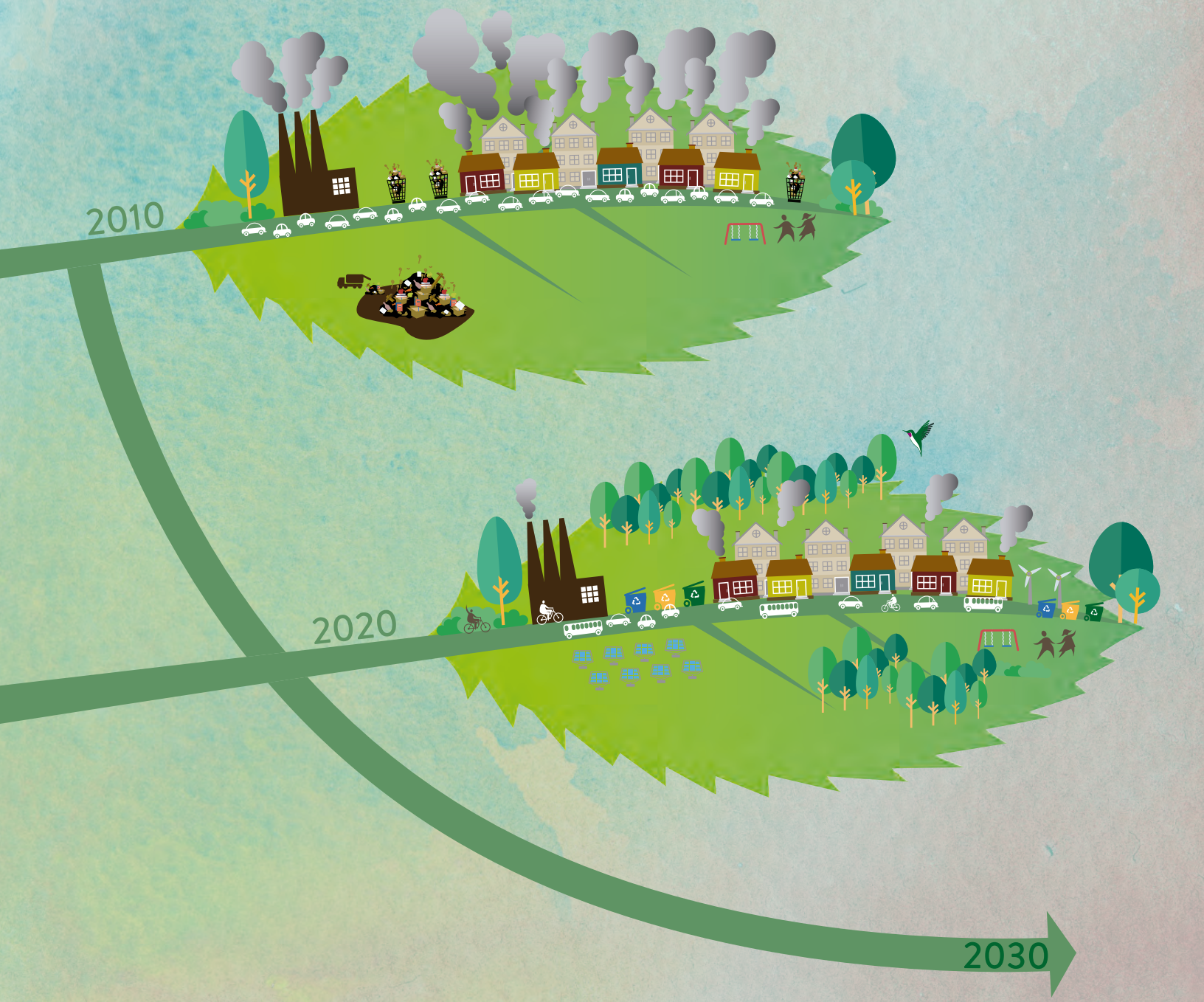
Ministerio del Medio Ambiente

Gobierno de Chile

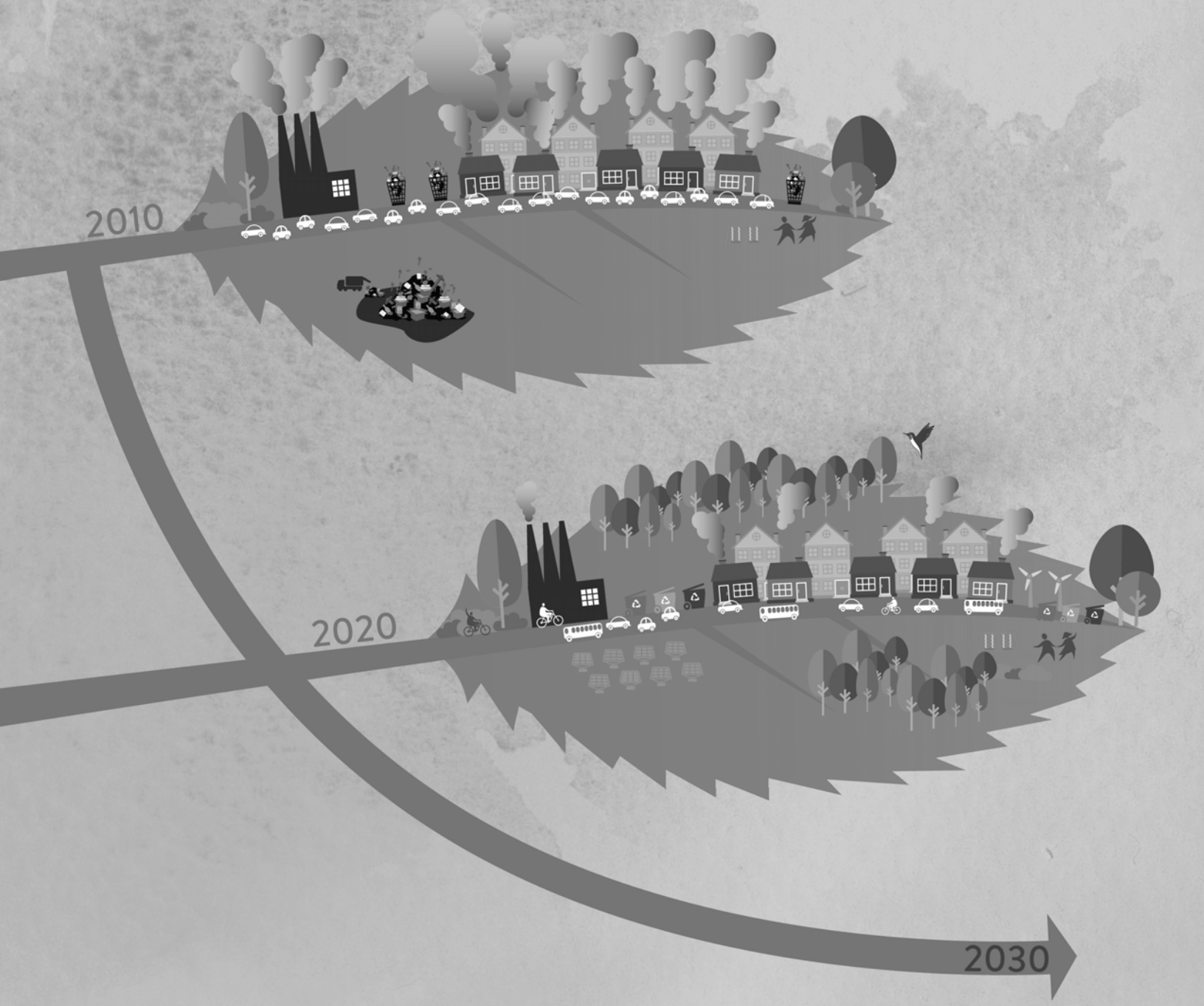
Guía Metodológica para

Evaluación *ex post* de programas y normativa ambiental



Guía Metodológica para

Evaluación *ex post* de programas y normativa ambiental



CONTENIDOS

1 | PRESENTACIÓN

En materia de Medio Ambiente, el logro de una mayor equidad ambiental constituye el eje de la gestión del Programa de Gobierno de la Presidente Michelle Bachelet para el periodo 2014-2018.

En el marco del desarrollo sustentable, la equidad social es un objetivo que acompaña el equilibrio entre crecimiento económico y la protección ambiental. Por ello, el Estado debe contar con las herramientas necesarias para adoptar decisiones en beneficio de la sustentabilidad y del bien común.

La Guía Metodológica para la Evaluación *Ex post* de Programas y Normativa Ambiental, provee de mecanismos que permitirán orientar el análisis de los resultados logrados y la medición del grado de cumplimiento de los objetivos de política pública propuestos en materia ambiental, retroalimentando y entregando insumos para perfeccionar los instrumentos, procesos y procedimientos de gestión y regulación vigentes.

Se espera que esta herramienta se transforme en un importante elemento en aras de la modernización del Estado, promoviendo la recolección y sistematización de información esencial para la toma de decisiones. Del mismo modo, constituye un esfuerzo por parte del Ministerio del Medio Ambiente enfocado a elevar la calidad de las políticas públicas, en un escenario en que las interacciones sociales e interdependencias se tornan cada vez más relevantes.

Pablo Badenier Martínez
Ministro del Medio Ambiente

2 | INTRODUCCIÓN

La política ambiental en Chile busca alcanzar el desarrollo sustentable con el objeto de mejorar la calidad de vida de los chilenos, tanto de esta generación como de futuras, a través de políticas públicas y regulaciones eficientes, que promuevan buenas prácticas y mejoren la educación ambiental ciudadana[1]. Una política ambiental intenta atender los problemas ambientales, formular soluciones de mediano y largo plazo, diseñar programas y definir prioridades en la asignación presupuestaria. Estas políticas pueden promover la calidad ambiental (enfocadas a la reducción de contaminantes), proteger la diversidad biológica o los recursos naturales. También, pueden enfocarse en aspectos preventivos como realizar diagnósticos de problemas ambientales, detallar soluciones técnicas, estudiar aspectos sociales relativos a cambios de políticas, entre otros. Para ello, la política ambiental está compuesta de diferentes instrumentos de gestión tales como programas, planes, normas, estrategias, entre otros.

Para que las políticas ambientales ayuden efectivamente a mejorar los problemas ambientales es necesario evaluar su calidad, incluyendo los planes, programas o medidas que las componen. La evaluación de las políticas ambientales actuales es útil para mejorar la eficiencia y eficacia, así como contribuir en la elección de futuras políticas e instrumentos económicos (OECD, 1997). La evaluación de estas políticas puede realizarse antes o después de su implementación, es decir, con una evaluación *ex ante* o evaluación *ex post*, respectivamente. En el primer caso se realizan estimaciones de escenarios basados en supuestos de comportamiento futuro para programas nuevos, mientras que en el segundo se utilizan técnicas estadísticas para medir el real impacto del programa o política[2]. Debido a su enfoque diferente las evaluaciones *ex ante* y *ex post* deberían ser pensadas como parte de un todo y no como herramientas separadas (Agnolucci, 2004).

En Chile, el Ministerio del Medio Ambiente (MMA) es el encargado del diseño y aplicación de políticas, planes y programas ambientales, así como de la protección y conservación de la biodiversidad biológica y de los recursos naturales renovables e hídricos, promoviendo el desarrollo sustentable, la integridad de la política ambiental y su regulación normativa[3]. Por esto, el Departamento de Economía Ambiental del Ministerio del Medio Ambiente ha desarrollado un programa de Evaluación Ambiental de Políticas Públicas, que contiene como uno de sus objetivos "evaluar el impacto, eficiencia y efectos distributivos de la política pública ambiental tanto *ex ante* como ex post para proponer mejoras para futuras implementaciones".

La presente guía está orientada a dar los lineamientos generales para una evaluación *ex post* de políticas ambientales, desde los objetivos y la utilidad de realizar este tipo de evaluaciones hasta los mecanismos y herramientas de análisis que pueden y deben ser consideradas para ello. La Guía reúne información teórica, ejemplos de aplicaciones ambientales, tanto de revisión nacional como internacional, y adjunta en los capítulos finales una recopilación bibliográfica de estudios de evaluación *ex post* para una mejor documentación de lo que se ha hecho hasta la fecha en esta materia.

1 http://portal.mma.gob.cl/vision-y-mision

2 Este escenario hipotético es denominado el contrafactual (ver capítulo 2).

3 http://portal.mma.gob.cl/vision-y-mision

3 ¿QUÉ ES LA EVALUACIÓN DE IMPACTO *EX POST*?

La evaluación de la política pública se puede definir como un proceso integral de observación, medida, análisis e interpretación respecto a una intervención pública que permite emitir un juicio, basado en evidencias, respecto a su pertinencia, implementación, resultados e impacto[4].

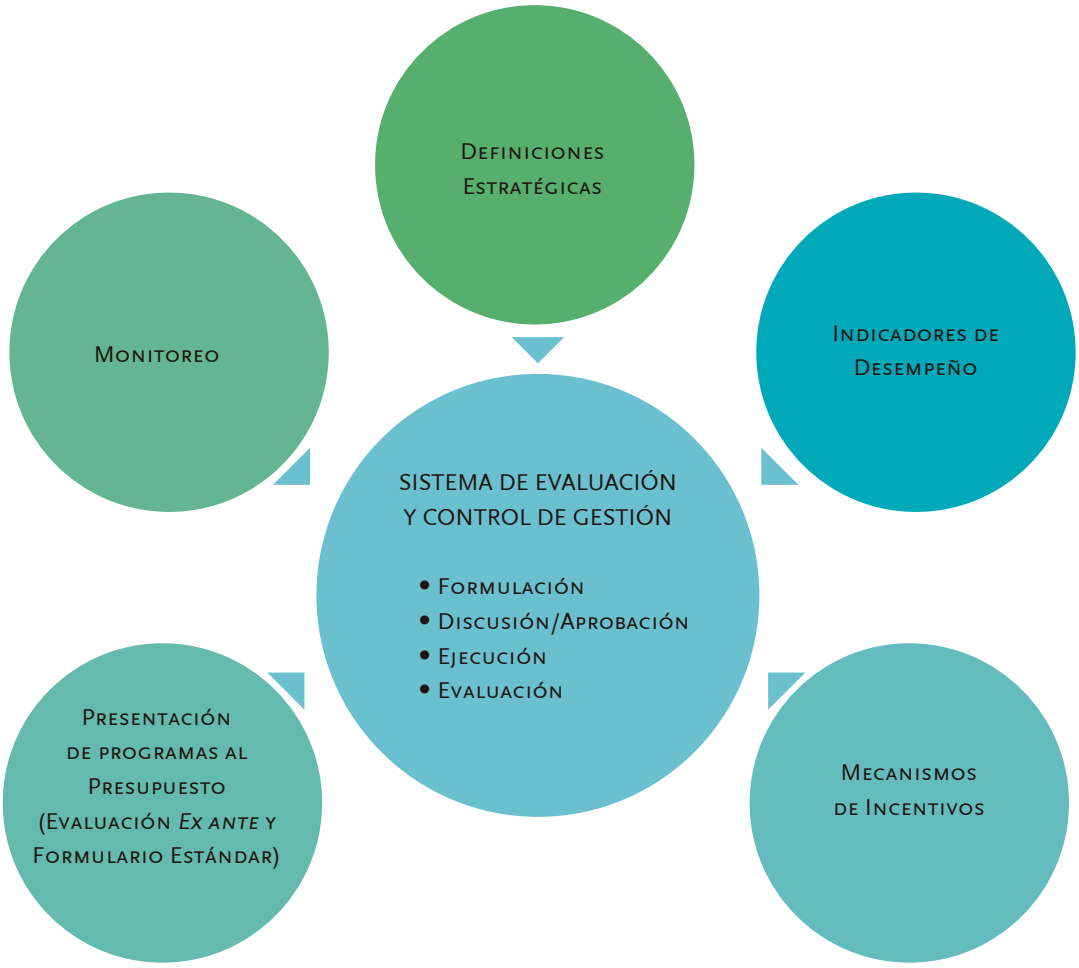
Específicamente, una evaluación de impacto *ex post*[5] es el análisis de los resultados logrados una vez que una política, programa o proyecto entra en operación, con el propósito de medir el grado de cumplimiento de los objetivos propuestos. En este contexto, permite retroalimentar y actualizar las metodologías, parámetros y supuestos del análisis técnico-económico (evaluación *ex ante*), para así entregar insumos que permitan efectuar las correcciones tendientes a perfeccionar los procesos y los procedimientos de inversión pública vigentes (Ministerio de Desarrollo Social, 2015).

Así, la evaluación *ex post* es la última fase -de evaluación- en el proceso del desarrollo de políticas o programas, cuyos resultados permiten retroalimentar el cumplimiento de los objetivos planteados. En el caso de Chile, el Sistema de Evaluación y Control de Gestión comprende distintos instrumentos asociados a las cuatro etapas de la gestión presupuestaria (Formulación, Discusión/Aprobación, Ejecución y Evaluación), ver Figura 3-1.

4. AEVAL (2010). "Fundamentos de Evaluación de Políticas Públicas".

5. En la presente guía se hablará indistintamente de evaluación de impacto y evaluación *ex post*.

Figura 3-1. Fases en el proceso de políticas o programas



Fuente: DIPRES

A diferencia de una evaluación *ex ante*, la evaluación *ex post* no pretende estimar efectos antes de la introducción de los programas ni simular diversos diseños hipotéticos que permitan mejorar el programa a implementar o evitar la introducción de programas inefectivos. Más bien, el objetivo es cuantificar los efectos generados por un programa existente aislando el efecto de otros factores no relacionados con el programa. Para ello se requiere responder a la pregunta **¿Qué hubiera sucedido si el programa o política no se hubiera realizado?** La complejidad de esta evaluación radica en establecer el **efecto causal** del proyecto, programa o política sobre los resultados de interés, es decir, si la intervención ha influido en los mismos, aislándolo de otros factores externos. Esto implica tener una teoría de cómo la intervención genera una cadena de resultados sobre los indicadores a medir y posteriormente la generación y recopilación de evidencias creíbles para responder a esa pregunta.

En particular, una evaluación de impacto *ex post* evalúa el efecto de una política, programa, proyecto o tratamiento sobre las unidades expuestas a ésta. Las unidades en evaluaciones económicas pueden ser individuos, familias, mercados, empresas, ciudades, regiones o países, pero en otras disciplinas las unidades pueden ser animales, áreas geográficas, aire, entre otros (Imbens & Wooldridge, 2009). Los tratamientos pueden ser programas, vouchers, leyes, regulaciones, medicamentos, entre otros. Por lo anterior, las políticas o programas ambientales encajan perfectamente en el concepto tradicional de la evaluación de impacto.

El impacto (τ) es definido matemáticamente como la diferencia en una variable de interés denominada "Y" entre el escenario con implementación del programa $Y(1)$ y la situación sin el programa $Y(0)$.

$$Impacto: \tau \equiv \Delta Y = Y(1) - Y(0)$$

De esta manera, para estimar el impacto de un programa (o política), además de determinar el resultado para los participantes del programa (tratados), se requiere estimar la situación contrafactual, es decir, **cuál habría sido el resultado para los participantes en el programa si no hubieran participado en él** $(Y(0))$**.** El contrafactual es por definición una situación hipotética dado que no es factible conocer el resultado del individuo sin el tratamiento si éste ya lo obtuvo, lo que es llamado el "problema fundamental de inferencia causal" (*Holland, 1986*). Es precisamente este el desafío de la evaluación *ex post* y existen técnicas para resolverlo que son desarrolladas en la presente guía.

