corresponde al escenario de efectos heterogéneos. Nuevamente los patrones observados no varían entre el conjunto de entrenamiento y el de testeo, coincidiendo con las métricas medidas para estas superficies.



Conjunto de entrenamiento Conjunto de testeo

Fig. 5.14: Comparación entre ITE verdadero y su estimación a través del CATE para las superficies B. La diagonal roja representa la relación ideal entre el valor y su estimación, la línea punteada verde representa el ATE teórico = 4. Los puntos azules son las estimaciones cuyo IC contiene el ITE real y los puntos fucsias son los puntos que no alcanzan a cubrir el ITE.

Se evidencia que el modelo con mejor desempeño predictivo resulta ser BART por ser

el modelo con puntos más cercanos a la diagonal, además de tener claramente mayor proporción de puntos celestes, indicando individuos cuyo intervalo de confianza para la estimación del CATE cubre al ITE verdadero.

Para BCF se observan regiones de puntos fucsias que si bien están cercanos a la diagonal no alcanzan a cubrir el valor real, lo cual se corresponde con los intervalos de confianza "cortos" observados en la Figura 5.10. Se visualizan también zonas donde las predicciones están notoriamente alejadas del valor real, en algunos casos hasta a más de cinco unidades de distancia, lo cual coincide con la dispersión observada para el RMSE en la Figura 5.8.

Para CF, por su parte, las zonas de bajo cubrimiento son visiblemente más grandes que las que cubren el valor verdadero, tal como se había notado también en la Figura 5.12. Incluso se visualizan valores cercanos a la diagonal cuyo intervalo de confianza fue demasiado corto y no alcanzó a cubrir el valor verdadero. Esto demuestra que los intervalos cortos visualizados en la Figura 5.10 están subestimando el error de la estimación, lo cual también se había evidenciado al medir el coverage. La concentración de la nube de puntos en una zona más reducida que los demás coincide con el patrón de RMSE observado en la Figura 5.8, donde se observó una tendencia del modelo a subestimar sistemáticamente los efectos individuales heterogéneos. El modelo presenta un bajo nivel de generalización para los valores más extremos del ITE, mostrando un peor poder de estimación para los valores de ITE más alejados del 0. Parecieran existir zonas donde el modelo está sesgado a sobreestimar el efecto causal (izquierda de la diagonal) y zonas donde, si bien la estimación no es lejana a la real, el tamaño del intervalo de confianza no logra cubrir el valor real (derecha de la diagonal).

5.3.6. Análisis de criterios de convergencia

Para el caso de BART, se visualiza la convergencia de las cadenas MCMC generadas en la Figura 5.15.

En líneas generales, la convergencia observada es muy buena, con gran densidad de valores cercanos al 1 para ambas superficies y conjuntos. Si bien para las superficies de tipo B se visualiza una mayor dispersión, no dejan de ser valores "buenos" de la métrica.

Para BCF se visualiza la convergencia de la cadena en la Figura 5.16. Para este modelo se observa una métrica de convergencia muy pobre: La totalidad de los valores es superior al 0.6, un valor ya de por sí alto para esta métrica, indicando que las cadenas generadas no aportan nueva información en cada sampleo y no alcanzan a explorar satisfactoriamente el espacio. Es importante aclarar que, de todos modos, los hiperparámetros seleccionados fueron elegidos específicamente por tener el mejor desempeño en cuanto a métricas para la cadena de Markov, por lo que en general el desempeño obtenido en cuanto a cadenas MCMC fue muy pobre para este modelo. De todas formas, se destaca que a pesar de generar "malas" cadenas su desempeño predictivo no fue despreciable respecto a los demás modelos.