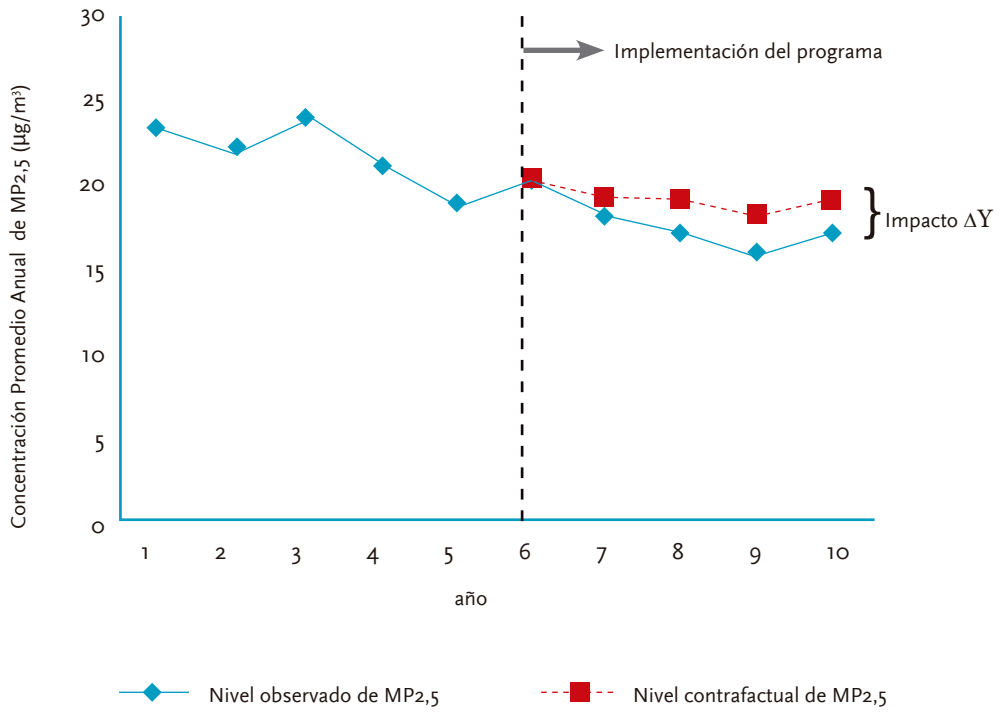
3.1 ESTIMACIÓN DEL ESCENARIO CONTRAFACTUAL

La generación de un escenario contrafactual permite corregir por factores externos que pueden haber influenciado el resultado durante el periodo de evaluación. Por ejemplo, no es posible atribuir al proyecto Plan de Recuperación Ambiental de Talcahuano (*PRAT iniciado en 1994*[6]) todo el mejoramiento a la calidad de las aguas en las bahías de la zona, dado que en el periodo posterior a su implementación también ocurrió una baja en la producción y contaminación debido a la crisis pesquera por la sobreexplotación de recursos.

Un ejemplo gráfico del contrafactual es presentado en la Figura 3-2, la cual muestra la evolución de las concentraciones del contaminante material particulado fino (MP2,5) ante la implementación en el año 6 de una medida que reduce emisiones. El contrafactual, representado por la línea roja, representa la evolución estimada de las concentraciones de MP2,5 en ausencia de la medida señalada.

6. En el marco de este programa se generaron una serie de estudios orientados a establecer las magnitudes y alcances de los procesos de deterioro ambiental en la ciudad de Talcahuano, y además, se inició el diseño de un plan de acción con el compromiso voluntario de las industrias para reducir su contaminación.

Figura 3-2. Impacto y escenario contrafactual



Fuente: Elaboración propia

Aun cuando la figura anterior corresponde a un ejemplo, es posible sacar algunas conclusiones generales:

- El impacto de la medida evaluada no corresponde a la diferencia entre el indicador (concentración de MP2,5) entre el año 10 y el año 6, dado que igualmente podría existir una variación en el indicador dado por el escenario contrafactual.
- El impacto varía dependiendo del año de evaluación elegido, por lo que corresponde a una decisión del evaluador cuándo se va a medir.

Afortunadamente la literatura estadística y econométrica ha desarrollado muchas estrategias para construir escenarios contrafactuales, tal como se explica en la siguiente sección. Sin embargo, cabe destacar que las estrategias de identificación son hipótesis sobre escenarios contrafactuales que nunca se observarán y, por lo tanto, no pueden ser probadas empíricamente. Por ejemplo, en una ciudad donde se haya implementado un Plan de Descontaminación Atmosférica[7] es imposible observar la evolución de las concentraciones del contaminante MP2,5 en ausencia del plan de descontaminación. En consecuencia, los cuestionamientos sobre los resultados de las evaluaciones de impacto *ex post* casi siempre recaen en la robustez de la hipótesis contrafactual.
Si bien una estrategia para construir el escenario contrafactual (no medible) es utilizar un grupo de control observable (medible), en la práctica es imposible identificar controles perfectos para las unidades tratadas. Sin embargo, existen técnicas que permiten definir grupos de unidades de control que sean estadísticamente indistinguibles, para lo cual se requiere:

- Que los grupos de tratamiento y control sean estadísticamente idénticos en las variables relevantes previo al programa (lo cual se puede evaluar con un test de diferencia de medias).
- Que ambos grupos reaccionen de la misma forma al tratamiento (este supuesto no es testeable).
- Que los grupos no sean afectados de forma distinta por factores exógenos u otros tratamientos. En este caso podría testearse si ambos grupos siguen comportamiento similares antes del tratamiento, aunque para ello se requiere contar con información de al menos dos periodos previos al tratamiento.

Bajo estas condiciones y supuestos, la diferencia entre los grupos de tratamiento y de control posterior a la aplicación del programa puede ser atribuible completamente a éste.

Por ejemplo, el objetivo de un programa de educación ambiental llevado a cabo en los establecimientos públicos de una comuna puede ser aumentar el reciclaje. Una alternativa para estimar un escenario contrafactual podría ser medir el nivel de reciclaje en establecimientos escolares de otras comunas con características similares. Así, el desafío de establecer estrategias de identificación consiste en encontrar unidades que posean características similares a las unidades que participan en el programa (por ejemplo: edades, tamaño de cursos, grado de vulnerabilidad social, etc.). Sin embargo, la dificultad surge porque, a pesar del esfuerzo por controlar las diferencias en las características, algunas de ellas son no observables (grado de motivación de los alumnos, conciencia ambiental del entorno familiar, etc.).

7. Instrumento de gestión ambiental establecido por la Ley 19.300 de Bases del Medio Ambiente que obliga a la reducción de emisiones de contaminantes mediante una batería de medidas de gestión.

3.2 ERRORES COMUNES EN LA ESTIMACIÓN DEL ESCENARIO CONTRAFACTUAL

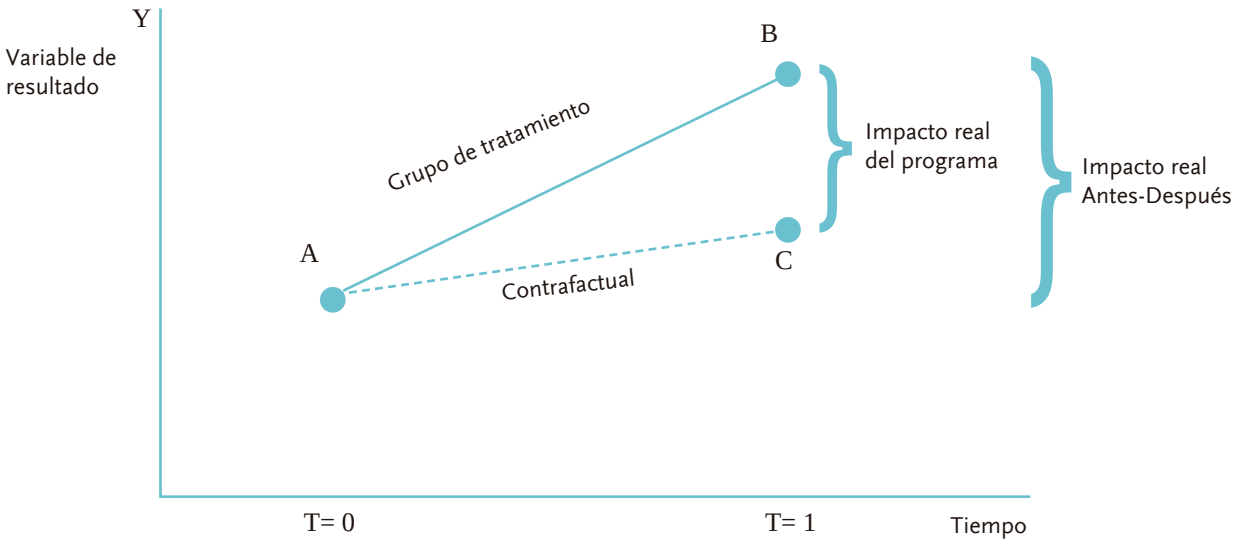
Es posible ver en diferentes estudios algunas imprecisiones en la aplicación de una evaluación de impacto, como por ejemplo, comparar las mismas unidades tratadas antes y después de la introducción de una política o programa (diseño "antes-después"), o bien, se comparan unidades que deciden participar o no participar del programa en un mismo momento (diseño "con-sin"). En estos casos, ambos contrafactuales son, en principio, inválidos.

3.2.1 Diseño "antes-después"

El diseño antes-después asume que si el programa no hubiera existido el resultado para los participantes del programa habría sido equivalente a su situación previa al programa. De esta forma, el impacto se obtiene simplemente al calcular la diferencia entre la media para el grupo después del tratamiento y la media para el grupo antes del tratamiento. **El supuesto central es que no existe ningún otro factor, salvo el programa, que haya podido afectar el resultado.**

En la Figura 3-3 el punto A representa el indicador Y en el tiempo T=0, es decir, antes de la aplicación del programa. La evaluación antes-después compara A con B (Y_B-Y), sin embargo, el indicador de todas maneras hubiera aumentado, lo que se ve reflejado en el punto C, contrafactual del análisis. De esta manera, la evaluación antes-después está adicionando un incremental de Y_C-Y_A que no seríacorrecto.

Figura 3-3 Efecto gráfico del tratamiento bajo diseño antes - después



Fuente: Elaboración propia

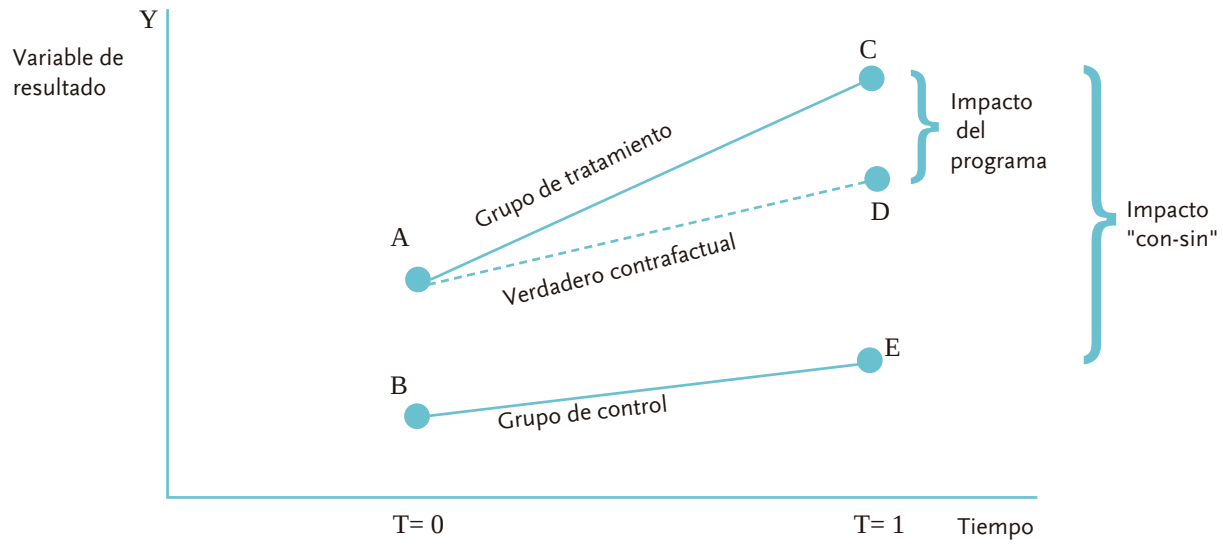
Por ejemplo, si por mercados internacionales el precio de la bencina sube, esto podría incidir en un menor uso de automóviles particulares, disminuyendo las emisiones y en consecuencia mejorar en parte la calidad del aire. Si se está evaluando el efecto en la calidad del aire por una medida de un Plan de descontaminación, esta debe considerar el escenario descrito, pues de otro modo, se le atribuiría un impacto adicional que no le corresponde.

3.2.2 Diseño "con-sin"

El diseño "con-sin" compara el resultado de los tratados con los no tratados sin considerar la posible existencia de diferencias entre los grupos. Aquellos que no se inscriben o no participan del programa podrían tener ciertas actitudes estratégicas dada su situación particular, ya que no se podría determinar si la diferencia de los resultados se debe al programa o a las diferencias previas entre los grupos. Este problema se denomina "sesgo de autoselección".

La Figura 3-4 representa la situación señalada gráficamente. El resultado con-sin (Y_C - Y_E) podría ser muy diferente al contrafactual (Y_D) producto principalmente a las diferencias que existían previamente a la aplicación del programa y su cambio en el tiempo, recalcando la importancia de establecer un buen grupo de control.

Figura 3-4 Efecto gráfico del tratamiento bajo diseño con-sin



Fuente: Elaboración propia

Un ejemplo de lo comentado es la medición de la efectividad de programas de "Pagos Por Servicios Ambientales" (PSA) para la protección de bosques. Si la variable de resultado del PSA es la superficie de bosque adicional protegida, es de especial relevancia controlar por el sesgo de autoselección debido a la estrategia que los individuos tienen a la hora de postular. En efecto, existe un incentivo para los propietarios que, desde el inicio, ya habían decidido destinar sus predios para la conservación, para recibir beneficios del programa aun cuando sus tierras no son amenazadas por otras actividades (Ver Arriagada *et al*., 2012) .

3.2.3 Confundentes

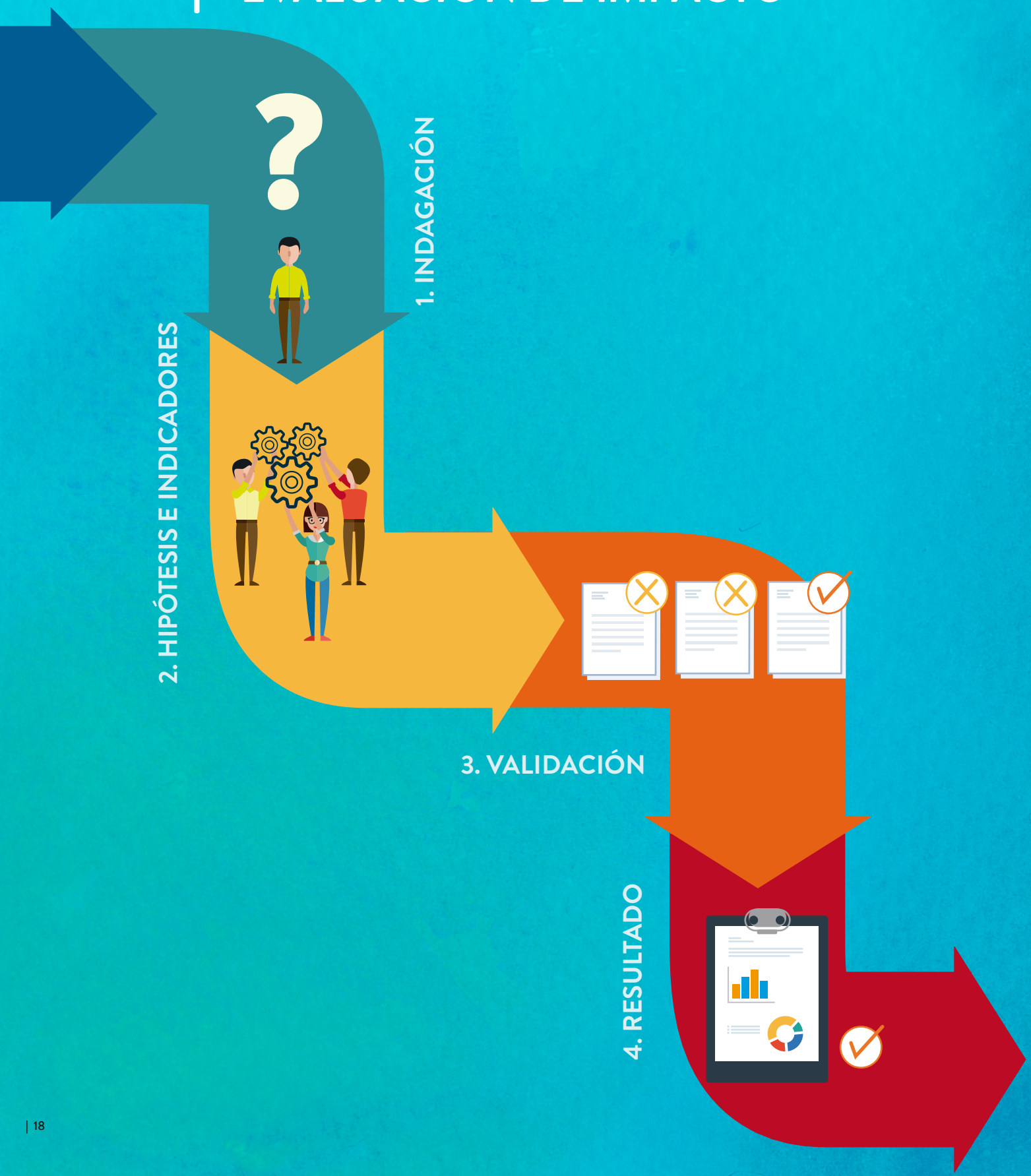
Cuando se intenta evaluar la relación entre una variable de tratamiento (independiente) y una variable de resultado (dependiente), pueden existir otras variables o factores externos a la relación causal que se evalua que también afectan a la variable de resultado en el mismo periodo bajo análisis. Así, la presencia de estos factores externos o variables de confusión (confundentes o confundidoras) puede generar sesgos en la estimación del efecto sobre la variable de resultado si no son controladas a través de técnicas estadísticas.

3.2.4 Asociación vs Causalidad

Como se ha mencionado, evaluar el impacto de un programa requiere ser capaz de aislar el efecto de la política o programa de otros factores que afectan los indicadores de resultados. Esto significa que **no se debe confundir asociación con causalidad**, lo cual nos puede llevar a una falacia en los resultados, y concretamente, asignar más o menos impacto a un programa del que corresponde.

Como ya se ha explicado, es imposible establecer la causalidad de un programa sobre una variable de resultado, porque no es posible observar el escenario con y sin tratamiento en un mismo momento del tiempo sobre las mismas unidades bajo análisis. Sin embargo, las diferentes técnicas de la evaluación *ex post* nos permiten establecer una causalidad desde el punto de vista estadístico (inferencia causal).

4 | ETAPAS PARA LA EVALUACIÓN DE IMPACTO



En esta sección se describen en detalle las diferentes etapas que debe cumplir una evaluación de impacto. Estas consisten en una serie de pasos que van desde la definición de lo que queremos evaluar (pregunta de evaluación) hasta la selección de la variables de resultado que se pretenden conseguir.

4.1 PREGUNTAS DE EVALUACIÓN

Cualquier evaluación parte con una pregunta relativa a la efectividad de una determinada política o programa. La típica pregunta en una evaluación *ex post* es ¿Cuál es el impacto o efecto causal de un programa sobre una variable de resultado? Sin embargo, también pueden existir otras preguntas como por ejemplo ¿Qué nivel o periodo de tratamiento genera los mayores impactos sobre una variable de resultado? o ¿Cuál combinación de programas genera los mayores impactos sobre una variable de resultado?

Una formulación adecuada de la pregunta o las preguntas es clave para diseñar adecuadamente la evaluación. También es relevante para determinar cuáles técnicas cualitativas y/o cuantitativas son las mejores opciones para realizar la evaluación.

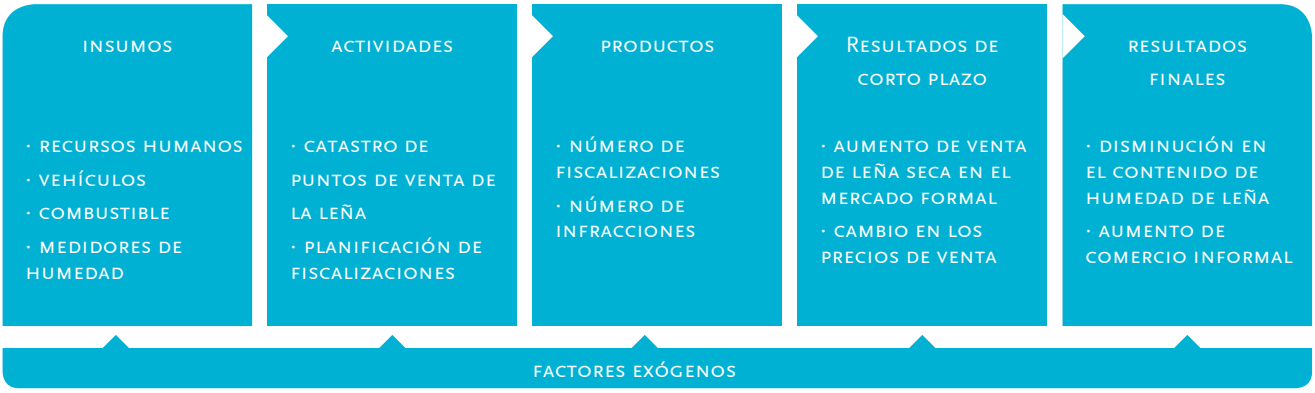
4.2 CADENA DE RESULTADOS

Uno de los aspectos fundamentales de cualquier evaluación de impacto es formular una teoría del cambio que describa en términos lógicos cómo se supone que el programa y las influencias ajenas al él cambiarán los indicadores de resultado.