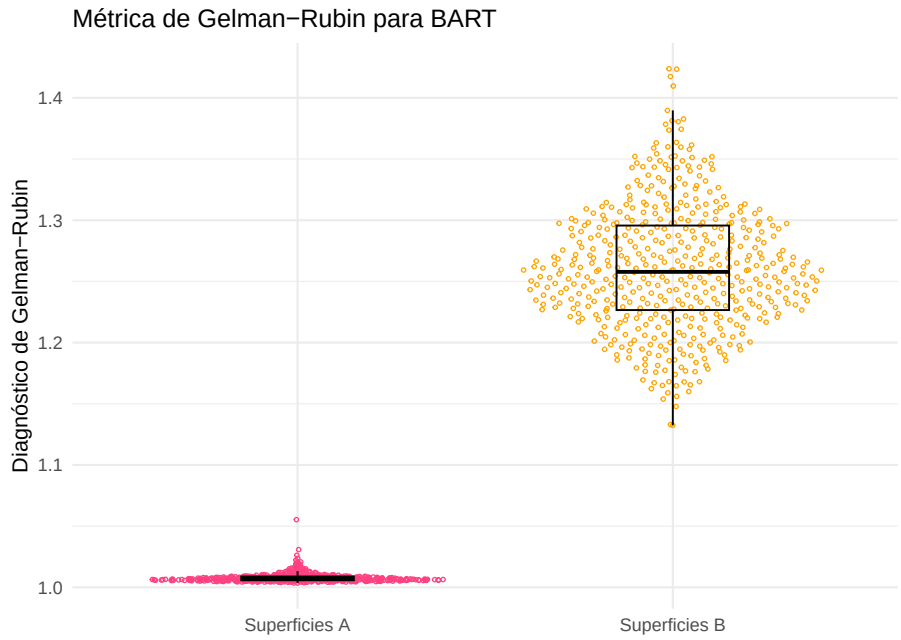


Fig. 5.15: Distribución del estadístico de Gelman-Rubin para las cadenas de Markov generadas por BART.

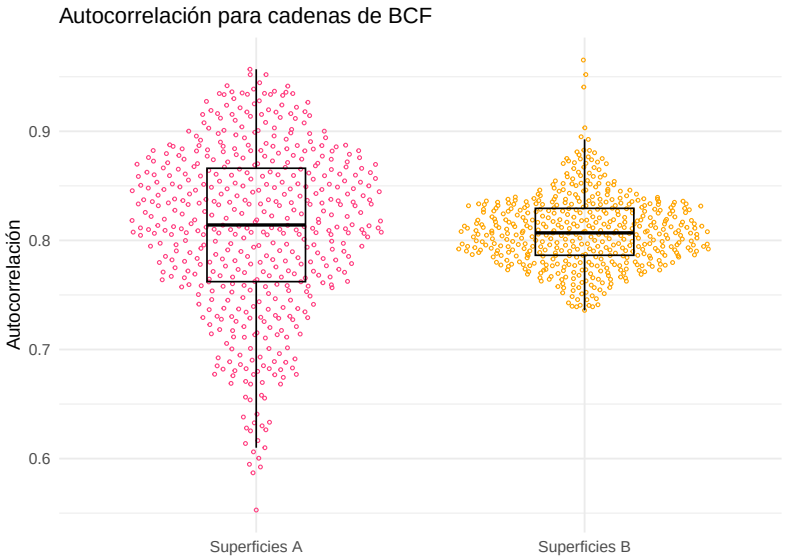


Fig. 5.16: Distribución de la autocorrelación, en promedio, de la cadena de Markov generada por cada BCF.

5.3.7. Discusión de resultados en IHDP

Para hablar de los resultados visualizados en esta sección es necesario remontarse a la primera etapa de la experimentación: la elección de hiperparámetros. En esa instancia, se eligió a los hiperparámetros de los modelos bayesianos según la convergencia de cadenas MCMC generadas, y a los de Causal Forest según su RMSE en la predicción del ATE.

Teniendo en cuenta que los hiperparámetros de los modelos bayesianos fueron elegidos únicamente por criterios de convergencia, es destacable la capacidad de éstos para estimar los efectos causales tanto individuales como muestrales. La *performance* superior de BART sobre BCF puede deberse al mejor sampleo de las cadenas MCMC generadas (Figura 5.15), y el sesgo observado en la Figura 5.14 para BCF puede deberse a la mala convergencia de las cadenas, llevando a la baja generalización del modelo en ciertas regiones del espacio.

Se atribuye la baja *performance* observada en todas las métricas estudiadas para BCF a la baja calidad del paquete utilizado: la falta de posibilidad de generar más de una cadena MCMC, la ausencia de hiperparámetros que muestren una convergencia aceptable y la nula documentación del paquete para calcular el *propensity score* son algunas de las razones por las que se atribuye el mal desempeño al paquete y no al modelo en sí, que como se explicó anteriormente es uno de los más aceptados actualmente para la estimación de efectos heterogéneos. Era esperable que este modelo tuviera mejor desempeño que sus competidores, al ser el ganador de las últimas competencias ACIC.