Específicamente, una cadena de resultados define la secuencia en la cual los recursos del programa ayudan a realizar las actividades que permiten generar acciones (u ofrecer bienes o servicios) sobre el grupo de tratamiento en el marco de la implementación del programa y determinan los resultados de corto, mediano y largo plazo.

La Figura 4-1 ejemplifica una cadena de resultados de un programa piloto que aumenta la fiscalización para prevenir la venta de leña húmeda en el mercado formal.

Figura 4-1. Ejemplo de cadena de resultados en programa de fiscalización



Fuente: Elaboración Propia

- Los insumos incluyen el personal administrativo, personal operativo, materiales y otros recursos necesarios para la aplicación del programa.
- Las actividades incluyen todos los bienes, servicios e incentivos ofrecidos a los participantes durante la ejecución del programa.
- Los resultados de corto plazo son todos los cambios que ocurren inmediatamente después que se recibe el tratamiento del programa.
- Los resultados de mediano plazo no son causados directamente por los componentes del programa sino que son generados posteriormente debido a los cambios iniciales.
- Los resultados finales o de largo plazo son los que requieren un mayor periodo de tiempo para que se generen sus efectos.
- Los factores exógenos se asocian a cualquier shock o variables independientes a la participación del programa que puede incidir en los resultados y confundirse con los impactos del programa porque ocurren en el periodo sujeto a evaluación.

En este caso, los resultados de corto plazo y resultados finales no necesariamente están bajo control y pueden generarse por múltiples factores exógenos, los cuales influyen de forma positiva o negativa durante todo el proceso, desde la implementación del programa hasta la generación de los resultados finales.

Como se observa, las cadenas de resultados son útiles porque permiten entender la lógica causal y la secuencia de eventos en los que se basa el programa, y además, ayudan a definir la información necesaria de levantar para medir los cambios y resultados. En consecuencia, **una evaluación de impacto *ex post* permite establecer si la teoría del cambio funciona en la práctica.**

Una buena alternativa para establecer los supuestos de una cadena de resultados es revisar la literatura especializada en busca de experiencias similares.

4.3 HIPÓTESIS PARA LA EVALUACIÓN

Una vez que se ha descrito la cadena de resultados, es posible definir una lista con los efectos esperados de la aplicación del programa y se puede formular la hipótesis que se quiere comprobar mediante la evaluación de impacto. Sin embargo, la hipótesis también puede basarse en el conocimiento generado por experiencias previas con programas similares.

Típicamente, la hipótesis estadística se basa en comprobar **si el valor esperado del indicador es igual entre los grupos de tratamiento y el grupo de control**. Definiendo como el indicador de resultado para el individuo i cuando recibe el tratamiento como $Yi(1)$ y el

indicador de resultado para el individuo i cuando no recibe el tratamiento como $Y_i(0)$, el test de hipótesis es:

$$H_0: E[Y_i(1)] = E[Y_i(0)]$$
$$H_1: E[Y_i(1)] \neq E[Y_i(0)]$$

Abadie (2002) aborda otra hipótesis al preguntar si la distribución de $Y_i(1)$ difiere de la distribución de $Y_i(0)$, es decir, no solo se evalúa la media sino todos los momentos de la distribución[9].

Otro tipo reciente de hipótesis aborda la heterogeneidad del efecto del tratamiento, ya que incluso si el efecto promedio es cero, podría ser importante establecer si en determinados subgrupos existe algún efecto del tratamiento (Angrist, 2004).

4.4 INFERENCIA CAUSAL

El impacto causal (τ_i) de una política o programa sobre una unidad[10] i es la diferencia entre el resultado que habría ocurrido si la unidad i no hubiese participado y el resultado que habría ocurrido si la unidad hubiese participado. Así la fórmula del impacto sobre una unidad es:

$$\tau_i = Y_i(1) - Y_i(0)$$

Si la unidad i participa en el programa o política el resultado observado será $Y_i(1)$ mientras $Y_i(0)$ será un resultado contrafactual . Alternativamente, si la unidad i no participa en el programa, el resultado observado será $Y_i(0)$ mientras que $Y_i(1)$ será un resultado contrafactual. Como se escribió anteriormente, es imposible observar a la misma unidad en dos situaciones diferentes en un mismo momento del tiempo.

Cuando se realiza una evaluación de impacto *ex post*, es relativamente fácil obtener $Y_i(1)$ del grupo de unidades tratadas, ya que es el resultado con el programa o política. Sin embargo, $Y_i(0)$ no es observable directamente. Por ello, se recurre a los denominados "grupos de control" que es un segmento de la población objetivo que no ha recibido el tratamiento y comparte características comunes con el grupo de tratamiento. A partir de diferentes técnicas estadísticas se pueden establecer grupos de control válidos que ayuden a determinar el impacto de un programa.

9. Los momentos son funciones que sirven para caracterizar a las distribuciones de probabilidad. Estos momentos son obtenidos a partir de los valores esperados de ciertas funciones de una variable aleatoria.

10. En este contexto unidad se refiere a persona, hogar, industria u otro elemento expuesto o no al tratamiento.

4.5 INDICADORES DE TRATAMIENTO

La evaluación de impacto típicamente se ha centrado en los efectos promedios de un tratamiento P_i. A continuación se describen los efectos más utilizados:

Tabla 4-1. Indicadores de impacto comúnmente utilizados

Nombre indicador	Descripción	Expresión estadística
ATE Average Treatment Effect	El efecto promedio del tratamiento considera la esperanza poblacional del efecto causal a nivel de unidades. Este efecto es útil si se desea evaluar la alternativa de aplicar el tratamiento a todas las unidades o a ninguna de ellas.	$\tau_{ATE} = E[Y_i(1) - Y_i(0)]$
ATT Average Treatment Effects on the Treated	El efecto promedio sobre los tratados considera la esperanza sobre el grupo de unidades tratadas. Este efecto es útil si se desea evaluar el efecto sólo sobre las unidades expuestas.	$\tau_{ATT} = E[Y_i(1) - Y_i(0) \mid P_i = 1]$
ATNT Average Treatment Effects on Non-Treated	El efecto promedio sobre los no tratados considera la esperanza sobre el grupo de unidades no tratadas. Este efecto es útil si se desea evaluar el efecto que existiría sobre las unidades no expuestas.	$\tau_{ATNT} = E[Y_i(1) - Y_i(0) \mid P_i = 0]$

P_i : participación en el programa / $P_i=1$: participa / $P_i=0$: no participa

Fuente: Imbens & Wooldridge (2009)

Recientemente, se ha generado un creciente interés sobre la distribución del efecto del tratamiento, lo cual ha llevado particularmente al estudio de dos nuevos efectos que se muestran en la Tabla 4-2.

Tabla 4-2. Indicadores de impacto para medir cambios locales o marginales

Nombre indicador	Descripción	Expresión estadística
LATE Local Average Treatment Effect	Es utilizado para analizar el efecto del tratamiento en un pequeño subgrupo de unidades expuestas. El efecto promedio local del tratamiento asume que la participación o no participación cambia cuando una variable exógena Z cambia desde un valor Z* a Z**. Por ejemplo, Z puede ser una variable que describa si se realizó o no una promoción sobre un grupo de unidades para fomentar la participación en el programa.	$\tau_{LATE} = E[Y_i(1) \mid P_i(Z^*) = 1] - E[Y_i(0) \mid P_i(Z^{**}) = 0]$
MTE Marginal Treatment Effect	El efecto marginal del tratamiento estima el cambio en el resultado ante un cambio infinitesimal en la probabilidad de participación. También, es utilizado para extrapolar los resultados a toda la población cuando se asume una distribución de probabilidad.	$\tau_{MTE} = dE(Y_i)/dP$

Fuente: Imbens & Wooldridge (2009)

4.6 SELECCIÓN DE INDICADORES DE DESEMPEÑO

Uno de los elementos centrales en el proceso de evaluación de impacto es determinar los indicadores de desempeño que medirán los efectos. Dependiendo del contexto del programa los indicadores pueden medir efectos directos (para los beneficiarios del programa) e indirectos (otros actores que sean afectados por externalidades del programa).

Ambos tipos de indicadores pueden dividirse en indicadores que miden *inputs* u *outputs*. Por ejemplo, los indicadores que miden *inputs* pueden identificar el consumo de ciertos combustibles de las fuentes industriales que están reguladas bajo una nueva normativa, mientras que los indicadores de *outputs* identificarían el nivel de emisiones.

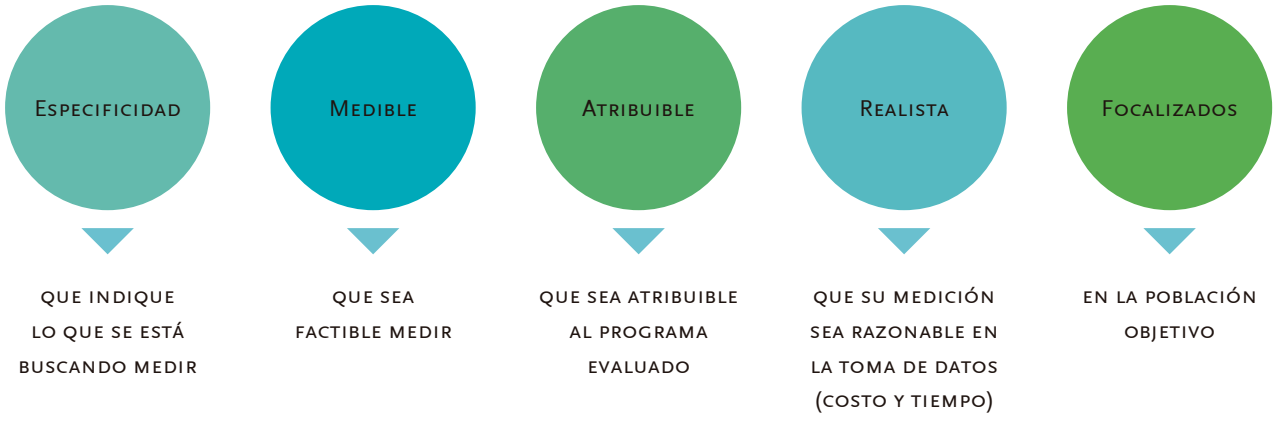
Los indicadores de desempeño pueden medir tanto la implementación del programa como los resultados. Por ello, es útil que la selección de los indicadores sea acordada con los involucrados en el diseño y ejecución del programa.

Si el programa tiene objetivos múltiples es necesario seleccionar el objetivo cuyo impacto más interesa evaluar, o decidir si se desea evaluar más de uno. No obstante, algunos objetivos pueden ser multidimensionales por lo cual hay que decidir cuál o cuáles de las dimensiones se medirán.

Independientemente de los indicadores elegidos, éstos deben cumplir con algunas características básicas[11] que son descritas por el Banco Mundial en su guía titulada "La evaluación de impacto en la práctica" y que se presentan en la Figura 4-2. Luego, en la tabla 4-3 se presentan algunos ejemplos de indicadores de impacto.

11. Estos criterios son también llamados EMARF.

Figura 4-2. Características básica para indicadores



Fuente: Elaboración propia

Tabla 4-3. Ejemplo de indicadores de impacto

Finalidad	Ejemplos de indicadores
Gestión del programa (indicador de implementación)	• N° de profesionales que participan en el programa • Grado de satisfacción con la coordinación de los diferentes equipos de trabajo • N° de reuniones del equipo ejecutor • Porcentaje de ejecución presupuestaria del programa
Actividades realizadas (indicador de implementación)	• Porcentaje de realización de las actividades planificadas • N° de informes • Número de comunidades intervenidas
Caracterización (indicador de implementación)	• Características de la población objetivo (socioeconómicas, conductas ambientales, patrones de uso o consumo, entre otras) • Tasa de adopción del programa • Deserción del programa • Tasa de cobertura por zona geográfica
Eficacia del programa (indicador de resultado)	• Reducción en el uso de combustibles fósiles • Reducción en las quemas al aire libre • Reducción en las atenciones de urgencia por problemas respiratorios • Reducción en las descargas de RILES • Reducción en el número de días con pre-emergencia ambiental • Aumento en la adopción de tecnologías limpias • N° y tipos de quejas de los beneficiarios • Cambio en la percepción en la calidad ambiental

Fuente: Elaboración Propia

4.7 RECOLECCIÓN DE INFORMACIÓN

La recolección de información permite la medición y cuantificación de los indicadores de impacto, existiendo métodos diferentes según la naturaleza cualitativa o cuantitativa de la información.

En el caso de métodos cuantitativos podemos mencionar los siguientes:

- Registros administrativos de los participantes en el programa.
- Registros internos relacionados con la gestión del programa, los cuales pueden incluir información sobre insumos, actividades y/o servicios.
- Bases de datos estadísticas que fueron desarrolladas para otros fines (por ejemplo: Encuesta ENIA, Encuesta CASEN, entre otras) pero que pueden contener información de la población afectada y potenciales grupos de control, tanto de años anteriores como posteriores a la implementación del programa.
- Encuestas realizadas mediante un cuestionario a una muestra o bien a toda la población objetivo, **tanto tratados como no tratados con el programa.**

En el caso de métodos cualitativos podemos mencionar los siguientes[12]:

- Revisión de documentos oficiales, informes, investigaciones, evaluaciones, libros, artículos periodísticos, entre otros.
- Levantamiento de información mediante la observación de las actividades y participantes en el programa.
- Entrevistas en profundidad a los actores involucrados en el diseño, implementación del programa, así como también, a los beneficiarios del programa que permiten recoger sus puntos de vista y experiencias sobre el programa.
- *Focus groups* guiados por profesionales dirigidos a los actores involucrados en el diseño, implementación del programa, así como también a los beneficiarios del programa.
- Paneles de expertos con experiencia en la aplicación previa de programas similares.
- Estudio de casos para el análisis de situaciones particulares pero que por su contexto permitan determinar factores críticos en el logro de los objetivos del programa.

12. De acuerdo a la DIPRES las técnicas cualitativas que más ha utilizado son las entrevistas (semi-estructuradas) y los *focus groups*.

4.8 VALIDEZ INTERNA Y EXTERNA DE UNA EVALUACIÓN DE IMPACTO

Un aspecto importante del diseño de evaluación de impacto es su validez, la cual tiene que ver con la robustez de las conclusiones de la evaluación. Los principales tipos son la validez interna y externa.

4.8.1 Validez interna

Una evaluación de impacto posee validez interna cuando el impacto estimado ha condicionado todos los otros factores que pueden afectar el resultado (confundentes), y por lo tanto, se puede estimar el verdadero impacto del programa. Esto ocurre cuando el grupo de control utilizado representa adecuadamente la situación contrafactual, es decir, es capaz de reflejar qué habría pasado con las unidades tratadas en ausencia del programa.

En términos prácticos, si no existe validez interna se está asignando parcial o totalmente el impacto a un programa o política, cuando en realidad se debe a factores exógenos no atribuibles al programa o política. En este caso surgen los llamados sesgos por omisión y sesgos de selección los cuales afectan la validez interna de la evaluación, como por ejemplo:

- Atribuir la reducción en la compra de vehículos que emiten altos niveles de CO_2 a la introducción de impuestos verdes cuando ha ocurrido una contracción en la actividad económica (omisión).
- Atribuir la reducción a las emisiones de fuentes industriales a la introducción de normas de calderas, cuando al mismo tiempo se amplía la oferta de gas natural producto de la decisión de instalar plantas de regasificación en el país (omisión).
- Las empresas que desean participar en un programa voluntario de producción limpia, pueden ser las menos contaminantes o que tengan las opciones más atractivas para realizar el cambio en su proceso productivo (selección).

Otra forma de sesgo de selección es la atrición de la muestra. Esto significa que a lo largo de la implementación del programa, algunos participantes del grupo de tratamiento abandonan el programa o algunos participantes del grupo de control se rehúsan a

seguir respondiendo los cuestionarios. Este cambio en la composición de los grupos puede afectar el impacto observado.

En algunas evaluaciones se levanta información a través de cuestionarios, pero los temas abordados en el cuestionario pueden modificar la conducta de las personas en el futuro al darse cuenta de su poca preocupación en estas temáticas o, alternativamente, en encuestas futuras estarán mejor preparadas para responder las preguntas. Por otra parte, el efecto Hawthorne es un cambio en el resultado que experimentan los tratados por el simple hecho que se les pone atención a su comportamiento, y no por el programa. Así, esta situación también podría afectar la validez interna de la evaluación de impacto.

Una modificación en el instrumento de medición (por ejemplo, el cuestionario) entre el periodo previo y posterior a la implementación del programa, puede generar una variación en los resultados que se confunde con los impactos del programa.

También, la generación de *spillovers*[13] entre el grupo de tratamiento y el grupo de control que se producen por la interacción, contacto o aprendizaje entre ellos, pueden generar una subestimación del programa.

La validez interna **no es una propiedad de las metodologías**, sino que se asocia a las aplicaciones de inferencia causal realizadas en cada evaluación. Un mismo método puede entregar conclusiones con mayor o menor validez interna dependiendo de las circunstancias y características del programa evaluado.

En resumen, los elementos a considerar que afectan la validez interna son los siguientes:

- **Autoselección** de los participantes para postular al programa.
- **Atrición de la muestra**.
- **Factores no observables** o shocks que afectan de forma distinta al grupo de tratamiento y grupo de control.
- **Cambios de conducta** o bien aprendizaje derivado de las preguntas de los cuestionarios con los cuales se levanta la información (efecto Hawthorne).
- **Interacción o aprendizaje** entre los grupos de tratamiento y control (spillovers).

4.8.2 Validez externa

Una evaluación de impacto posee validez externa cuando el impacto estimado puede generalizarse a toda la población elegible de la política o programa. Naturalmente, esto requiere que la muestra utilizada en la evaluación sea representativa de la población elegible.

Alternativamente, se puede interpretar como la generalización a otros programas similares, situaciones o momentos. Sin embargo, en la medida que las condiciones del programa sean más controladas para facilitar la evaluación, menos generalizables serán las conclusiones. En estos casos es importante que al menos exista validez interna.

4.9 TEMAS RELEVANTES EN LA EVALUACIÓN *EX POST*

Existen varias consideraciones importantes que hay que considerar a la hora de diseñar una evaluación *ex post* y en la interpretación de resultados de la misma que son remarcadas a continuación.

13. Interacciones que se pueden dar entre ambos grupos que pueden alterar el comportamiento y alterar los resultados asociados al impacto del programa.

4.9.1 Diseño de la evaluación

Las evaluaciones de impacto se podrían diseñar mejor antes de que comience la aplicación de un programa o política. Cuando ya se implementó no es posible influir en el mecanismo de asignación del tratamiento, y posiblemente, ya no se recogió información de línea base, por lo cual las opciones de técnicas cuantitativas para realizar la evaluación se reducen considerablemente.

4.9.2 Efectos temporales

El impacto de diferentes programas o políticas puede mostrar diversos patrones a través del tiempo. Una intervención puede generar fuertes impactos en el corto plazo, pero irse diluyendo a través del tiempo. Alternativamente, el impacto de un programa puede aparecer sólo después de cierto periodo, incluso incrementándose con el transcurso del tiempo. Por lo anterior, se pueden obtener conclusiones y recomendaciones equivocadas si no se considera adecuadamente la temporalidad de los impactos.

A modo de ejemplo, un programa de conservación de especies en peligro de extinción probablemente no presentará impactos de corto plazo, pero sí pueden ser significativos luego de un par de años, una vez que se han recuperado las poblaciones de las especies.