El bajo cubrimiento de los Causal Forests (Figuras 5.11, 5.12), así como el sesgo observado para ciertas regiones de las estimaciones individuales bajo el escenario de efectos causales heterogéneos (Fig 5.14) puede deberse a una bajo poder de generalización entre los árboles generados que no alcanza a cubrir la heterogeneidad real del espacio, como sí lo hicieron los modelos bayesianos.

6. CONCLUSIONES

6.1. Discusión acerca del objetivo de la tesis

El objetivo principal de este trabajo fue estudiar el área de inferencia causal replicando y expandiendo los hallazgos del artículo de Hill (2011) [17], a través del análisis de la estimación del efecto causal no sólo del modelo BART, protagonista del artículo original, sino también de Bayesian Causal Forest (BCF) y Causal Forest (CF), considerados actualmente como parte del estado del arte en inferencia causal.

Previo a la implementación de modelos sobre distintos *datasets*, se llevó a cabo una recolección exhaustiva de bibliografía y artículos en el tema, para tener una base sólida del estado del arte en cuanto a modelos y métricas a considerar. Fue en esta instancia de la tesis donde se comenzó a profundizar en el trabajo de Hill, no sólo en su artículo de 2011, que fue una gran base para la misma, sino también con las competencias ACIC organizadas posteriormente e impulsadas principalmente por ella, cuyo objetivo era evaluar modelos de inferencia causal en *datasets* con características diversas, propósito que también guía el presente trabajo.

Toda esta información fue clave para construir el marco teórico, definir qué modelos se utilizarían y sobre qué *datasets* teniendo en cuenta el poder de cómputo y restricciones temporales existentes.

En este sentido, se consiguió estudiar todos los modelos seleccionados. Además, se efectuó un análisis sistemático de distintas combinaciones de hiperparámetros para maximizar su desempeño en la estimación tanto del efecto causal promedio como del individual, un aspecto que no había sido abordado en el artículo original [17].

Los modelos fueron puestos a prueba sobre escenarios desde más sencillos hasta más complejos, comenzando por un pequeño *dataset* simulado con una sola covariable, hasta la simulación de superficies de respuesta con menor y mayor complejidad del efecto del tratamiento, cubriendo un espectro de complejidades factible para las condiciones de realización del trabajo.

Se pusieron a prueba utilizando implementaciones con soporte y aceptación en la comunidad académica, documentando previamente los modelos a estudiar y los paquetes tenidos en cuenta.

En el contexto de los entornos experimentales considerados, se generaron estimaciones del efecto promedio y predicciones del efecto individual del tratamiento, verificando el desempeño predictivo de cada modelo, sus debilidades y sus fortalezas ante distintos *ground-truth*. Para medir el desempeño, se tuvieron en cuenta métricas como el error absoluto de las estimaciones del ATE y el RMSE de las predicciones del ITE, así como el *coverage* de los intervalos de confianza, una métrica ampliamente utilizada en este campo.

Dentro de los márgenes de factibilidad, se puede concluir que este trabajo ofrece un análisis ordenado y técnicamente fundamentado de modelos de inferencia causal que hoy ocupan un lugar central en la literatura.

6.2. Trabajo a futuro

Esta tesis deja abiertas diversas posibilidades para profundizar el análisis y ampliar los alcances de las conclusiones obtenidas. Si bien se logró cumplir con los objetivos planteados dentro de los márgenes de factibilidad del proyecto, existen aspectos metodológicos y experimentales que podrían enriquecerse en futuras investigaciones.

Sobre el modelo BCF y su implementación

Una de las principales limitaciones estuvo dada por la implementación del modelo Bayesian Causal Forest utilizada (3.3.1). En este trabajo se optó por una versión simplificada del modelo, con menor flexibilidad en términos de parámetros y sin acceso directo a funciones de ajuste fino que están presentes en el artículo original del modelo [28].

Si bien el paquete utilizado fue desarrollado por los mismos autores, es posible que la implementación del paquete `bcf` permita explorar mucho mejor el potencial del modelo, especialmente en la estimación de efectos heterogéneos, ya que este paquete sí permite la generación de múltiples cadenas MCMC. Se atribuye el bajo desempeño de BCF a las pocas cadenas generadas, en contraposición con BART, que con más cadenas paralelas obtuvo mejores estimaciones.

Por otro lado, la documentación del paquete no especifica cómo se realiza la estimación interna del *propensity score* (que, como se mencionó anteriormente, no debería ser realizada por el modelo tal como fue planteado originalmente, sino ser recibida como un parámetro de entrada, pero el paquete tampoco permite esto). La estimación generada podría ser muy pobre, lo que opacaría la ventaja de contar con el *propensity score* poniendo al modelo por detrás de BART que no cuenta con esta característica. Esto es contrario al resultado observado en las competencias ACIC, donde BCF tenía una notoria ventaja sobre BART en la estimación de efectos causales.

Sobre la ampliación de *datasets* y escenarios

Una línea natural de extensión consiste en evaluar los modelos en un mayor número y variedad de *datasets*.

En particular, aplicar los modelos a datos observacionales reales (más allá de simulaciones controladas) permitiría observar su comportamiento ante problemas como la violación del supuesto de ignorabilidad, la presencia de variables confusoras no observadas, y la existencia de estructuras de dependencia más complejas.

También podría profundizarse la simulación de escenarios más desafiantes, por ejemplo incorporando un mayor número de variables de confusión, mecanismos de asignación de tratamiento más sofisticados o relaciones no lineales más abruptas entre covariables y efectos.

Sobre la dificultad inherente a la inferencia causal

Es importante destacar que la estimación del efecto causal verdadero es, en la mayoría de las ciencias empíricas, una tarea altamente desafiante. A diferencia de problemas puramente predictivos, donde el *ground-truth* puede observarse directamente, en inferencia

causal el contrafactual es por definición inobservable, lo cual introduce una fuente estructural de incertidumbre. Esto implica que toda evaluación depende fuertemente del diseño del estudio, la calidad de los supuestos y la selección de variables relevantes.

Este hecho refuerza la necesidad de combinar enfoques estadísticos sólidos con conocimiento sustancial del problema bajo estudio, así como de utilizar múltiples modelos y métricas para validar resultados y comprender su sensibilidad.

Sin embargo, cabe preguntarse: **¿hasta dónde llegan realmente las limitaciones en la inferencia causal?** En la próxima sección se comparten algunas reflexiones en torno a esta pregunta central, clave para comprender cómo se abordan —y con qué grado de certeza— los problemas de estimación causal.

6.3. Reflexiones sobre el área de inferencia causal

En los dos primeros capítulos de este trabajo se han discutido algunos aspectos importantes del área de la inferencia causal. En particular, se subrayó la importancia que tiene en todas las ciencias empíricas la correcta evaluación de teorías causales alternativas. A pesar de ello, el elevado costo computacional asociado a calcular el posterior de los modelos $P(\text{Modelo}|\text{Datos})$ ha hecho que esta tarea sea todavía un aspecto marginal en la literatura de inferencia causal. Por este motivo, la disciplina actualmente está limitada casi exclusivamente a la estimación de efectos causales entre una causa y un efecto. En esta sección se busca reflexionar sobre la importancia de conocer la realidad causal subyacente, evaluando los modelos causales alternativos, incluso para este tipo de estimación de efectos causales pares de variables.

Primera limitación: antes de evaluar efectos causales

La necesidad de conocer la estructura causal subyacente es fundamental antes de comenzar el proceso de estimación de efectos causales entre el tratamiento y la variable objetivo. Cuál es el conjunto de variables que eliminan correctamente las correlaciones espurias de los datos observados depende de la estructura causal subyacente oculta en la cual están inmersas las variables tratamiento y objetivo. Debido a que no es posible computar el posterior de los modelos, el área de inferencia causal se ha visto obligada, en el mejor de los casos, a proponer estructuras causales basadas en el conocimiento experto, y tomarlas como verdades reveladas, asumiendo que son verdaderas (a pesar de la incerteza) para poder definir el conjunto de covariables de control.

En las competencias de datos de la ACIC, la discusión respecto a la selección de variables de control queda absolutamente fuera del alcance. Allí ni siquiera se transita el camino obligado en cualquier análisis de inferencia causal basado en datos observados sin intervenciones, en el que se requiere proponer una estructura causal para justificar las variables de control elegidas. En las competencias de datos no se propone ninguna estructura causal subyacente. En todos los casos, las simulaciones de las variables objetivo a partir de las covariables reales se diseñan siempre para garantizar que ellas funcionen como buenos conjuntos de control.