4.9.3 Multiplicidad de tratamientos

En algunos programas que se han diseñado para incorporar diversos tratamientos se puede determinar el impacto de cada tratamiento, pero también su impacto combinado. Este análisis es interesante porque el efecto total puede ser mayor o menor que la suma de los tratamientos individuales.

4.9.4 Heterogeneidad del tratamiento

El nivel o intensidad del tratamiento puede variar a través de las unidades de un el grupo de tratamiento. Por ejemplo, un programa de educación ambiental no tendrá el mismo efecto en personas que hayan asistido a un 40% de las clases, respecto a otras personas que hayan asistido a un 90% de las clases.

4.9.5 Heterogeneidad del impacto

El efecto de un programa probablemente no será homogéneo en el grupo de tratamiento. Por lo anterior, las evaluaciones de impacto no se limitan a estimar la magnitud promedio de los efectos, sino también pueden determinar el efecto en diferentes subgrupos de la población objetivo. Conocer si el efecto del programa es heterogéneo en estos subgrupos es útil para redefinir a los beneficiarios, así como también intensificar o introducir nuevos componentes en el programa que aumenten el impacto.

4.9.6 Intensidad de la intervención

Típicamente se ha enfatizado el caso binario de participación o no participación en un determinado programa. Sin embargo, el tratamiento puede presentar diferentes niveles o bien ser uniforma. Estas características del tratamiento deben ser definidas en el diseño del programa, lo cual permitiría en la etapa de evaluación determinar si existe un "nivel óptimo" de intervención.

5 | METODOLOGÍAS CUANTITATIVAS DE EVALUACIÓN

Las metodologías cuantitativas son claves en la evaluación de impacto, ya que el objetivo central que se intenta responder es de naturaleza cuantitativa. Dependiendo del tipo de selección entre individuos con y sin tratamiento, las metodologías se clasifican en los siguientes grupos:

Experimentales: para evitar sesgo por autoselección el método asigna la participación de forma aleatoria (por ejemplo a través de una lotería), asegurando que las unidades pertenecientes al grupo de tratamiento y control tengan características iguales (en términos estadísticos) ya que no se les permite elegir si participan o no.

Cuasi – experimentales (o diseños naturales): en este caso existe un evento fortuito que asigna la participación con características similares a un experimento aleatorio controlado (por ejemplo diferencias en el momento de inicio del tratamiento, diferencias regionales del programa, un cambio no esperado en la legislación, entre otros).

No experimentales: es un conjunto de técnicas estadísticas que tratan de identificar el impacto de un programa aun cuando no se haya realizado una asignación aleatoria del tratamiento.

FIGURA 5-1 CLASIFICACIÓN DE TÉCNICAS DE EVALUACIÓN DE IMPACTO Y METODOLOGÍAS ASOCIADAS[14]

Fuente: Elaboración propia

Para controlar las características que originan el sesgo de selección, algunas técnicas usan datos antes y después del programa, tanto para el grupo de tratamiento como para el grupo de control, con el objetivo de controlar por las características no observables. Estas técnicas aprovechan la aleatorización creada a través de un evento externo al evaluador, controlando diferencias de las características observables entre los grupos intentando restablecer condiciones experimentales en un contexto no experimental.

Las diferentes metodologías tienen un objetivo común: cómo encontrar un individuo (o grupos de individuos) comparable al individuo (o grupos de individuos) con tratamiento, eliminando o minimizando todos los sesgos para aislar exclusivamente el efecto del programa evaluado.

14. Aunque esta es la clasificación estándar de metodologías, en la práctica se han ido generando combinaciones de técnicas, como por ejemplo: *matching* con diferencias en diferencias, diferencias en diferencias con variables instrumentales, *matching* con variables instrumentales, función de control con modelos estructurales, entre otras (Ver Figura 7-2).

A continuación se describen los principales métodos para la evaluación de impacto, su formulación estadística y sus ventajas y desventajas, lo cual tiene especial relevancia a la hora de elegir una o una combinación de ellas.

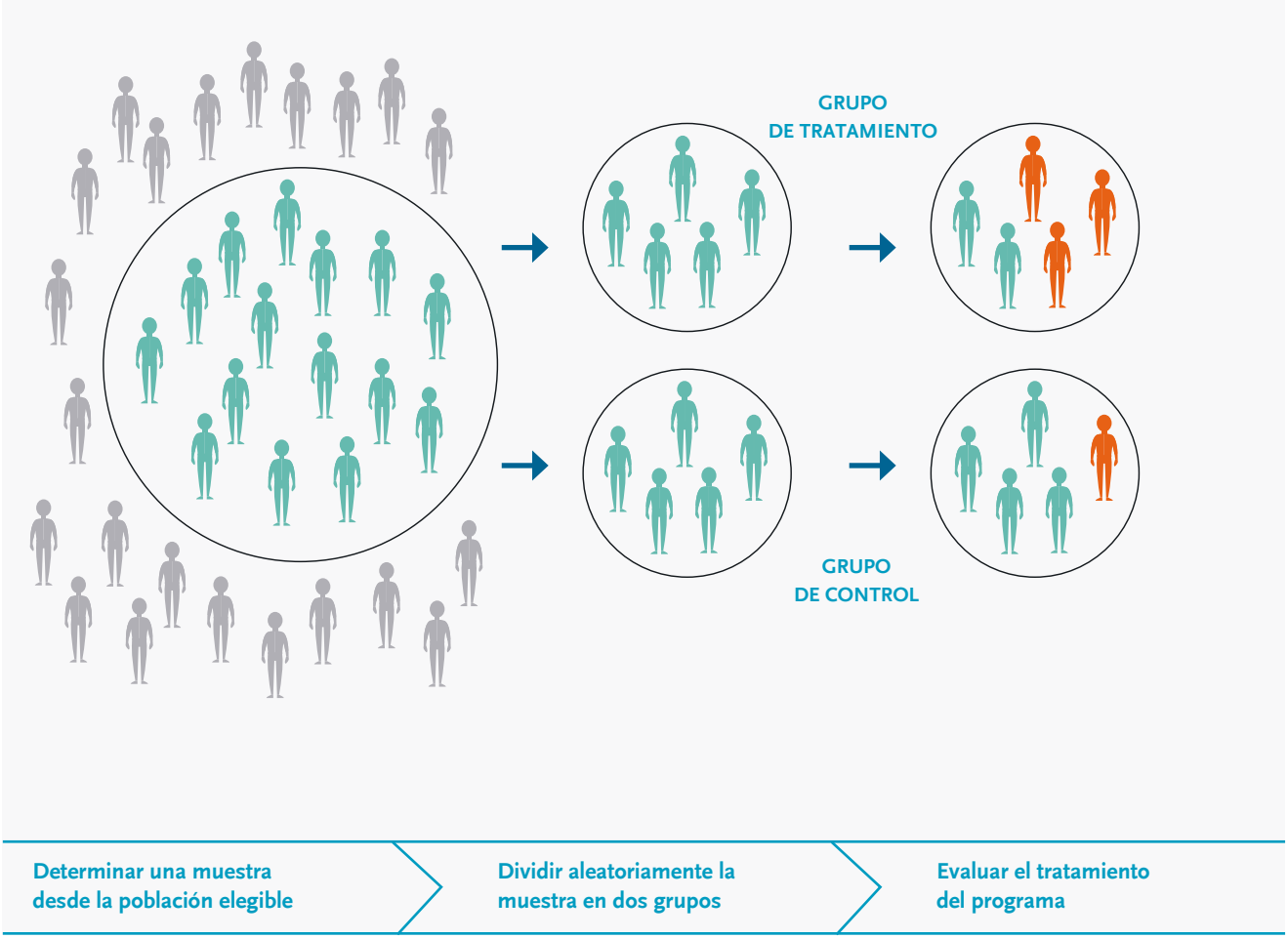
5.1 EXPERIMENTOS ALEATORIOS

5.1.1 Descripción

Evaluar el impacto de una política o programa con un experimento aleatorio consiste en seleccionar al azar un grupo de unidades elegibles (es decir, que puedan participar en el programa a evaluar), para formar dos grupos al azar: beneficiados y no beneficiados por el programa.

La Figura 5-2 representa los pasos mencionados, donde el color rojo de los individuos representa aquellos que han sufrido cambios en el indicador de resultado respecto a la situación base. Como ha sido mencionado, no sólo los tratados pueden presentar dicho cambio.

Figura 5-2 Diagrama de experimento aleatorio



Fuente: Elaboración propia

La robustez del método se basa en que la selección aleatoria permite que las características observables y no observables de las unidades se distribuyan de forma similar entre ambos grupos. Esto es posible siempre que el número al que se aplique la asignación aleatoria sea lo suficientemente grande, lo cual dependerá de la significancia estadística, potencia y varianza del indicador que se pretende medir.

Dado que ambos grupos son equivalentes en términos estadísticos en todas las características observables y no observables que pueden afectar sobre el resultado, es lógico atribuir causalidad a cualquier diferencia en el tratamiento. Sin embargo, se requiere que la equivalencia de las características entre ambos grupos se mantenga durante toda la aplicación del experimento.

Aun cuando los experimentos aleatorios son considerados la "regla de oro" para la evaluación de impacto, no existen muchos experimentos aleatorios en contextos económicos o sociales, y menos aún en un contexto netamente ambiental.

Esto se explica principalmente por el alto costo del seguimiento de una muestra representativa de la población elegible. La estimación del tamaño muestral con un cierto margen de error y nivel de significancia estadística podría requerir miles de personas en el grupo de tratamiento y control, lo cual implica una detallada planificación *ex ante*. También, se han argumentado consideraciones éticas asociadas a excluir a las unidades del grupo de control de los potenciales beneficios asociados al tratamiento.

Además, en los experimentos aleatorios bajo contextos sociales no siempre se consigue un cumplimiento del 100% de las asignaciones a pesar de que el evaluador haga todo lo posible. En este caso sólo es posible estimar una "intención de tratamiento", ya que se puede ofrecer el programa, pero no se puede obligar a participar en él. Cuando existe una asignación aleatoria, pero la participación es voluntaria, la estimación del impacto no es válida para toda la población, sino solamente para el subgrupo de los que sí aceptan participar. Una solución a esto es posible mediante el uso de variables instrumentales (ver capítulo 5.5).

5.1.2 Formulación Estadística

El caso típico de aleatorización con una muestra de la población de N unidades es que a $N_1 < N$ unidades se les asigne aleatoriamente el tratamiento y que $N_0 = N - N_1$ unidades sean asignadas al grupo de control. Sin embargo, existen variaciones como una aleatorización por pares, en la cual en una etapa inicial las unidades son emparejadas de a pares, y en una segunda etapa se escoja aleatoriamente una unidad en cada par para asignarle el tratamiento. Otra alternativa más sofisticada es una estratificación de la población, y luego, una aleatorización del tratamiento dentro de los estratos.

El impacto se mide luego de terminado el tratamiento mediante el promedio del resultado de interés entre las unidades que pertenecen al grupo de tratamiento y el promedio del resultado entre las unidades que pertenecen al grupo de control, este es el efecto promedio del tratamiento. Si la diferencia de medias entre ambos grupos resulta estadísticamente significativa se dice que el programa o política tiene un efecto (positivo o negativo) sobre el resultado.

$$\tau_{ATE} = E\,[Y_i(1) - Y_i(0)]$$

Si una proporción p de cada grupo no es voluntaria y además, las características de estas unidades no están relacionadas con el tratamiento, el estimador del impacto sobre los voluntarios será:

$$(1\text{-}p)\cdot E\,[Y_i(1) - Y_i(0)]$$

La expresión anterior es una fracción del efecto promedio del tratamiento (τ_{ATE}), por lo cual si la proporción no voluntaria es observable es posible identificar el τ_{ATE} y la asignación aleatoria podría seguir siendo una buena alternativa para estimar el impacto. Desafortunadamente, la falta de disposición para ser voluntario probablemente no es homogénea entre el grupo de tratamiento y grupo de control en la mayoría de los experimentos sociales.

5.1.3 ¿Cuándo utilizar?

La aleatorización de un programa o política puede considerarse un criterio de asignación equitativo cuando la falta de recursos no permite ampliar el tratamiento a toda la población, ya que permite asignar el tratamiento sólo a una parte de ella. También, cuando los recursos son insuficientes se puede realizar una aleatorización del tiempo en el cual se puede optar al tratamiento.

En general, los experimentos aleatorios tienden a aprovechar el hecho que exista una alta demanda por el tratamiento o cuando es necesario implementar el tratamiento a las unidades gradualmente en el tiempo hasta que cubra a toda la población elegible. En ambos casos, es más fácil asegurar que todas las unidades elegibles tienen la misma probabilidad de participar.

El experimento aleatorio se ha utilizado en evaluaciones rigurosas del impacto de programas a gran escala. Pero también se puede usar cuando se desea probar políticas o programas pilotos o costosos cuyos resultados son desconocidos.

Para lograr la validez interna del experimento aleatorio se debe chequear que no existan diferencias estadísticamente significativas en todas las características observables (apli-

cando un test de diferencia de medias para cada característica). También se requiere que todas las unidades tratadas sigan correctamente el protocolo del tratamiento (por ejemplo, asistir a todas las clases de educación ambiental) o que las unidades del grupo de control no tengan acceso a este tratamiento (por ejemplo, el material didáctico de las clases de educación ambiental no podría ser compartido por los asistentes con sus vecinos del barrio).

Otro problema, es que con el transcurso del tiempo algunas unidades del grupo de tratamiento y/o grupo de control abandonan el experimento, este "desgaste muestral" o "atrición de la muestra" puede sesgar las estimaciones del resultado si existen diferencias significativas entre las características de los que abandonan o permanecen en ambos grupos.

Por lo anterior, se requiere un monitoreo permanente para evitar situaciones que puedan afectar la validez interna del experimento. También es necesario incorporar en la planificación y diseño una anticipación a los efectos del desgaste muestral, externalidades u otros factores que puedan sesgar los resultados con el objetivo de minimizar sus efectos.

Aun cuando se haya diseñado una aleatorización del tratamiento, este tipo de técnica no debería ser utilizada si se demuestra que las características de las unidades del grupo de tratamiento y grupo de control previo a la aplicación del tratamiento son estadísticamente distintas. Tampoco debería utilizarse si finalmente en la ejecución la asignación del tratamiento no fue aleatoria.

Además, la validez externa del experimento aleatorio requiere que la muestra de unidades sea representativa de la población a la cual se pretende extrapolar los resultados, o bien, que las condiciones experimentales puedan reproducirse en otros contextos. Por ejemplo, el efecto de la aplicación de una regulación específica de un PDA en una ciudad con un número significativo de episodios críticos de emergencia o pre – emergencia podría no ser representativo de los efectos de ampliar esta política a nivel nacional.

Si un programa tiene inscripción abierta o universal es posible realizar una promoción del programa sobre ciertas unidades. Si la promoción fue aleatoria se pueden realizar estimaciones válidas del impacto de un programa en la medida que en los grupos con y sin promoción sean equivalentes en términos estadísticos previo a la promoción, que la promoción sea capaz de aumentar la inscripción en el programa (comparando las tasas de inscripción entre los grupos con y sin promoción) y, además, que la promoción no afecte los resultados de interés.

Los experimentos aleatorios son muy utilizados en bioestadística. Por ejemplo, la *Food and Drug Administration* en Estados Unidos requiere evidencia de este tipo para aprobar medicamentos o procedimientos médicos. Sin embargo, estos son mucho menos utilizados en experimentos sociales, lo cual se puede atribuir en parte a que las investigaciones de interés históricamente han involucrado programas de educación o programas laborales en los cuales es prácticamente imposible hacer experimentos ciegos (en los cuales la unidad no conoce al grupo que fue asignado pero si lo sabe el investigador) o de doble ciego (en los cuales la unidad no conoce al grupo que fue asignado ni tampoco lo sabe el investigador) abriendo la posibilidad de efectos placebo. Además no es posible aislar el efecto de las interacciones entre unidades pertenecientes al grupo de tratamiento y control.

No obstante, en años recientes ha existido un número creciente de experimentos aleatorios en países en desarrollo. Estos incluyen experimentos a gran escala como el programa Progresa en México (Schultz, 2001) y otros a escalas más pequeñas (Miguel & Kremer, 2003; Banerjee, Duflo, Cole, & Linden, 2007; Duflo & Hanna, 2006 y; Olken, 2007).

5.1.4 Ventajas/ Desventajas

Ventajas	Desventajas
• No se requieren técnicas econométricas sofisticadas de estimación para estimar el impacto (diferencia de medias). • Es un método robusto para la evaluación de impacto ya que el grupo de control es un subconjunto aleatorio de la población elegible. • Fuerte validez interna. • Experimentos bien diseñados pueden mejorar el conocimiento que existe sobre la aplicación de un programa o política. • Reduce los requerimientos de datos respecto a otras técnicas no experimentales, ya que sólo se requiere información de los resultados después de aplicado el tratamiento para el grupo de tratamiento y grupo de control. • Es útil en la aplicación de programas piloto o en aquellos en los cuales su implementación, en términos de cobertura, es gradual en el tiempo.	• En las evaluaciones de impacto bajo contextos sociales es difícil asegurar que las condiciones experimentales se hayan cumplido. • En los experimentos bajo contextos sociales no siempre se consigue el 100% de cumplimiento en los criterios de selección. • Se requiere un tamaño muestral mínimo que permita detectar el impacto del tratamiento, por lo cual en la medida que existe mayor variabilidad en el resultado esperado y/o nivel de confianza estadística mayor es el tamaño muestral requerido. • Es muy costoso e intensivo en tiempo. • Algunos programas o políticas son voluntarios, por lo cual las unidades elegidas aleatoriamente pueden decidir no participar en el programa. • Las unidades que deseen participar y no son escogidas aleatoriamente podrían sufrir un efecto de decepción alterando su conducta. • Es difícil en términos éticos excluir a unidades de una política que mejore el bienestar de la población. • Las unidades podrían decidir abandonar el experimento.

5.1.5 Ejemplo(s) de aplicación

Para ilustrar el método se puede mencionar un trabajo que estima los beneficios económicos obtenidos por los granjeros a partir de un programa llamado *Reuters Market Light (RML)* que entrega a sus teléfonos móviles un servicio de información sobre datos del mercado de productos agrícolas y clima (Fafchamps & Minten, 2012). Este estudio realizó un experimento aleatorio a 100 villas de Maharashtra en India que no estuvieran previamente afectadas por campañas de marketing de este servicio informativo. Además, se incluyó en el experimento solo a granjeros que al momento de la encuesta de línea base tuvieran celular. Se implementaron dos regímenes de tratamiento. En el primer tratamiento a todos los granjeros de cada villa se les ofreció el servicio de mensajería de forma gratuita. En el segundo tratamiento se les ofreció el servicio a un pequeño subconjunto de los granjeros de cada villa. El propósito del segundo tratamiento era evaluar si se generaban externalidades a otros granjeros de la misma villa. Los resultados mostraron que no había evidencia de un efecto sobre el precio que recibieron los granjeros por sus productos, pérdidas de cultivos por tormentas o el cambio de cultivos. Sin embargo, se encontró evidencia que los granjeros cambiaron los lugares donde vendieron sus cultivos.

En Chile fue realizada una evaluación experimental al Programa de Apoyo al Microemprendimiento (PAME) del Ministerio de Desarrollo Social[15]. Este programa tiene por objetivo entregar herramientas para que los hogares vulnerables superen la condición pobreza a través del desarrollo de emprendimientos. Para asegurar la participación de los no tratados, se les comunicó que igualmente recibirían el beneficio posterior a la investigación, y además, se entregó incentivos (*gifcards*) para facilitar la entrega de información. Los resultados preliminares de los participantes que postularon en el año 2010 y que cumplieron las condiciones de admisibilidad mostraron que luego de 10 meses del egreso tenían un 81% de ocupación versus un 69% de los que no participaron, y además, el programa aumentó en un 27% los ingresos percibidos. También, se observó que los egresados aumentaron en 40% la probabilidad de tener un emprendimiento y aumentó en 29% la probabilidad de que se generaran emprendimientos en el hogar.