Hoy, gracias a los descubrimientos del área de la inferencia causal, se sabe que para estimar efectos causales entre un tratamiento y un objetivo es necesario conocer el contexto (la estructura causal) en la que se encuentran ese par de variables. Posiblemente existan

algunas heurísticas, soluciones parciales, que sean útiles en términos prácticos. En efecto, existe una creencia muy extendida que afirma que a medida que se agregan más variables en el conjunto de control mayor chances hay de que ese conjunto cumpla con el criterio de *ignorability* que se considera necesario para evaluar el efecto causal. En el libro *Introduction to Causal Inference: a Machine Learning Perspective* [12], Brady Neal argumenta que no siempre se puede asegurar que no haya confusión en los datos observacionales, pero ajustar por muchas covariables ayuda a aproximar la validez causal[1]. Estas ideas aparecen incluso en el artículo de Hill 2011 [17][2].

En efecto, si fuera posible medir todas las variables de la estructura causal subyacente de interés, condicionar sobre todas las variables presentes en los caminos *backdoor* siempre garantizaría el cumplimiento del criterio de *ignorability*. Sin embargo, si alguna de esas variables forman parte del camino *frontdoor*, la estimación del efecto causal no será correcta, pues se estarían cerrando los caminos causales a través de los cuales el tratamiento impacta sobre la variable objetivo.

Condicionar sobre todas las variables presentes en los caminos traseros siempre hace que el tratamiento sea independiente de los contrafactuales, que es lo que se quiere (cerrar *backdoor*), pero también puede hacer que el tratamiento sea independiente de la variable factual, que es lo que no se quiere (cerrar *frontdoor*). Es decir, se hace ineludible la obligación de reflexionar y distinguir al menos cuáles son las variables que forman parte del *backdoor* y cuáles forman parte del *frontdoor* en la estructura causal subyacente oculta, evitando agregar las variables del *frontdoor* en el conjunto de variables de control.

Distinguir entre variable *backdoor* y *frontdoor* es un método efectivo para determinar el conjunto de variables de control, solo en los casos en los que se cuente con la capacidad de medir todas las variables del *backdoor*. Pero con que no se pueda medir una única variable del *backdoor*, el método puede dejar de ser efectivo y nuevamente se visibiliza la obligación de conocer la estructura causal subyacente para decidir qué variables incluir en el conjunto de control. Por ejemplo, en el segundo capítulo de esta tesis se exhibió una estructura causal en el camino *backdoor* se encuentra cerrado por la presencia de un *collider*, haciendo que el conjunto vacío sea un conjunto de control bueno para cortar las correlaciones espurias (Figura 2.16). En ese contexto, agregar una variable al conjunto de control puede ser contraproducente, pues si esa variable es el *collider* se estaría abriendo el flujo de asociación *backdoor*, generando correlaciones espurias que no existían previamente en los datos.

En teoría, la aplicación estricta del sistema de razonamiento en contextos de incertidumbre permite evaluar correctamente los argumentos causales alternativos. Si se contara con la capacidad de intervenir en ciertas variables, rápidamente sería posible distinguir cuál es el estructura causal subyacente oculta, pues la ventaja de los modelos causales es la preservación de su capacidad predictiva a diferentes cambios de contextos. A través de la Entropía de Shannon (Ecuación 1.6) es posible deducir que el modelo que mejor predice es el que mejor se corresponde con la estructura causal subyacente.

[1] "We often cannot know for certain if conditional exchangeability holds. There may be some unobserved confounders [...], it is something that we must always be conscious of in observational data. Intuitively, the best thing we can do is to observe and fit as many covariates [...] to try to ensure unconfoundedness"

[2] "The ability to include many potential confounding covariates as predictors can be quite helpful when trying to satisfy *ignorability*. [...] (for a study that conditions on a huge number of covariates to justify *ignorability* and still is critiqued for not including enough, see Bingenheimer, Brennan, and Earls 2005)"

En la práctica, aproximar la aplicación estricta de las reglas de la probabilidad para computar el posterior de los modelos dado los datos, $P(\text{Modelo}|\text{Datos})$, suele ser computacionalmente costoso por lo que todavía el área de inferencia causal no es capaz de calcular correctamente la incertidumbre asociada a los argumentos causales alternativos en base a la evidencia. Esta es la primer gran limitación que tiene actualmente la inferencia causal como disciplina. Incluso antes de evaluar el impacto entre un tratamiento y un objetivo es necesario evaluar los argumentos causales alternativos para determinar qué conjuntos de control eliminan la correlación espuria de los datos observados sin intervenciones.

Segunda limitación: después de evaluar efectos causales

Como se mencionó anteriormente, a pesar de que la evaluación de argumentos causales alternativos sea uno de los objetivos más importantes comunes a todas las ciencias empíricas, el elevado costo computacional asociado a la evaluación del posterior de los modelos dado los datos $P(\text{Modelo}|\text{Datos})$ ha hecho que la disciplina de inferencia causal haya tenido que restringirse a la evaluación de efectos causales entre pares de variables: el tratamiento y el objetivo. En la sección anterior se han discutido las dificultades que se generan para la estimación de efectos causales cuando se desconoce la estructura causal subyacente.

Ante esta dificultad, cuando se desconoce la estructura causal subyacente, los experimentos aleatorizados aparecen como la única alternativa que garantiza estimar efectos causales correctamente. En este contexto el tratamiento se selecciona realmente de forma aleatoria, sin tener en cuenta absolutamente ninguna de las causas que motivan naturalmente la selección del tratamiento. Los experimentos reales, sin embargo, no suelen ser experimentos ideales: la asignación del tratamiento puede ser aleatoria simplemente al interior de un subgrupo que requiere tal intervención; o la intervención puede no ser perfecta, modificando el valor de otras variables no intencionalmente; o la intervención es perfectamente aleatoria pero su cumplimiento no; entre otras posibles complicaciones.

A pesar de todas las posibles complicaciones que pueden tener los experimentos aleatorizados reales, en esta sección se supondrá que la estimación de efecto causal es correcta. Si ese fuera el caso, se contaría con la capacidad de evaluar efectos causales correctamente sin conocer en absoluto la estructura causal subyacente en la cual están inmersos el tratamiento y el objetivo. Lo que se busca discutir en esta sección son las limitaciones que tienen los estudios de efectos causales (sin modelos causales) en el paso posterior a la estimación.

En el último artículo científico publicado por Martín Rossi, profesor de inferencia causal en la Universidad de San Andrés y Secretario de Desregulación del actual gobierno, se analiza una política pública de asignación aleatoria de subsidios para la adquisión de vivienda en Argentina [37]. La conclusión principal del trabajo *The unintended effect of Argentina's subsidized homeownership lottery program on intimate partner violence* afirma, en resumen, que la asignación del subsidio para la adquisión de viviendas a mujeres en pareja produjo un aumento en la violencia de genero producida por su pareja[3].

[3] "We study a natural experiment in Argentina, where low-income women were selected through a lottery system to receive a house and a heavily subsidized long-term mortgage. We exploit the random assignment to estimate the causal link between subsidized homeownership programs and intimate partner violence (IPV). [...]. **We find that the program causes an increase in IPV for women under joint-ownership contracts"**

Si bien este resultado puede resultar contraintuitivo, como se ha mencionado, en esta sección no se discutirá la validez de la estimación del efeccto causal, suponiendo que el resultado es correcto. En efecto, el resultado no es sorprendente para las teorías antropológicas sobre las estructuras elementales de la violencia [38], que identifican la obligación social denominada "mandato de masculinidad" de exhibición pública de poder en cualquiera de sus formas (intelectual, económica, sexual, física, etc) como condición para adquirir y renovar el estatus de hombre. Se podría discutir aquí en mayor detalle el modelo causal general propuesto en el libro de Segato [38]. Sin embargo, la evaluación de efectos causales a través de experimentos aleatorizados se caracteriza justamente por prescindir de todo modelo causal.

El problema que se busca señalar aquí es que incluso una correcta estimación del efecto causal no es suficiente si no se cuenta con un modelo causal que permita comprender el contexto en el cual se produce tal relación causal. Luego de verificar que la asignación del subsidio para la adquisición de viviendas produce un aumento en la violencia de contra la mujer, ¿qué decisión se debería tomar? ¿Habría que dejar de asignarle el subsidio a la mujeres en pareja? ¿Ayudarlas a separarse antes de adquirir el subsidio? ¿Esperar a que haya violencia por parte de la pareja antes de ayudarla a separarse? ¿Tiene sentido alguna de estas ideas?

Sin un modelo causal que sirva de contexto para el efecto causal no es posible tomar decisiones. Esta es la segunda limitación de la disciplina de la inferencia causal actual. No sólo es necesario evaluar modelos causales antes de sacar conclusiones causales en datos observados sin intervenciones. Luego de lograr estimar correctamente el efecto causal también se necesitan modelos causales para poder tomar decisiones.