15. http://www.fosis.cl/index.php/estudios-y-evaluaciones/2420-evaluacion-experimental-del-programa-de-apoyo-al-microemprendimiento-pame

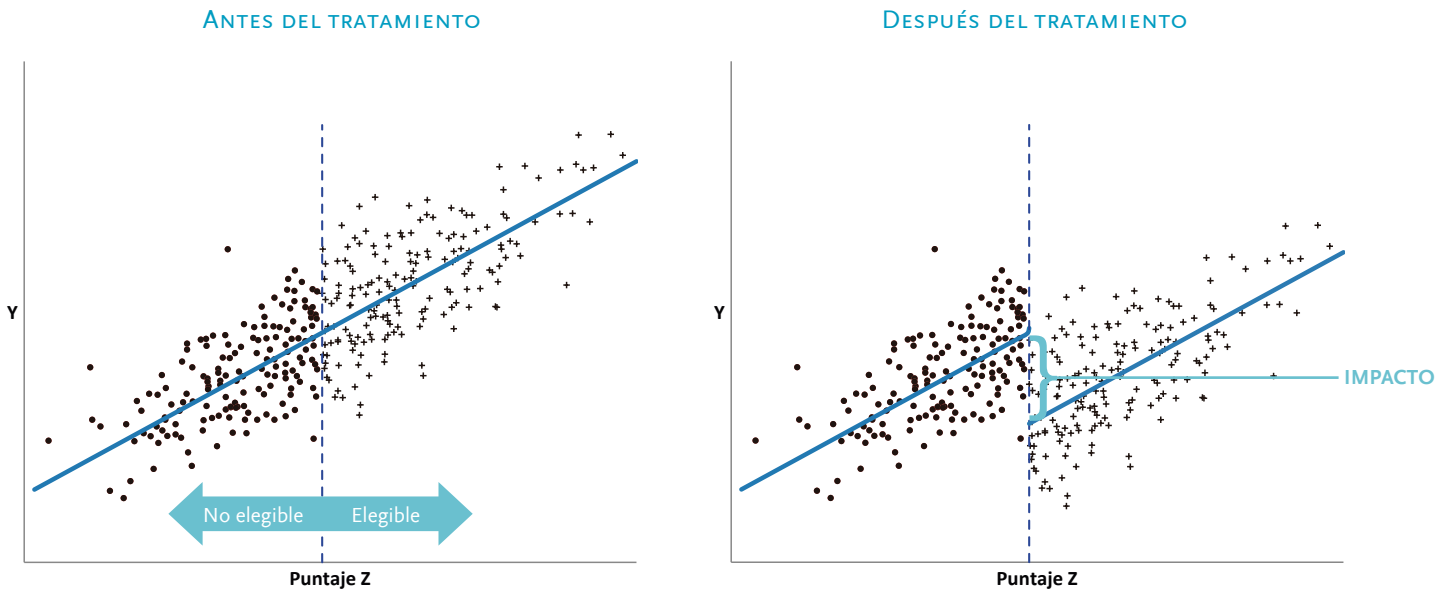
5.2 REGRESIÓN DISCONTINUA

5.2.1 Descripción

Esta técnica aprovecha los diseños de políticas o programas para **replicar localmente las condiciones aleatorias de un experimento.**

La formación de la variable de puntaje (Z) categoriza a los participantes del programa, con el supuesto fundamental que dos postulantes con un Z parecido son comparables entre sí. De esta manera, **si las unidades eran similares previo al programa y fueron sometidas a los mismos factores externos, entonces es posible atribuir cualquier diferencia en los resultados (efecto local) a la aplicación del programa.**

Figura 5-3 Representación gráfica de regresión discontinua antes y después del tratamiento



Fuente: Gertler et al. (2011)

El impacto se estima por medio de la diferencia en los resultados posteriores a la aplicación del programa entre las unidades que se encuentran próximas al puntaje de corte. El impacto estimado con esta técnica se conoce como efecto promedio local del tratamiento (LATE).

La ventaja de este proceso es que las unidades no pueden controlar la variable de asignación del tratamiento en la vecindad del punto de corte, que ocurre específicamente cuando la probabilidad de tratamiento cambia discontinuamente sobre la variable Z. Con la regresión discontinua el efecto del tratamiento puede ser medido por la media condicional del resultado en el valor límite por la izquierda y derecha del puntaje de corte, por lo cual un análisis gráfico puede ser muy útil para visualizar si este método es adecuado como herramienta de identificación.

Si las unidades están muy próximas al puntaje de corte son muy semejantes, la comparación de los resultados sería tan buena como si la asignación del grupo de tratamiento y grupo de control hubiese sido aleatoria. Sin embargo, el impacto del programa en torno al puntaje de corte no se puede extrapolar a unidades cuyo puntaje de corte esté más distante.

Dado que la evaluación se realiza solo cerca del puntaje de corte, se requiere de muchas observaciones bajo y sobre este límite para que las estimaciones estadísticas tengan suficientes grados de libertad. Aun cuando se puede ampliar el margen respecto al límite para incluir a más observaciones, a mayor margen las unidades serán menos similares.

La forma funcional específica para modelar la relación entre la variable de resultado y el puntaje no es tan relevante en la cercanía del puntaje de corte. No obstante, si se asume una relación lineal entre ambas cuando en realidad es no lineal el sesgo podría ser mayor.

Existen dos elementos que pueden invalidar los resultados de la regresión discontinua independientemente del tipo de diseño utilizado:

- Que existan variables que varíen en forma conjunta con el puntaje (ej: la edad), por lo que el impacto calculado ya no se atribuiría solo al tratamiento.
- Que las unidades realicen acciones para manipular su puntaje.

Una forma de testear estos elementos es observar si existen discontinuidades en los valores promedios de las características de las unidades alrededor del puntaje de corte. Si se observa discontinuidad en estas características seguir utilizando la técnica no sería adecuado para estimar el impacto del programa.

5.2.2 Formulación Estadística

Si la participación en el programa P depende del puntaje Z y de otros factores no observables, entonces el mecanismo para identificar el efecto del programa es la discontinuidad en la probabilidad de participar en el puntaje de corte ($\overline{Z}$). La discontinuidad puede ser aguda (*sharp*) o difusa (*fuzzy*) dependiendo de si la participación en el programa es una función determinística (sólo depende del puntaje de corte) o no determinística de Z (depende del puntaje de corte y de factores no observables).

En el diseño *sharp* el puntaje de corte $\overline{Z}$ determina completamente la participación, por lo cual la probabilidad de participar en el programa para todas las unidades es cero o uno, dependiendo del puntaje de la unidad respecto al puntaje de corte (por lo cual no es muy utilizado en la práctica para evaluación en contextos sociales). Esto implica que la selección se basa sólo en características observables, así el impacto del tratamiento en términos locales es probablemente independiente del proceso de selección. Por esto, en ausencia de traslape entre tratados y controles se debe asumir que existe continuidad de la variable de resultado en Z para asegurar que los no tratados a un lado del puntaje de corte son contrafactuales adecuados para los tratados en el lado opuesto.

$$\tau_{RD}(\overline{Z}) = E[Y_i(1) \mid Z_i = \overline{Z}^-] - E[Y_i(0) \mid Z_i = \overline{Z}^+]$$

En este caso, τ_{RD} puede ser interpretado como el impacto del tratamiento sobre una unidad seleccionada aleatoriamente en el punto de corte.

En un contexto de regresión, el modelo puede ser estimado al definir la variable dicotómica, o *dummy*, P_i que adopta el valor 1 si la unidad i tiene un puntaje igual o mayor a $\overline{Z}$ y el valor 0 si la unidad tiene un puntaje menor a $\overline{Z}$. Entonces el estimador del efecto del tratamiento condicionando por características observables X_i (aunque no es necesario incluirlas ayuda a reducir la varianza del estimador) puede ser obtenido como:

$$Y_i = \alpha + \beta \cdot X_i + \tau_{RD} \cdot P_i + \varepsilon_i$$

Alternativamente, se puede relajar el supuesto de linealidad con una función no lineal como por ejemplo una función polinomial:

$$Y_i = \alpha + \beta \cdot X_i + \gamma \cdot X_i^2 + \delta \cdot X_i^3 + \tau_{RD} \cdot P_i + \varepsilon_i$$

Incluso se pueden incorporar diferentes tendencias a cada lado del puntaje de corte al agregar términos de interacción entre las características observables y la variable dicotómica de participación.

$$Y_i = \alpha + \beta \cdot X_i + \gamma \cdot X_i^2 + \delta \cdot X_i^3 + \tau_{RD} \cdot P_i + \tau_{RD}' \cdot P_i \cdot X_i^2 + \tau_{RD}'' \cdot P_i \cdot X_i^3 + \varepsilon_i$$

En términos prácticos, se parte con especificaciones simples, y luego, se van incorporando variables polinomiales de grado superior, chequeando los test estadísticos, el ajuste del modelo y la robustez de los resultados.

Otra alternativa más compleja es que la función sea estimada por métodos no paramétricos, cuyos resultados pueden servir para contrastarlos con las estimaciones paramétricas mencionadas previamente.

En el segundo tipo de diseño de regresión discontinua, conocido como diseño *fuzzy*, la participación no es completamente determinada por el puntaje Z, ya que asume que existen otros factores no observables que determinan la participación. En este contexto, la participación o no de una unidad podría ocurrir a ambos lados del puntaje de corte. En la práctica se requiere que la discontinuidad del diseño *fuzzy* sea suficientemente grande como para ser visualizada gráficamente. Sin embargo, este diseño pierde su atractivo y simplicidad, ya que sólo un subgrupo de unidades se mueve al estado de tratamiento en el puntaje de corte y, además, asume que no es posible que factores no observables se relacionen localmente con la variable de resultado, el cual es un supuesto demasiado fuerte incluso a nivel local.

El estimador en este caso se puede obtener de la siguiente forma:

$$\tau_{RD}(\bar{Z}) = \frac{E[Y_i(1) \mid Z_i = \bar{Z}^-] - E[Y_i(0) \mid Z_i = \bar{Z}^+]}{P(Z_i = \bar{Z}^-) - P(Z_i = \bar{Z}^+)}$$

5.2.3 ¿Cuándo utilizar?

La regresión discontinua se puede utilizar cuando los programas poseen un índice continuo de elegibilidad y un puntaje de corte para determinar quiénes tienen o no derecho a participar.

La técnica es útil si la pregunta de interés es si se debe expandir marginalmente el programa. No obstante, debido a que sólo estima efectos locales cerca del puntaje de corte no es una técnica útil cuando se intenta decidir si continuar con el programa o implementarlo desde un área específica hacia todo el país, o bien, expandirlo a un grupo de población más grande, ya que en ambos casos lo que se requiere estimar es el efecto promedio del tratamiento para toda la población.

Para la validez interna de la técnica se requiere que las unidades no puedan manipular completamente su puntaje. Si la manipulación es imperfecta, es posible demostrar que la asignación del tratamiento se puede considerar aleatoria en el punto de corte manteniendo la validez (Lee, 2008), aunque otros autores afirman que se debe reinterpretar el efecto estimado (Imbens & Wooldridge, 2009).

Para que la técnica pueda ser utilizada deberían chequearse tres análisis gráficos (Imbens & Lemieux, 2008). Primero, observar la discontinuidad y relación de la forma funcional de la variable de resultado en el punto de corte, incluyendo en el grafico los valores predichos por el modelo después del puntaje de corte de tal modo de asegurar que se traslapen con los valores observados. Segundo, graficar diferentes características observables contra la variable de puntaje. La validez podría ser criticada si alguna de estas características presentan una discontinuidad en el puntaje de corte. Tercero, un gráfico de la distribución de la variable de puntaje podría ayudar a observar algún tipo de manipulación, representado por un salto al lado derecho del puntaje de corte.

Esta técnica no debería ser utilizada si la cantidad de observaciones no es suficientemente grande para que permita identificar visualmente un salto en el umbral del puntaje de corte, ni tampoco si se observan saltos en las características entre el grupo de tratamiento y grupo de control previo a la asignación del tratamiento.

5.2.4 Ventajas/ Desventajas

Ventajas	Desventajas
• Al utilizar un índice de elegibidad no es necesario que el evaluador deba excluir a un grupo de unidades para generar un grupo de control. • No es necesario controlar por otras variables observables. • El impacto puede ser identificado sin asumir supuestos sobre formas funcionales, por lo que perfectamente se pueden utilizar técnicas de regresión no paramétricas que permiten estimar formas funcionales flexibles.	• El impacto del programa estimado sólo es válido en la cercanía del límite del puntaje de corte y no puede extrapolarse al impacto de toda la población elegible. • La necesidad de moverse lejos del puntaje de corte para asegurar un tamaño de muestra grande introduce un sesgo ya que las unidades del grupo de tratamiento y grupo de control no son necesariamente comparables. • Requiere un tamaño de muestra bastante grande para tener suficientes observaciones cerca del puntaje de corte. • Si se ha producido algún incumplimiento del puntaje de corte para asignar el tratamiento el método pierde validez y se deberían utilizar técnicas más avanzadas para corregir esta discontinuidad difusa. • Existe la posibilidad que las unidades manipulen su puntuación con el objetivo de eludir o aumentar su probabilidad de participación lo cual invalidaría los resultados. • No deben existir otras variables o características de las unidades que varíen de forma conjunta con el puntaje de corte.

5.2.5 Ejemplo(s) de aplicación

Un informe desarrollado para la Unión Europea construye un indicador de innovación relacionado al cambio climático a nivel de firma (a partir de 700 entrevistas a firmas en 6 países europeos) con el objetivo de estimar el impacto del sistema de transacción de emisiones europeo (Martin, Muûls, & Wagner, 2012). Los autores aprovechan que algunos sectores industriales quedan exentos de la regulación ambiental por el valor del indicador de comercio que poseen, por lo cual utilizan este criterio para estimar una regresión discontinua. Sus resultados muestran que la mayoría de las firmas de la muestra fomentan la innovación relacionada al cambio climático, pero existen diferencias relevantes entre países, incluso después de controlar por estructura industrial. Además, aquellas firmas que esperan recibir una menor asignación de permisos en la nueva fase del sistema de transacción de emisiones europeo tienden a innovar más.

5.3 DIFERENCIAS EN DIFERENCIAS Y DATOS DE PANEL

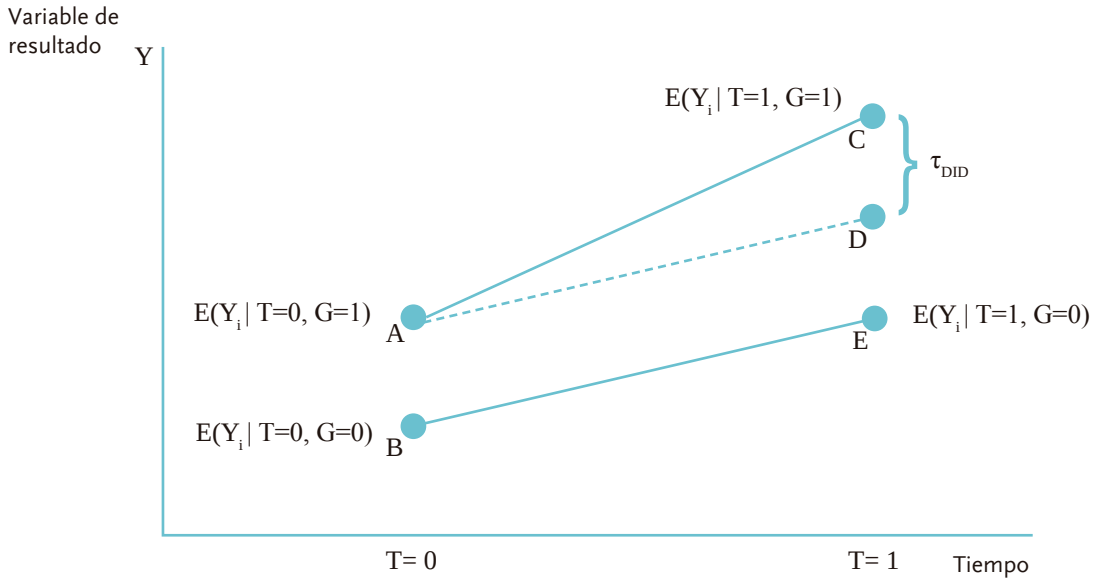
5.3.1 Descripción

El método de "diferencias en diferencias" es conocido por este nombre ya que compara la diferencia en el efecto promedio para el grupo de tratamiento antes (T=0, G=1) y después del tratamiento (T=1, G=1) con respecto a la diferencia en el efecto promedio para el grupo de control antes (T=0, G=0) y después del tratamiento (T=1, G=0). Es un caso particular de datos de panel cuando existen solo dos periodos de observaciones sobre las unidades analizadas. También es conocido como "experimento natural" ya que intenta encontrar un grupo de control generado naturalmente.

La diferencia de los resultados antes y después del grupo de tratamiento (primera diferencia) es capaz de eliminar factores no observables que permanecen constantes en el tiempo para dicho grupo, ya que se compara al grupo con sí mismo. Sin embargo, aún persisten los factores externos temporales (ver capítulo 3.2.1).

La segunda diferencia entre los resultados antes y después de un grupo de control es capaz de eliminar los factores externos que varían con el tiempo y que interfieren en el grupo de tratados. Por lo tanto, al sustraer ambos resultados antes-después el método es capaz de condicionar por los efectos no observables individuales y efectos temporales lo cual permite generar una mejor estimación del impacto del tratamiento.

Figura 5-4. Efecto gráfico del tratamiento bajo diferencias en diferencias

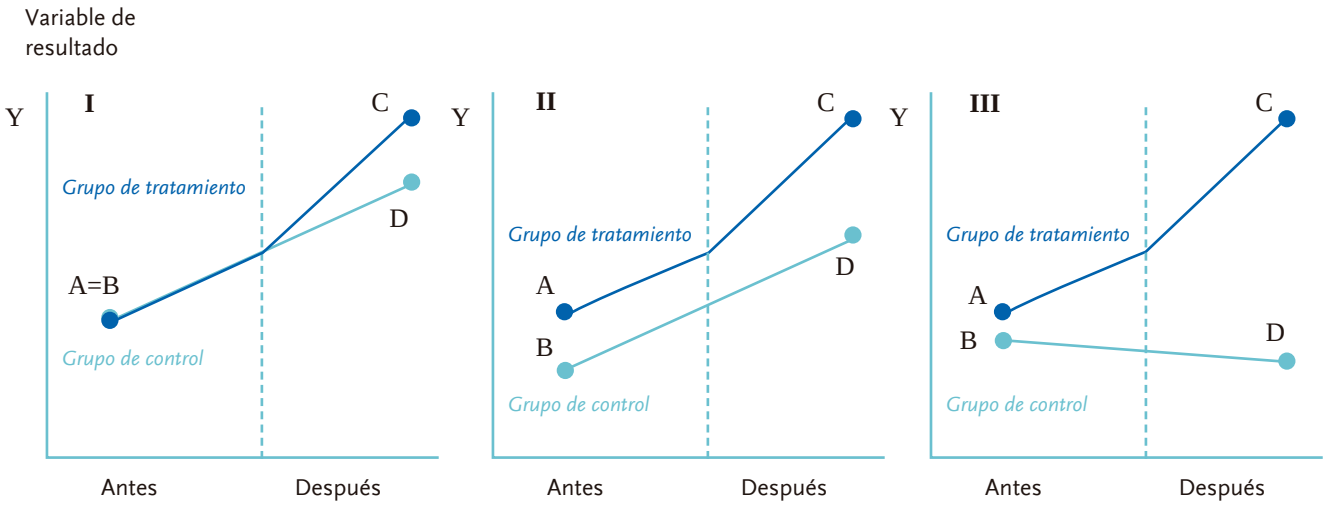


Fuente: Adaptado de Stock & Watson (2003)

Este método permite encontrar el efecto promedio sobre los tratados (ATT) al asumir que los factores no observables permanecen constantes en el tiempo (existen efectos temporales comunes a través de los grupos y no cambia sistemáticamente la composición dentro de cada grupo). Este supuesto permite superar las principales limitaciones de otros métodos estadísticos como la regresión discontinua y el pareamiento (*matching*). Además, es un supuesto no tan restrictivo, ya que no requiere que los resultados promedio de los grupos de tratamiento y control sean equivalentes antes del tratamiento, sino que difieran en una magnitud constante y **que sigan las mismas tendencias a través del tiempo.**

Aunque el método de diferencias en diferencias permite condicionar las características constantes en el tiempo entre el grupo de tratamiento y el grupo de control, **no elimina el sesgo producido por aquellas características que varían a lo largo del tiempo entre el grupo de tratamiento y el grupo de control.** Por ejemplo, pueden surgir diferentes tendencias temporales entre el grupo de tratamiento y control si los grupos están localizados en dos áreas geográficas o mercados diferentes. En la Figura 5-5 el punto A representa el indicador Y antes de la aplicación del programa en el grupo de tratamiento y el punto B representa el mismo indicador en el grupo de control, mientras los puntos C y D representan el indicador después del programa para el grupo de tratamiento y grupo de control, respectivamente. La evaluación de diferencias en diferencias compara el cambio en el grupo de tratamiento (Y_C-Y_A) respecto al cambio en el grupo de control (Y_D-Y_B). Esta comparación es válida mientras la tendencia en el indicador sea similar en el grupo de tratamiento y grupo de control antes del programa, pero si el grupo de control sigue una tendencia diferente no será un contrafactual válido (Figura 5-5 III).