Futuro de la inferencia causal

La inferencia causal actual está enfocada únicamente en la estimación de efectos causales entre pares de variables, prescindiendo de la evaluación de los argumentos causales alternativos. Esta es en sí misma una limitación de la disciplina, pues el objetivo más importante de todas las ciencias empíricas es la evaluación de la teorías alternativas dada la evidencia como un todo. En las dos secciones anteriores se buscó señalar que, incluso restringida a la estimación de efecto causal entre pares de variables, sigue vigente la necesidad de evaluar modelos causales.

Es necesario evaluar modelos causales antes de estimar el efecto causal en datos observados sin intervenciones para poder determinar cuál es el conjunto de variables de control que cortan las correlaciones espurias que se transmiten por el *backdoor*, sin cortar los efectos causales que se transmiten por el *frontdoor*. Y se necesita evaluar modelos causales después de estimar efectos causales para poder tomar decisiones.

Por ese motivo se argumenta que el futuro de la inferencia causal depende en gran medida del desarrollo de métodos eficientes de inferencia que permitan computar el posterior de los modelos dados los datos, es decir, $P(\text{Modelo}|\text{Datos})$.

Bibliografía

[1] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[2] Dan RC Thal and Mariel M Finucane. Causal methods madness: Lessons learned from the 2022 acic competition to estimate health policy impacts, 2023.

[3] Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition, 2019.

[4] Christopher M Bishop and Hugh Bishop. *Deep learning: Foundations and concepts*. Springer Nature, 2023.

[5] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

[6] Edwin T Jaynes. *Probability Theory: The Logic of Science (edited by Larry Bretthorst)*. Cambridge University Press, April 2003.

[7] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[8] Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty in artificial intelligence*, pages 46–54. Elsevier, 1994.

[9] Judea Pearl. From bayesian networks to causal networks. In *Mathematical models for handling partial knowledge in artificial intelligence*, pages 157–182. Springer.

[10] Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, 53(3):1071–1104, 2024.

[11] Matheus Facure. *Causal Inference for the Brave and True*. Self-published, 2023.

[12] Brady Neal. Introduction to causal inference. *Course Lecture Notes (draft)*, 2020.

[13] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[14] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[15] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[16] L Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[17] Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[18] P Richard Hahn, Vincent Dorie, and Jared S Murray. Atlantic causal inference conference (acic) data analysis challenge 2017. *arXiv preprint arXiv:1905.09515*, 2019.

[19] Carlos Carvalho, Avi Feller, Jared Murray, Spencer Woody, and David Yeager. Assessing treatment effect variation in observational studies: Results from a data challenge. *Observational Studies*, 5(2):21–35, 2019.

[20] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[21] P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

[22] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979.

[23] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), March 2010.

[24] Vincent Dorie, Hugh Chipman, and Robert McCulloch. *dbarts: Discrete Bayesian Additive Regression Trees Sampler*, 2025.

[25] Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457 – 472, 1992.

[26] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006.

[27] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[28] P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, 15(3):965 – 2020, 2020.

[29] Drew Herren, Richard Hahn, Jared Murray, Carlos Carvalho, and Jingyu He. *stochtree: Stochastic Tree Ensembles (XBART and BART) for Supervised Learning and Causal Inference*, 2025.

[30] Nikolay Krantsevich, Jingyu He, and P. Richard Hahn. Stochastic tree ensembles for estimating heterogeneous effects. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6120–6131. PMLR, 25–27 Apr 2023.

[31] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests, 2019.

[32] Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2):37–51, 2019.

[33] P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

[34] Julie Tibshirani, Susan Athey, Erik Sverdrup, and Stefan Wager. *grf: Generalized Random Forests*, 2024.

[35] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. EconML: A python package for ml-based heterogeneous treatment effects estimation, 2019.

[36] Aleksander Molak and Ajit Jaokar. *Causal Inference and Discovery in Python: Unlock the secrets of modern causal machine learning with DoWhy, EconML, PyTorch and more*. 2023.

[37] Bruno Cardinale Lagomarsino and Martin A Rossi. Jue insight: The unintended effect of argentina's subsidized homeownership lottery program on intimate partner violence. *Journal of Urban Economics*, 142:103612, 2024.

[38] Rita Laura Segato. *Las estructuras elementales de la violencia: contrato y status en la etiología de la violencia*, volume 334. Universidade de Brasília, Departamento de Antropologia Brasília, 2003.