Figura 5-5 Uso correcto e incorrecto del método de diferencias en diferencias



Fuente: Elaboración propia

A pesar de que es imposible demostrar el supuesto que el grupo de tratamiento y grupo de control siguen tendencias iguales en ausencia del tratamiento, una forma de evaluar qué tan razonable es este supuesto sería comparar la tendencia de ambos grupos antes del tratamiento, **lo cual involucra disponer de al menos dos rondas de datos antes del tratamiento.** Una segunda opción sería realizar una estimación de diferencias en diferencias con un grupo de tratamiento falso que se sabe no ha sido afectado por el tratamiento, o alternativamente, ocupar diferentes grupos de control.

El método no controla por efectos temporales no observables específicos a nivel de unidades, por ejemplo si se genera una caída en el indicador de interés justo antes que comience el tratamiento, es razonable esperar que el indicador crezca entre los tratados incluso sin haberse aplicado el tratamiento (Ashenfelter, 1978).

Recientemente, se ha propuesto una generalización al método estándar de diferencias en diferencias conocido como "cambios en cambios" (Athey & Imbens, 2006), al demostrar que el método de diferencias en diferencias es un caso particular del método de cambios en cambios. A pesar de su atractivo no es muy utilizado porque el código para implementarlo no ha sido incorporado en *softwares* estadísticos populares. No obstante, este método permite que los efectos temporales y del tratamiento difieran sistemáticamente a través de las unidades, estimar toda la distribución contrafactual de los efectos del tratamiento sobre el grupo de tratamiento, y también, la distribución de los efectos sobre el grupo de control, incluso en el caso que ambas distribuciones varíen de forma arbitraria. Además, este método entrega una pequeña guía sobre cómo serían los efectos de la política en el caso que fuera aplicada al grupo de control.

5.3.2 Formulación Estadística

La estimación del método de diferencias en diferencias bajo el esquema más simple es con dos periodos (T = 0 y T = 1) y dos grupos. El impacto del programa es calculado como la diferencia del cambio observado antes y después en el grupo de tratamiento (G = 1) y el grupo de control (G = 0).

$$\tau_{DID} = (E[Y_i \mid G_i = 1,\ T_i = 1] - E[Y_i \mid G_i = 1,\ T_i = 0]) \\ - (E[Y_i \mid G_i = 0,\ T_i = 1] - E[Y_i \mid G_i = 0,\ T_i = 0])$$

En un contexto de regresión, esto es equivalente a la siguiente especificación:

$$Y_i = \alpha + \beta \cdot G_i + \delta \cdot T_i + \tau_{DID} \cdot I_i + \varepsilon_i$$

El efecto del tratamiento τ_{DID} puede ser calculado a través del coeficiente estimado para la interacción entre el indicador para el periodo "1" y el grupo "1" ($I_i = G_i \cdot T_i$).

El método se puede extender fácilmente al caso con múltiples grupos, periodos de tiempo y también incluir variables de control. Sea T el número de periodos de tiempo y G el número de grupos, entonces la regresión sería:

$$Y_i = \alpha + \sum \beta_t \cdot I_{Ti=t} + \sum \delta_g \cdot 1_{Gi=g} + \tau_{DID} \cdot I_i + \gamma \cdot X_i + \varepsilon_i$$

Donde I_i ahora es un indicador del tratamiento para la unidad i que está en el grupo g y en el periodo t. Por otra parte, X_i es una variable o un vector de características observables.

Esta versión del método permite imponer restricciones testeables sobre los datos. Por ejemplo, si los dos grupos no fueron expuestos al tratamiento en dos periodos consecutivos, entonces el efecto del tratamiento debería ser cero, lo cual puede testearse con un test típico de significancia estadística.

El estimador de diferencias en diferencias es un caso particular de un estimador de efectos fijos (datos de panel). El estimador de efecto fijo elimina los factores no observables al estimar una regresión lineal utilizando como la variable dependiente las diferencias en los resultados entre dos periodos contiguos $\Delta Y_{it} = Y_{it} - Y_{it-1}$ (estimador de primeras diferencias), al sustraer el promedio individual a través del tiempo, o alternativamente al agregar variables *dummy*[16] individuales (método conocido como LSDV).

El modelo de cambios en cambios (changes-in-changes) asume que las características no observables del individuo serán las mismas en un periodo dado independientemente del grupo al cual pertenezca. Para ello estima de forma no paramétrica la distribución de probabilidad de los resultados en el grupo de control tanto para el período antes como para el período después, así es capaz de estimar el cambio ocurrido en el grupo de control ocurrido a través del tiempo. Luego, asumiendo que la distribución de probabilidad de los resultados en el grupo de tratamiento debería haber experimentado el mismo cambio en ausencia del tratamiento, es posible estimar la distribución contrafactual para el grupo de tratamiento en el segundo periodo (con un enfoque similar se puede estimar el efecto del tratamiento sobre el grupo de control). Así es posible estimar el efecto del tratamiento sobre cualquier cuantil de la distribución (Imbens & Wooldridge, 2009).

Por otra parte, el método de cambios en cambios relaja el modelo lineal aditivo al asumir que en ausencia de la intervención el resultado satisface $Y_i(0) = h_0(U_i, T_i)$ con $h_0(u,t)$ creciente en u. La variable aleatoria U_i representa todas las características no observa-

16. Las variables dummy son variables cualitativas, también conocidas como indicativas, binarias o dicotómicas.

das del individuo i y, además, incorpora la idea que el resultado de un individuo con U_i=u será el mismo en un periodo de tiempo dado independientemente de a qué grupo pertenece. La distribución de U_i se permite que varíe a través de los grupos pero no a través del tiempo dentro de los grupos (Imbens & Wooldridge, 2009).

El efecto promedio del tratamiento para el grupo de tratamiento en el segundo periodo es:

$$\tau_{CIC} = E[Y_i(1) - Y_i(0) \mid G_i = 1,\ T_i = 1]$$

El primer término de esta expresión puede obtenerse directamente de los datos porque:

$$E[Y_i(1) \mid G_i = 1,\ T_i = 1] = E[Y_i \mid G_i = 1,\ T_i = 1]$$

Sin embargo, la dificultad está en estimar el segundo término. Bajo ciertos supuestos es posible demostrar que la distribución completa de $Y(0)$ dado G_i=1 y T_i=1 se puede identificar a través de la siguiente ecuación $F_{Y11}(y) = F_{Y10}(F^{-1}_{Y00}(F_{Y01}(y)))$, donde $F_{Ygt}(y)$ denota la función de distribución de Y_i dado G_i=g y T_i=t (Athey & Imbens, 2006). Así, el resultado esperado para el grupo de tratamiento en el segundo periodo sin tratamiento es:

$$E[Y_i(0) \mid G_i = 1,\ T_i = 1] = E[F^{-1}_{01}(F^{-1}_{00}(Y_{i10}))]$$

Finalmente, para analizar el efecto contrafactual de la intervención sobre el grupo de control se asume que en presencia de la intervención $Y_i(1) = h_1(U_i,T_i)$ con $h_1(u,t)$ creciente en u.

5.3.3 ¿Cuándo utilizar?

La técnica de diferencias en diferencias es útil cuando se ha realizado un cambio o se ha introducido una política o programa y existe disponibilidad de información a nivel de unidades antes y después de este evento. En particular, como se requiere información de un grupo de control antes y después se puede aprovechar el caso de una política o programa introducido en áreas geográficas específicas con respecto a áreas en las cuales no se ha introducido pero que poseen unidades que son afectadas por factores temporales similares. Por ejemplo, podría ser el caso de un estudio piloto de un subsidio para el cambio de calefactores a pellets, que está siendo realizado en una ciudad, pero que no se realiza en el mismo periodo en otra ciudad con características similares.

La técnica no debería ser utilizada si el grupo de tratamiento y grupo de control son afectados por tendencias temporales diferentes entre sí.

5.3.4 Ventajas/ Desventajas

Ventajas	Desventajas
• La facilidad de obtener un número grande de observaciones incrementa los grados de libertad y eficiencia de las estimaciones. • Permite controlar tanto por variables observables como no observables invariantes en el tiempo pero heterogéneas entre individuos. • Facilidad de estimación.	• Si existe algún factor que afecta las tendencias de los dos grupos de forma diferente la estimación será sesgada. • Pueden existir problemas para la recolección de información asociados a la atrición de la muestra o al error de medición. • Dificultad de conseguir datos de panel (longitudinales) una vez iniciado el tratamiento, es decir, datos que permitan el seguimiento de las mismas unidades a través del tiempo.

5.3.5 Ejemplo(s) de aplicación

Rogan *et al.* (2011) evalúan el impacto de una política ambiental en Irlanda que fue diseñada para influenciar la tendencia en la adquisición de vehículos que emitan menos CO2. Los impactos evaluados fueron emisiones de CO2, tamaño de motor, combustible, precios de los autos e ingresos generados. No obstante, existieron factores asociados a la crisis económica europea que afectaron la estimación del impacto del programa. Por ello, el estudio compara las tendencias en el periodo previo a la aplicación del impuesto y al año posterior de su aplicación aprovechando una base de datos del tipo panel. Los resultados muestran que en el primer año de la aplicación del impuesto las emisiones de los autos nuevos cayeron 13%. Sin embargo, esto no se generó por una reducción en el tamaño de los motores comercializados sino por un significativo cambio de tipo de combustible (a diésel con las nuevas normas europeas). Además, el impacto *ex post* fue mayor que el impacto estimado antes de la implementación del impuesto, aunque su recaudación fue 33% menor a la prevista.

Tanaka (2015) explora el impacto de las regulaciones ambientales en China sobre la mortalidad infantil. Estas regulaciones se remontan a 1998 cuando el gobierno chino creó las llamadas zonas de control dos, abarcando 175 localidades que superaban los estándares de contaminación permitidos; en estas zonas las industrias fueron obligadas a instalar tecnologías de control y reducir sus emisiones. Lo anterior, permitió realizar un experimento natural en el cual las localidades tienen o no la regulación. El supuesto clave para identificar el impacto de esta política con el método de diferencias en diferencias es que las localidades que no tenían regulación entregaban un contrafactual válido asociado a los cambios en la mortalidad infantil, si es que ellas hubiesen tenido la regulación. Los resultados muestran que la mortalidad infantil cayó en 20% en las ciudades tratadas donde se aplicaron las denominadas "zonas de control". Además, las mayores reducciones se obtuvieron durante el periodo neonatal y en infantes que tenían madres con bajo nivel educacional.

5.4 *MATCHING* O PAREAMIENTO

5.4.1 Descripción

Los métodos de *matching* o pareamiento crean para cada unidad tratada el escenario contrafactual mediante un emparejamiento con una unidad no tratada. La metodología utiliza las características observables de cada unidad de manera que cada par sea lo más similar posible[17].

Este método **supone que no existen diferencias no observables entre ambos grupos**, es decir, que todo lo medible es suficiente para generar el escenario contrafactual. Esto es posible si las características observables no predicen exactamente la participación[18].

En la Figura 5-6 se muestra como para cada unidad tratada por el programa (por ejemplo cada hogar al cual se le asignó un subsidio para el aislamiento térmico de la vivienda) se busca una unidad no tratada con las características más parecidas en la situación base (ingreso per cápita, equipo de calefacción, consumo de energía para calefacción, número de personas, nivel de aislación de la vivienda, entre otros).

FIGURA 5-6. EJEMPLO DE *MATCHING* PARA UN PROGRAMA DE SUBSIDIO TÉRMICO DE VIVIENDAS

HOGARES TRATADOS			
Ingreso per capita ($/mes)	Leña consumida (kg)	N° personas	Aislación vivienda
$400.000	2000	4	C
$250.000	1600	3	C
$150.000	1200	6	D
$350.000	2200	2	C
$250.000	1600	6	C
$100.000	800	3	C
$200.000	2000	4	C

HOGARES NO TRATADOS			
Ingreso per capita ($/mes)	Leña consumida (kg)	N° personas	Aislación vivienda
$155.000	1300	6	D
$640.000	3200	2	C
$200.000	2500	4	C
$390.000	2000	4	C
$350.000	1600	3	C
$580.000	900	1	C
$250.000	1600	6	C

Fuente: Elaboración propia

La Figura 5-6 también muestra la dificultad de encontrar una unidad no tratada con todas las características iguales a cada unidad tratada. Este problema puede ser más complejo si el número de características observables o los valores que pueden adoptar estas características se incrementan. Incluso si todas las variables son discretas la dimensionalidad combinada se incrementa exponencialmente con el número de características, lo cual hace casi imposible encontrar un emparejamiento para cada observación dentro de la muestra.

17. En la práctica el método utiliza técnicas estadísticas que construyen para cada unidad tratada una o varias unidades sin tratamiento, las cuales tienen las características observables lo más parecidas a la unidad que recibió el tratamiento. Así, las unidades parecidas sin tratamiento se convierten en el grupo de control para estimar la situación contrafactual.

18. Por ejemplo, si todas las mujeres están participando pero los hombres no, entonces la variable género predice exactamente la probabilidad de participación.

Una solución (no siempre factible) puede ser ampliar la muestra de unidades no tratadas para conseguir un conjunto más amplio de dónde escoger controles. Sin embargo, otra solución más sencilla fue desarrollada a través del método ***propensity score matching (PSM)*** (Rosenbaum & Rubin, 1983).

Bajo este método ya no se requiere que cada unidad tratada sea emparejada con otra unidad no tratada con un valor idéntico en todas las características observables. En realidad el problema de dimensionalidad se resuelve fácilmente al **calcular una probabilidad de propensión a participar en el programa, el *propensity score matching*, el cual es estimado a partir de todas las características observables.**

El PSM adopta un valor entre 0 y 1, y por lo tanto, trata de emparejar unidades con y sin tratamiento con un puntaje lo más parecido posible entre unidades tratadas y no tratadas. Tal como se aprecia en la Figura 5-7, existe una disminución en la dimensionalidad desde 4 características a solo el valor del PSM, pero además, es más fácil encontrar unidades no tratadas con puntajes similares a aquellos de las unidades tratadas, estas unidades semejantes se convierten en el grupo de control y se usan para estimar el contrafactual. El impacto del programa se estima comparando los resultados promedio del grupo de tratamiento con el resultado promedio del grupo de control, este último es estimado estadísticamente a partir de las características observadas.

Figura 5-7. Ejemplo de *propensity score matching*

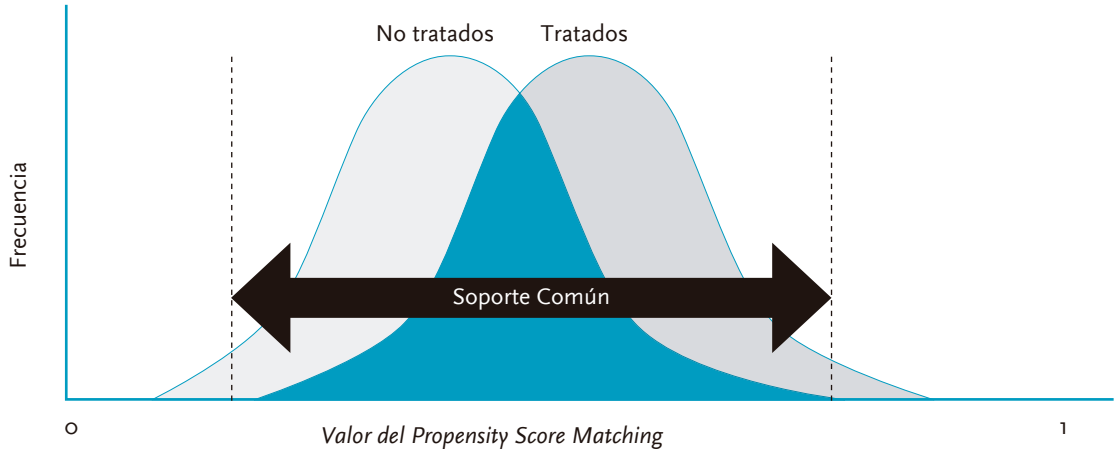
Hogares tratados					Hogares no tratados				
Ingreso per capita ($/mes)	Leña consumida (kg)	N° personas	Aislación vivienda	PSM	PSM	Ingreso per capita ($/mes)	Leña consumida (kg)	N° personas	Aislación vivienda
$400.000	2000	4	C	0,820	0,360	$155.000	1300	6	D
$250.000	1600	3	C	0,671	0,942	$640.000	3200	2	C
$150.000	1200	6	D	0,352	0,392	$200.000	2500	4	C
$350.000	2200	2	C	0,384	0,820	$390.000	2000	4	C
$250.000	1600	6	C	0,484	0,654	$350.000	1600	3	C
$100.000	800	3	C	0,271	0,912	$580.000	900	1	C
$200.000	2000	4	C	0,451	0,484	$250.000	1600	6	C

Fuente: Elaboración Propia

No obstante, puede ocurrir que para algunas unidades tratadas no se pueda encontrar una unidad no tratada que tenga puntuación suficientemente parecida. A esto se le conoce como **falta de soporte común.**

En la Figura 5-8 se observa por separado la distribución del *PSM* para las unidades tratadas y para las unidades no tratadas. La falta de soporte común se refleja en que para probabilidades cercanas a 0 y 1, no existen unidades tanto tratadas como no tratadas simultáneamente. En la práctica, es común observar una falta de rango común en los extremos de la distribución.

FIGURA 5-8. FALTA DE SOPORTE COMÚN EN EL *PROPENSITY SCORE MATCHING*



Fuente: Elaboración propia

Existen distintos algoritmos de *matching* que utilizan distintas ponderaciones para asociar el conjunto de unidades no tratadas a cada unidad tratada.

- El algoritmo más sencillo es el "vecino más cercano" o "*nearest neighbor matching*" que consiste en emparejar cada unidad tratada solo con una unidad no tratada en la que la diferencia en los puntajes del *propensity score* sea menor a cierto valor. Sin embargo, este algoritmo tiene el problema que la distancia entre los puntajes para cada par de tratados y controles puede llegar a ser grande. Si se permite el reemplazo, es decir, que los individuos no tratados puedan ser usados más de una vez en el *matching*, se genera una mejora en calidad promedio del *matching*. Además, se puede usar un cierto número de unidades no tratadas que sean más cercanas, para reducir la variabilidad de las estimaciones cuando las muestras de unidades no tratadas son pequeñas. Sin embargo, ambas opciones generan un *trade-off* entre sesgo y varianza (Caliendo & Kopeinig, 2008). El *matching* también puede especificar un radio o distancia máxima sobre la cual se puede escoger unidades no tratadas para realizar el emparejamiento, pero en este caso es difícil conocer *a priori* cual es un nivel de tolerancia razonable.

- Otra alternativa más sofisticada, es la utilización de *kernel*, el cual usa todas las unidades no tratadas y no solo la unidad más cercana. El *kernel* asigna una ponderación positiva a todas las observaciones, aunque diferentes esquemas de ponderaciones generan diferentes estimadores. Un *kernel* uniforme asigna un mismo peso a cada observación dentro de una cierta distancia o vecindario, mientras que otros tipos utilizan ponderaciones dependiendo de la distancia entre la unidad tratada y las unidades no tratadas que son emparejadas. En términos simples, el *kernel* asigna pesos que dependen de la distancia entre cada individuo del grupo de control respecto a cada individuo del grupo de tratamiento, por ello los pesos de los individuos del grupo de control cercanos a un individuo del grupo de tratamiento serán altos y los pesos serán menores con individuos más distantes. Su mayor ventaja es una menor varianza porque se utiliza más información.

El uso de *propensity score matching* requiere técnicas de *bootstraping* para calcular el error estándar del efecto del tratamiento. Sin embargo, el uso de estas técnicas no aseguran que el algoritmo del vecino más cercano entregue estimadores consistentes. Por otra parte, el uso de *kernel* reduce la variabilidad del estimador y entrega estimaciones más precisas que el *matching* del vecino más cercano. Sin embargo, no existe un algortimo de *matching* que sea siempre el mejor en cualquier situación, ya que el desempeño varía caso a caso y depende de forma muy importante de la estructura de datos, por lo cual se sugiere intentar diferentes algoritmos (ver Caliendo & Kopeinig, 2008).

A pesar que los diferentes algoritmos de *matching* utilizan de forma diferente la información de la muestra, los impactos estimados no deberían ser tan dependientes del algoritmo utilizado. Por ello, se requiere realizar análisis de sensibilidad basado en diferentes algoritmos para asegurar la robustez de los resultados. Además, es importante mostrar que las características observables son similares sobre las unidades efectivamente utilizadas en el *matching*, así como también que las distribuciones de los *propensity score* son similares usando test estadísticos como el test de Kolmogorov-Smirnov.

5.4.2 Formulación Estadística

Formalmente, el estimador del *matching* es la diferencia de los resultados entre las unidades tratadas y no tratadas con similares características usando ponderadores w_i de la distribución de las características X entre los tratados. En este caso T y C representan el grupo de tratamiento y grupo de control, $\hat{w}_{ij}$ es la ponderación que se le atribuye a la unidad no tratada j para la unidad tratada i, y w_i es la nueva ponderación que reconstruye la distribución del resultado para la muestra de tratados finalmente utilizada.

$$\tau_M = \sum_{i \epsilon T} \{ y_i - \sum_{j \epsilon C} \hat{w}_{ij} y_i \} w_i$$

Alternativamente, se demostró que si se conocen todos los factores relevantes que determinan la participación, el procedimiento de *matching* puede basarse en la probabilidad condicional a participar o *propensity score* (Rosenbaum & Rubin, 1983):

$$Prob\ (X) = Prob\ (P=1|\ X)$$

El estimador del *propensity score matching* se obtiene con el siguiente algoritmo (Dehejia & Wahba, 1999):

i. Partir de un modelo *logit* o *probit* que cumpla con el principio de parsimonia[19]. La variable dependiente es si la unidad participó o no en el programa (P = 1 ó P = 0), y las variables explicativas corresponden a todos los factores observables (X).
ii. Ordenar las observaciones de acuerdo a *propensity score* estimado (menor a mayor).
iii. Estratificar las observaciones en grupos en donde los propensity score entre tratados y controles sean lo más próximos posible. Comenzar dividiendo en bloques de igual rango.
iv. Realizar un test de diferencias de medias entre los grupos de control y tratamiento dentro de un mismo bloque, estos no deben presentar diferencias significativas sobre las características de cada observación. Si las características de los individuos están balanceadas detener el proceso. Si las características no están balanceadas en algunos bloques, dividirlas en bloques de menor rango y hacer nuevamente las pruebas correspondientes. Si las características no están balanceadas para ningún bloque es necesario rediseñar el modelo *logit* o *probit*.

Adicionalmente, Ichino *et al.* (2006) propusieron un análisis de sensibilidad para el *propensity score matching* para evaluar la robustez de los estimadores del efecto medio de tratamiento ante fallas del supuesto de independencia condicional (*Conditional Independence Assumption, CIA*[20]). El análisis supone que la CIA no se mantiene para el conjunto de variables observables X, pero sí se mantiene dado X y una variable binaria no observable U. Los autores sugieren especificar cuatro parámetros para caracterizar la distribución de U. Luego, asignar un valor de U a cada individuo, de acuerdo a su estado de tratamiento y resultado, así se incluye a U en el conjunto de variables usadas para estimar el *propensity score* y calcular el efecto del tratamiento, repitiendo el procedimiento muchas veces (ej. 1.000) para obtener el efecto del tratamiento a través de la distribución de U. Si los resultados son relativamente insensibles a través de un rango plausible de U la inferencia causal es más defendible.

19. Utilizar un mínimo de variables posibles para explicar gran parte de la variabilidad de los datos del modelo.

20. También llamado selección en observables.

5.4.3 ¿Cuándo utilizar?

En la práctica, el *matching* se emplea típicamente cuando no es posible utilizar los métodos de selección aleatoria, el diseño de regresión discontinua, ni diferencias en diferencias. Esto se explica porque **el método es restrictivo al asumir que las características en las cuales se basa el *matching* es lo único que determina la participación en el programa.**

Además, es necesario justificar con la teoría o intuición que las características no observables no afectan el resultado ni la probabilidad de participación. Por esto, el *matching* es apropiado cuando el evaluador tiene una clara descripción del proceso de selección y una base de datos con muchas características que afectan la participación.

Algunos estudios utilizan el *matching* cuando no se dispone de datos de línea base, pero en las encuestas realizadas posterior a la aplicación del programa es posible extraer características observables a partir de las cuales se pueden deducir las características de las unidades en la línea base (por ejemplo: edad, género, educación, u otras). Obviamente, este análisis compromete seriamente la validez de los resultados, ya que muchas de las características observables podrían ser afectadas por el tratamiento, lo cual invalidaría los resultados.

Es importante notar que el supuesto de independencia condicional, sobre el cual se basan los métodos de *matching* requiere usar características observables de la línea de base, por lo cual las encuestas realizadas a las unidades tratadas deben ser muy similares a las encuestas realizadas a las unidades no tratadas. Además, estas encuestas deben poseer una gran cantidad de variables caracterizadoras, ya que se asume que no hay ninguna variable no observable que difiera sistemáticamente entre el grupo de tratamiento y grupo de control.

A pesar de que exista una base de datos bastante rica en términos de características, es difícil escoger las variables caracterizadoras. En este contexto, Heckman & Navarro-Lozano (2004) estudiaron cuán importante y al mismo tiempo, lo difícil que es escoger el conjunto apropiado de variables para el *matching*. En particular, los autores destacan que variables que generen una bondad de ajuste mayor podrían generar más sesgos que otras variables con menor bondad de ajuste, por lo cual sugieren que es muy relevante la correcta distinción entre variables exógenas y endógenas[21].

Finalmente, esta técnica no debería utilizarse si existen pocas características observables por las cuales condicionar al grupo de tratamiento y grupo de control, ya que como sólo se controla por las características observadas podrían existir muchas características no observables afectando el resultado.

21. Las variables endógenas se explican dentro de un modelo económico a partir de sus relaciones con otras variables (que a su vez pueden ser endógenas o exógenas). Las variables exógenas están determinadas fuera del modelo, es decir, están predeterminadas, el modelo las toma como fijas y mantienen siempre el mismo valor

5.4.4 Ventajas/ Desventajas

Ventajas	Desventajas
• Las técnicas de *matching* pueden aplicarse en muchas situaciones independientemente de las reglas de asignación del programa. • Se puede aplicar con posterioridad a la implementación del programa cuando se dispone de información tanto de una muestra de unidades tratadas por el programa como de otra muestra de unidades que no lo han sido. • Las técnicas de *matching* pueden ser usadas con datos de corte transversal o datos longitudinales (panel).	• El *matching* solo se basa en las características observadas, por lo cual se asume que las características no observables no afectan el resultado (supuesto muy fuerte). • Requiere grandes muestras para evitar la falta de soporte común entre los tratados y no tratados. • Es muy demandante de datos ya que requiere condicionar por una gran cantidad de características observables. • Si la distribución del PSM es muy diferente entre el grupo de tratamiento y el grupo de control existirán serias dudas sobre la credibilidad de la evaluación. • El *matching* es generalmente menos robusto que los otros métodos de evaluación porque sus resultados son más sensibles al número de observaciones y características observables por las cuales se condiciona.

5.4.5 Ejemplo(s) de aplicación

Andam *et al.* (2008) analizan la efectividad del sistema de áreas protegidas para reducir la deforestación tropical. El estudio se restringe sólo a Costa Rica, porque es el país que tiene uno de los sistemas de protección más desarrollados a nivel mundial a través de pagos por servicios ambientales, específicamente para evitar la deforestación.

El estudio trató de controlar por las diferencias en las imágenes satelitales de las áreas protegidas y no protegidas, basándose en la historia del programa y en la literatura sobre deforestación tropical. Las variables de *matching* fueron indicadores de uso productivo de la tierra, distancia a fronteras, caminos y ciudades, también se controló por otros factores causales menos claros como distancia a ríos, líneas férreas, densidad poblacional, inmigrantes, educación, pobreza y tamaño del distrito administrativo.

El estudio demuestra que la medición de la efectividad puede mejorarse sustancialmente al controlar por factores observables (efectos *spillovers* sobre áreas vecinas) y realizando análisis de sensibilidad. Los resultados arrojan que un 10% de las áreas protegidas en Costa Rica hubiesen estado deforestadas sin el sistema de protección.

Alix-Garcia & Sims (2010) investigan la efectividad y los efectos secundarios de un programa federal mexicano que compensa a los terratenientes por protección ambiental. Para evaluar la deforestación evitada gracias al programa, se debe comparar cómo los beneficiarios se hubiesen comportado si no hubieran recibido los pagos. Se utilizó un *matching* para región, tipo de propiedad, área enrolada, pendiente del terreno, elevación, tipo de base forestal, tasa de deforestación previa, densidad poblacional, grado de marginalidad y acceso a mercados. Las estimaciones muestran una reducción de 33% a 37% en la probabilidad de deforestación. Además, el programa tiene efectos heterogéneos, ya que parece ser más efectivo para evitar la deforestación donde la pobreza es baja y en los estados del sureste y noreste de México.

5.5 VARIABLES INSTRUMENTALES

5.5.1 Descripción

En algunos casos, las variables dependientes y/o explicativas se correlacionan con el término de error, ya sea porque existe una relación inversa entre la variable resultado y las explicativas ($Y \rightarrow X$), o bien cuando se han omitido variables explicativas importantes, como por ejemplo, no observables (motivación, análisis costo-beneficio que realiza cada unidad para determinar su participación, conciencia ambiental u otras). En esta situación, estimar el efecto de un programa estaría sesgado y los resultados posiblemente erróneos.

El método de variables instrumentales **busca corregir el problema al incorporar información de una nueva variable exógena que no debe estar correlacionada con las características no observables pero sí con la participación** (por ejemplo promoción para participar en el programa).

La técnica de variables instrumentales permite obtener estimaciones consistentes a partir de una regresión, aún en presencia de variables explicativas que estén correlacionadas con factores no observables incluidos en el término de error de una regresión.

Uno de los artículos más famosos por el uso de variables instrumentales utiliza el trimestre de nacimiento como un instrumento de la escolaridad para estimar el retorno de la educación, es decir, el incremento porcentual en el salario asociado a un año más de escolaridad (Angrist & Krueger, 1991). El atractivo de este estudio es precisamente encontrar una variable instrumental plausible, ya que el trimestre de nacimiento no debería afectar las remuneraciones, salvo a través de su efecto sobre los años de escolaridad obligatorios debido a que se puede cumplir antes la mayoría de edad establecida por las diferentes leyes estatales en Estados Unidos, y además, el trimestre no debería estar relacionado con otros factores como inteligencia, motivación, habilidades, contexto familiar, etc.

5.5.2 Formulación Estadística

El estimador de variables instrumentales puede ser estimado de forma simple con el método de mínimos cuadrados en dos etapas (*two-stage least squares*, 2SLS). En la primera etapa, se realiza una regresión lineal de la variable de participación en el programa P con respecto a la variable instrumental Z y las características observables X. En la segunda etapa, se realiza una regresión lineal de la variable de resultado Y con respecto a los valores de P predichos en la primera etapa junto con otras características observables X. Sin embargo, los errores estándar obtenidos con este método no son eficientes, por lo cual se sugiere estimar mediante rutinas de máxima verosimilitud pre-programadas en algún software estadístico que estiman ambas ecuaciones de forma conjunta.

$$P_i = \lambda + \gamma \cdot Z_i + \delta \cdot X_i + u_i$$

$$Y_i = \alpha + \beta \cdot X_i + \tau_{VI} \cdot \hat{P}_i + \varepsilon_i$$

Mientras más fuerte sea la relación entre la variable instrumental y la participación en el programa será mejor el instrumento (instrumento fuerte). Además, se requiere que la variable instrumental no esté correlacionada con otras variables que afecten el resultado

(condición de exogenidad). Estos requisitos aseguran que el instrumento es capaz de replicar las condiciones de asignación aleatoria. En la práctica, la fuerza de la relación entre la variable instrumental y la participación se puede testear con una regresión lineal. Sin embargo, el supuesto de exogenidad no es testeable.

Sólo si el efecto del tratamiento es igual para todas las unidades, el método permitirá identificar el efecto promedio del tratamiento (τ_{VI}) para toda la población. No obstante, en la práctica lo normal es que el efecto varíe entre las unidades (efecto heterogéneo). Bajo este contexto el método de variables instrumentales sólo permite estimar el impacto sobre un subgrupo de unidades cuya conducta es afectada por el tratamiento, este es el llamado efecto local promedio del tratamiento (LATE), ya que los impactos de las unidades cuya participación no se ve influenciada por la variable instrumental no están siendo considerados.

El concepto del LATE requiere supuestos más fuertes que el método tradicional de variables instrumentales ya que son utilizados para permitir heterogeneidad del efecto del tratamiento (Imbens & Angrist, 1994). Aunque el LATE es analíticamente muy similar al estimador de variables instrumentales, en términos conceptuales es diferente. Por ejemplo, cuando una variable dicotómica que representa un cambio de políticas es utilizada como instrumento, el LATE medirá sólo el efecto del tratamiento en el subconjunto de la población que cambia su estado de participación en respuesta al cambio en el instrumento.

Otro concepto llamado efecto marginal del tratamiento (MTE) fue introducido por Heckman & Vytlacil (1999),(2001),(2007) y por Carneiro *et al*. (2010) quienes reinterpretan el método de variables instrumentales y LATE bajo el contexto de un modelo de selección. El MTE utiliza una variable instrumental continua para recobrar completamente la distribución de probabilidad de la participación mientras todas las unidades tengan probabilidades positivas de ser tratados o no tratados ocasionadas por cambios en la variable instrumental Z. El atractivo de la técnica es que si los datos son suficientemente ricos los estimadores ATE, ATT y LATE pueden ser expresados a partir de MTE usando distintas ponderaciones.

5.5.3 ¿Cuándo utilizar?

Una variable instrumental es válida sólo cuando la participación en el programa no es determinada por los resultados potenciales, es decir, no está correlacionada con la variable dependiente sobre la que se evalúa el impacto del programa. Como esta condición es muy difícil de satisfacer el método es poco utilizado.

Por lo anterior, se requiere justificar la utilización de las variables instrumentales seleccionadas, enfatizando el cumplimiento de los supuestos necesarios para la identificación.

Existen algunos tests estadísticos para sustentar el uso de esta técnica, ya que se requiere demostrar que las variables instrumentales utilizadas no son débiles, es decir, que tienen suficiente capacidad para explicar el comportamiento de la variable endógena. En el caso de múltiples instrumentos se puede realizar un test de sobre - identificación o test de Hausman, mientras que para testear instrumentos débiles se puede utilizar el procedimiento descrito por Stock y Yogo (2005).

La técnica no debería ser utilizada si no es posible encontrar un instrumento que esté correlacionado con la participación en el programa y no correlacionado con el resultado. Tampoco puede utilizarse si la correlación con la participación es pequeña ya que la estimación del efecto del tratamiento podría estar seriamente desviada del efecto verdadero.

Por ejemplo, en los casos de programas con inscripción abierta o universal es posible realizar una promoción aleatoria del programa a algunas unidades (la promoción puede ser una campaña informativa o incentivos para la inscripción); en este caso no se necesita excluir a ninguna unidad elegible. En la medida que existan unidades inscritas si se promociona habrá una variación entre el grupo con promoción y el grupo sin promoción que permite identificar el impacto del programa sobre los inscritos si se promociona (se crea una variable instrumental "promoción del programa"), ya que la promoción aleatoria genera el equivalente a un grupo de control.

5.5.4 Ventajas/ Desventajas

Ventajas	Desventajas
• Si se dispone de un instrumento adecuado la técnica de variables instrumentales es una herramienta muy útil y creíble. • Si se dispone de una variable instrumental válida el método es de fácil estimación. • El estimador es consistente (propiedad estadística) cuando el tamaño tiende a infinito.	• No es fácil encontrar una variable instrumental creíble. • Una variable instrumental útil para un programa, no es necesariamente un buen instrumento en otros contextos. • Los resultados son sesgados en muestras pequeñas. • Si sólo existen pequeñas correlaciones entre la variable instrumental y la participación en el programa (instrumentos débiles), las estimaciones son incluso peores que las realizadas por mínimos cuadrados ordinarios. • Los resultados del impacto son válidos sólo para un subgrupo particular de la población. • Si las unidades están conscientes de los beneficios del tratamiento sobre ellas, entonces estas unidades tienen una decisión de participación mucho más informada, lo cual genera una correlación entre el resultado y la variable instrumental invalidando el estimador.

5.5.5 Ejemplo(s) de aplicación

Frankel & Rose (2005) analizan cuál es el efecto de mayor comercio exterior de un país sobre el ambiente. A diferencia de estudios previos consideran simultáneamente la endogeneidad entre el comercio exterior, PIB y medioambiente, justificando un instrumento para el comercio exterior y otro para el PIB. En particular, los autores construyen una variable instrumental para el comercio exterior a partir de un modelo que explica los patrones comerciales con variables geográficas (que son plausiblemente exógenas) y una variable instrumental para el PIB con valores rezagados del PIB, población y tasa

de inversión en capital humano. También, utilizan siete medidas de daño ambiental que incluyen emisiones de CO_2, SO_2, NO_2 y MP, deforestación, energía y acceso a agua en zonas rurales. Debido a que existen algunos datos no disponibles en las variables dependientes, la cantidad de observaciones en cada modelo estimado varía entre 35 y 40. La utilización de variables instrumentales no cambió los resultados obtenidos por mínimos cuadrados ordinarios, y además, sus resultados muestran que el comercio exterior tiene un efecto positivo sólo en tres variables ambientales, SO_2, NO_2 y MP. Así, se rechaza la idea que el comercio internacional tiene un efecto perjudicial sobre el medio ambiente y se rechaza la hipótesis del refugio de la contaminación, la cual afirma que el comercio fomentaría a algunos países a especializarse en actividades contaminantes.

Jeffords & Minkler (2014) buscan determinar si los derechos ambientales garantizados constitucionalmente son necesarios para obtener buenos indicadores ambientales. Para lo anterior, utilizan tres instrumentos, el primero relacionado con que las constituciones más nuevas probablemente deberían tener derechos ambientales, la segunda variable incluye el número de otros derechos sociales y económicos incorporados en la constitución, y el tercero es el mínimo entre la antigüedad de la constitución y de la provisión de CER. Los datos incluyen observaciones para 169 países, de los cuales 110 incluyen derechos ambientales en sus constituciones y 59 no los incluyen. Debido a que existen algunos datos no disponibles en las variables dependientes o independientes, la cantidad de observaciones en cada modelo estimado varía entre 109 y 147. Con la base de datos construida el estudio encuentra evidencia de que las constituciones de los países tienen efecto sobre los indicadores ambientales. Además, concluyen que los tres instrumentos utilizados para predecir la inclusión de derechos ambientales son válidos y no son débiles.

5.6 MODELOS ESTRUCTURALES

5.6.1 Descripción

Los modelos estructurales han sido introducidos para explicar a través de modelos de comportamiento económico fenómenos generados por programas o políticas públicas. Este método se basa en **descripciones matemáticas estilizadas (sencillas) que describen el comportamiento de agentes** (maximización de utilidad de los individuos, minimización de costos o maximización de beneficios de las firmas), estas ecuaciones se calibran con microdatos de hogares y/o firmas con el objetivo representar el equilibrio parcial en un mercado, con datos de la matriz insumo-producto para representar un equilibrio general (representación de todos los mercados), datos macroeconómicos para desarrollar modelos de equilibrio general dinámico estocásticos (DSGE), así como también, existen modelos estructurales muy simplificados para desarrollar modelos econométricos de forma reducida.

Una vez descrito el equilibrio se modifica algún parámetro de política para realizar predicciones sobre alguna variable de resultado. Así, los modelos estructurales típicamente son usados para realizar evaluaciones *ex ante*, pero a la vez generan escenarios contrafactuales que pueden ser usados para evaluación *ex post* cuando los resultados contrafactuales no pueden ser identificados de otra forma.

El desarrollo de modelos estructurales es un área muy amplia en economía, obviamente también ha abordado problemáticas ambientales para evaluaciones *ex ante*. Por ejemplo, en el caso de Chile existe una amplia literatura sobre modelos de optimización que

simulan regulaciones ambientales en el contexto de sistema de permisos transables, sistema de permisos ambientales, mientras que la simulación de impuestos ambientales es escasa.

5.6.2 Formulación Estadística

La formulación del modelo es muy dependiente del tipo de situación que se pretenda representar por lo cual en esta sección no se describirán formulaciones específicas. Sin embargo, a modo de ejemplo, en algunos casos se puede modelar la minimización de costos totales de las firmas sujeto a restricciones de regulación y disponibilidad de combustibles, en otros casos, la modelación de mercados específicos como el eléctrico, o bien ser modelos de equilibrio general en los cuales los consumidores maximización su utilidad y las firmas maximización de beneficios, generándose un conjunto de precios que equilibran los mercados y que se ajustan ante *shocks* ambientales. En general, estos modelos son determinísticos pero también existen modelos estocásticos que incluyen algunas variables aleatorias.

5.6.3 ¿Cuándo utilizar?

Aunque es poco probable que el evaluador disponga de un modelo estructural que describa la situación específica que se desea simular, en el caso que existiera un modelo desarrollado previamente se le podría solicitar a su autor que desarrolle las simulaciones contrafactuales requeridas en el contexto del estudio que se pretende abordar.

También, podría ser útil para estimar la ampliación desde un programa a pequeña escala hacia una hipotética implementación a gran escala.

La técnica no debería ser utilizada si las simulaciones del modelo tienen un bajo poder predictivo al contrastarlo con datos reales.

5.6.4 Ventajas/ Desventajas

Ventajas	Desventajas
• Es una alternativa cuando no existe la posibilidad de obtener un grupo de control o escenarios contrafactuales. • Si se dispone de un modelo ya construido (un modelo internacional, desarrollado por alguna investigación científica local o un modelo comercial) se requiere poco tiempo para generar escenarios contrafactuales. • Permite realizar múltiples simulaciones de escenarios contrafactuales.	• Desarrollar un modelo estructural puede ser muy demandante de tiempo y datos. • Típicamente las predicciones de los modelos teóricos no tienen un buen nivel de ajuste comparado con los datos reales, por lo cual pueden ser un pobre contrafactual.

5.6.5 Ejemplo(s) de aplicación

Pilavachi *et al.* (2008) comparan los datos predichos a través de modelos económicos y energéticos con respecto a datos reales luego de 15 años de sus proyecciones para establecer la relevancia de estos modelos en términos de políticas energéticas. Las predicciones analizadas corresponden a las efectuadas en el año 1985 con los modelos MEDEE y EFOM, para proyectar el consumo de energía al año 2000. Para comparar los resultados con los datos reales se requirió obtener información de consumo energético de EUROSTAT. El estudio concluye que los modelos utilizados son importantes para hacer predicciones energéticas. No obstante, las diferencias entre los datos reales y proyectados demuestran lo importante que son los supuestos sobre las decisiones políticas, incentivos económicos y comportamiento social, pero a la vez lo difícil que resulta predecirlos en los ejercicios de simulación.

Webber *et al.* (2015) aportan evidencia *ex post* con datos a gran escala sobre la efectividad de programas de *retrofit* o reacondicionamiento de hogares para mejorar la eficiencia energética. Los datos de consumo energético a nivel de hogar no están públicamente disponibles en el Reino Unido, sólo están disponibles a nivel local para ciertas áreas geográficas pequeñas que incluyen entre 200 y 6.000 viviendas, por lo cual los autores debieron combinar datos de uso de energía con datos anónimos a nivel de hogar ofrecidos por las autoridades del programa. El conjunto de datos finalmente utilizado incluyó características físicas de los hogares como número de habitaciones, tipo de construcción, antigüedad, localización de la propiedad, nivel de aislación antes y después de participar en el programa, fecha en la cual se realizó la aislación y características socioeconómicas. Los resultados sugieren que los impactos reales de un programa de reacondicionamiento de viviendas para eficiencia energética fueron más altos que lo predichos. En particular, los impactos en el uso de la energía en áreas de bajos ingresos es concordante con las predicciones, pero en áreas de ingresos medios y altos los impactos son mayores que los predichos.

5.7 FUNCIÓN DE CONTROL

5.7.1 Descripción

El método de la función de control **analiza conjuntamente el efecto del programa y la elección de una unidad cuando evalúa decidir su participación en el programa.** El individuo participará en el programa si los beneficios de participar superan los costos. Sin embargo, la decisión depende de variables observables y no observables. Por ello, se especifica una distribución de probabilidad conjunta de la regla de asignación y el tratamiento.

La función de control soluciona el problema de endogeneidad de la regla de asignación como un problema de variable omitida. Al asumir que toda la información relevante para la asignación depende de variables no observables. Específicamente, la función de control se construye al modelar la asignación del tratamiento a partir de variables instrumentales, luego esta función de control es incluida en una regresión para estimar el resultado.

5.7.2 Formulación Estadística

Este método está directamente relacionado con el tradicional estimador de sesgo de selección de Heckman (1979), ya que en las primeras aplicaciones de la función de control $f(\bullet)$ se utilizaba el supuesto de distribución conjunta normal de los términos de error en ambas ecuaciones (u y v) lo cual permitía incluir correlación de los errores ρ y, además, utilizaba una forma funcional tipo Probit para la regla de asignación, tal como se especifica a continuación.

$$Y_i = \alpha + \tau \cdot P_i + \delta \cdot X_i + u_i$$

$$P_i = f(\gamma \cdot Z_i + v_i)$$

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right]$$

El método inicialmente fue estimado usando el procedimiento Heckit en dos etapas. En la primera etapa se calculaban los valores predichos de la función de control a través de una regresión de P sobre Z. En la segunda etapa se estimaba la variable de resultado con una regresión de mínimos cuadros ordinarios incorporando como variable explicativa los valores predichos de la primera etapa en la función de distribución $f(\bullet)$ y distribución acumulada $F(\bullet)$.

$$Y_i = \alpha + \beta \cdot X_i + \tau_{FC} \cdot P_i + \delta \cdot \left[P_i \cdot \frac{f(\hat{\gamma} \cdot Z_i)}{F(\hat{\gamma} \cdot Z_i)} + (1 - P_i) \cdot \frac{f(\hat{\gamma} \cdot Z_i)}{1 - F(\hat{\gamma} \cdot Z_i)} \right] + u_i$$

La estructura que impone el proceso de selección sobre las características no observables permite extrapolar los resultados a escenarios de políticas alternativas. Sin embargo, esta misma estructura ha sido fuertemente criticada por ser muy restrictiva. Para superar estas críticas se han propuesto estimadores semi - paramétricos o no paramétricos.

5.7.3 ¿Cuándo utilizar?

Cuando se desea modelar el proceso de selección o asignación del tratamiento basado en factores no observables. Sin embargo, sus resultados tienden a ser bastante sensibles a la estrategia de identificación y a los supuestos de la distribución de los errores.

5.7.4 Ventajas/ Desventajas

Ventajas	Desventajas
• Se puede entender el proceso de selección y confirmar expectativas previas. • El modelo es muy similar al método de variables instrumentales, pero sondea información sobre factores no observables. • El método de función de control es probablemente más eficiente y preciso que utilizar variables instrumentales, aunque menos robusto.	• Ha sido criticada por asumir una estructura para la regla de selección muy restrictiva. • Los supuestos paramétricos pueden ser inapropiados para delinear el efecto de factores no observables, aunque en especificaciones más complejas también se podrían utilizar técnicas no paramétricas.

5.7.5 Ejemplo(s) de aplicación

Brouhle & Ramirez (2010) analizan los factores que influyeron en la participación de las empresas en un programa de registro y cambio voluntario (VCR) en Canadá y determinan cómo los gases de efecto invernadero se redujeron por los niveles de participación en el VCR desde 1995 a 2003. Se utilizó una base de empresas obligadas a reportar los gases de efecto invernadero durante 2004. Luego de filtrar los datos por observaciones perdidas la muestra se redujo a 150 firmas. También, se utilizaron datos sobre la participación voluntaria en los registros de VCR. Los autores asumen que el proceso de participación dinámica en el programa se explica por la participación pasada y otras variables rezagadas. En particular, afirman que en 1998 las firmas que decidieron reportar o no reportar, también escogieron a qué nivel de calidad de aire comprometerse. El estudio concluye que las empresas participaron en el VCR para señalizar su responsabilidad ambiental a los reguladores e inversionistas, pero no a los consumidores. Sin embargo, de acuerdo al registro obligatorio de gases de efecto invernadero en 2004 no se observaron diferencias significativas entre las empresas que participaron en el VCR y las que no participaron.

5.8 TÓPICOS ADICIONALES

A continuación se describen algunos tópicos adicionales relacionados con algunas de las técnicas ya mencionadas. En particular, se analiza el caso en el cual existen diferentes niveles para un tratamiento y múltiples tratamientos.

5.8.1 Métodos combinados

Cuando se dispone de datos de línea de base y datos *ex post*, el *matching* puede combinarse con el método de diferencias en diferencias, lo cual permite condicionar por la heterogeneidad no observable que permanece constante a través del tiempo. Para el caso de dos periodos y dos grupos el estimador de *matching* de diferencias en diferencias consiste en aplicar el *propensity score matching* usando el cambio en el resultado, $\Delta Y_{it} = Y_{it} - Y_{it-1}$, en vez del resultado en niveles (Heckman, Ichimura, & Todd, 1997). Con varios periodos pretratamiento, los valores rezagados de la variable de resultado pueden ser incluidos en la estimación del *propensity score matching* para garantizar que las unidades siguen la misma tendencia antes del programa. Sus ventajas y desventajas

son las siguientes:

Ventajas	Desventajas
• El estimador del impacto es efectivo para eliminar el sesgo de selección cuando existen factores no observables invariantes en el tiempo. • Permite testear los supuestos realizados en un *matching* tradicional.	• Al igual que otros métodos asume que las variables no observables para los grupos de tratamiento y control son invariantes con el tiempo, lo cual a menudo no se cumple.

5.8.2 Diferentes niveles de tratamiento

La mayor parte de la literatura de evaluación de impacto se ha enfocado en el caso de un tratamiento binario (tratado o no tratado), por lo cual **se ha puesto poca atención en el caso de tratamientos discretos, continuos (con distintas intensidades del tratamiento) o el efecto de participar múltiples veces.** Sin embargo, estos casos no son poco usuales en contextos sociales o ambientales, en donde los tratamientos rara vez son completamente homogéneos.

Por ejemplo, se puede dar el caso de aplicar o no un subsidio, cuyo valor no necesariamente es constante, este es el caso de programas de aislamiento térmico de viviendas que realiza el MINVU. En otro contexto, los esfuerzos de un programa de fiscalización no necesariamente son homogéneos para todas las fuentes emisoras de contaminantes, ya que el fiscalizador puede concentrar más sus esfuerzos en algunos tipos de fuentes. También podría ser de interés cómo programas de fiscalización repetidos afectan el comportamiento de las fuentes emisoras fiscalizadas.

Diseñar una evaluación de impacto para un programa con niveles variables de tratamiento en términos conceptuales es relativamente fácil. En un contexto experimental se puede realizar una asignación aleatoria para decidir a qué parte de la muestra se va a asignar el tratamiento P_1, el tratamiento P_2, y así, hasta el tratamiento P_K. Obviamente, a una parte de la muestra no se le asigna ningún tratamiento, en consecuencia existirán $K+1$ opciones de tratamiento.

Si la asignación aleatoria fue implementada correctamente, este diseño garantiza que los diferentes grupos de unidades son similares. Por lo cual, se puede estimar el impacto del tratamiento, mediante la comparación del resultado promedio del grupo k con el resultado promedio del grupo sin tratamiento. También, se pueden estimar los diferentes niveles de tratamiento entre sí. Por ejemplo, evaluar si el efecto del tratamiento tipo k tiene mayor impacto que el tipo j.

Cuando la variable de tratamiento toma múltiples valores, discretos o continuos, las técnicas de regresión lineal, datos de panel o variables instrumentales son perfectamente válidas. Sólo se debe tener precaución en la interpretación de los resultados, con un tratamiento binario el coeficiente asociado al tratamiento en un contexto de regresión estima la diferencia entre unidades tratadas y no tratadas, en el caso de tratamiento con niveles múltiples de tratamiento la introducción de variables *dummy* permite capturar

efectos no lineales, mientras en el caso de un tratamiento continuo se mide el impacto de un incremento de una unidad en el tratamiento respecto a la variable de resultado (por ejemplo un año más de tratamiento, mil pesos extra de subsidio, etc.)

En el caso de las técnicas de *matching* la situación es algo distinta. Bajo un tratamiento en el cual existen múltiples niveles discretos asumiendo independencia condicional en la asignación del tratamiento, la extensión metodológica del *propensity score matching* para poder realizar estimaciones de impacto es directa desde el caso del tratamiento binario. Por ejemplo, supongamos que existen tres niveles de tratamiento para cada unidad i, estos son $P_i = 0$, $P_i = 1$ ó $P_i = 2$. Para estimar el impacto del tratamiento nivel 2 relativo al nivel 1, simplemente se pueden dejar de lado las unidades expuestas al tratamiento nivel 0. Un problema práctico es que el supuesto de soporte común[22] probablemente será violado con más de dos tratamientos.

También, se puede dar el caso en el cual las unidades pueden ser expuestas a una secuencia de tratamientos binarios, por ejemplo el programa se puede llevar a cabo en tres periodos, así en cada periodo se puede asumir independencia condicional en la asignación del tratamiento, dadas ciertas características observables de la unidad que no varían a través del tiempo.

En el caso de un tratamiento continuo bajo el supuesto de independencia condicional se requieren modificaciones más importantes relacionadas con la literatura de la estimación de ecuaciones simultáneas. El supuesto clave utilizado es que si se ajusta por diferencias previas al tratamiento es posible eliminar cualquier sesgo. Imbens (2000) introdujo el *propensity score* generalizado para el caso de múltiples tratamientos, el cual es la probabilidad condicional de recibir un tipo particular de tratamiento dado las variables observables pre – tratamiento, es decir, $\text{Prob}\,(P_i = p_i | X_i = x_i)$. Hirano & Imbens (2004) para utilizar la metodología se basan en estimar el *propensity score* generalizado usando una distribución lognormal.

5.8.3 Múltiples tratamientos

Cuando existen múltiples programas disponibles el evaluador puede estar interesado en estimar sólo los efectos individuales de cada tratamiento o bien la interacción entre ellos.

Diseñar una evaluación de impacto para un programa con múltiples tratamientos es un poco más complejo que el caso con diferentes niveles de tratamiento. La principal diferencia es la necesidad de generar varias asignaciones aleatorias independientes, lo cual produce un diseño cruzado. En el simple caso de dos tipos de tratamiento, una vez escogida la muestra de unidades elegibles dentro de la población, se asigna aleatoriamente a las unidades que formarán parte del grupo de tratamiento y el grupo de control. Luego, se hace una segunda selección aleatoria dentro del grupo de tratamiento para escoger a las unidades a las cuales se les aplicará conjuntamente el segundo tratamiento. Finalmente, se realiza otra selección aleatoria dentro del grupo de control para escoger las unidades no tratadas a las cuales se les aplicará el segundo tratamiento, mientras que las unidades restantes serán el grupo de control "puro". En consecuencia, se habrán generado cuatro grupos. Si la asignación aleatoria fue implementada correctamente, este diseño garantiza que los cuatro grupos de unidades son similares. Así, es posible estimar el impacto del primer tratamiento con respecto al grupo de control

22. El soporte común asegura la existencia de unidades tratadas que se "parecen" a las unidades no tratadas, es decir, las unidades que poseen el mismo valor del PSM tienen una probabilidad positiva de ser participantes y no participantes.

puro, así como también, el impacto del segundo tratamiento con respecto al grupo de control puro. También, es posible estimar el impacto de recibir el segundo tratamiento cuando ya se recibió el primer tratamiento.

Una forma sencilla de estimar los resultados en este contexto es con métodos de regresión lineal o datos de panel, para lo cual es necesario agregar variables *dummy* para cada tratamiento, incluyendo variables *dummy* para los efectos de interacción entre ellos. Por ejemplo, en el caso de dos programas, P_1 y P_2, la especificación bajo una regresión lineal es la siguiente:

$$Y_i = \alpha + \beta \cdot X_i + \tau_1 \cdot P_1 + \tau_2 \cdot P_2 + \tau_3 \cdot P_3 + \varepsilon_i$$

Donde, los coeficientes τ_1 y τ_2 capturan los efectos individuales de cada tratamiento, y el coeficiente τ_3 captura el efecto de la interacción entre los dos programas. Si τ_3 no es estadísticamente diferente de cero, el efecto simultáneo de ambos programas es simplemente la suma de los efectos individuales. En otro caso el efecto combinado podría reforzar o disminuir el efecto de cada programa.

Es obvio en este contexto la dificultad de utilizar otras técnicas, por ejemplo encontrar variables instrumentales válidas para las variables *dummy* del tratamiento.

Además, es importante notar que la evaluación de más de una intervención generará dificultades de diseño, ya que la complejidad se incrementará exponencialmente con el número de los distintos tipos de tratamientos. En este caso para poder distinguir los resultados entre los grupos, se requiere una gran cantidad de observaciones para la detección de diferencias estadísticamente significativas de las combinaciones de las diferentes intervenciones. Sin mencionar las dificultades prácticas para el funcionamiento del programa, y la necesidad de controlar la interacción o contaminación entre las unidades.

5.9 ¿QUÉ MÉTODO CUANTITATIVO UTILIZAR?

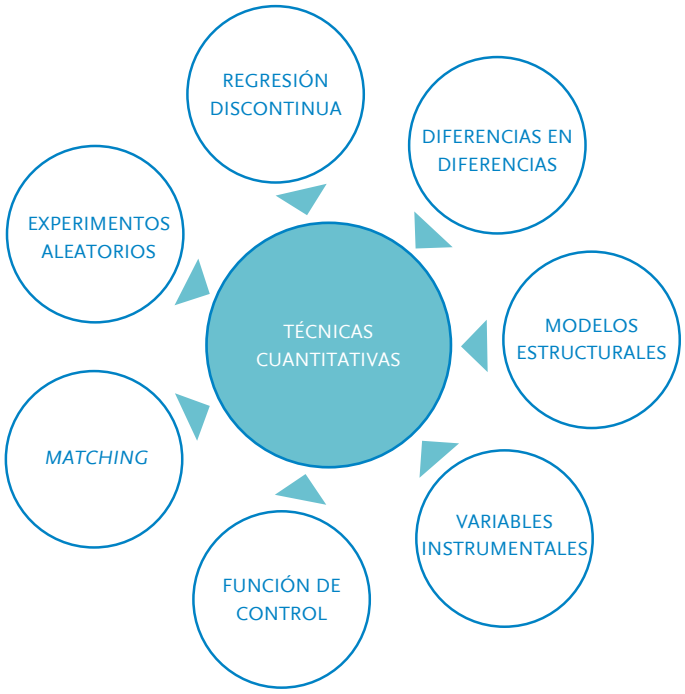
Según Blundell & Costa-Dias (2009) uno de los elementos clave para escoger el diseño de evaluación más apropiado es **comprender claramente la regla de asignación de la política o programa**. En particular, deben existir argumentos convincentes para demostrar que la regla de asignación se basa en variables observables, o bien, que existen variables observables y no observables que determinan la asignación. Esto porque algunos diseños asumen sólo la selección sobre variables observables, mientras que otros son capaces de condicionar el impacto cuando existe heterogeneidad no observable. La selección en observables se puede interpretar simplemente como que lo único que explica el resultado son las variables o características conocidas y para las cuales se dispone de información de las unidades (por ejemplo: género, edad, tamaño familiar, entre otras). Al contrario, la heterogeneidad no observable se refiere a variables o características que influyen en el resultado para las cuales no hay datos (por ejemplo: nivel de conciencia ambiental, aversión al riesgo, habilidad, entre otras). Pero incluso si los argumentos son convincentes respecto a la regla de asignación, algún diseño particular sólo permitirá responder un conjunto limitado de preguntas sobre el programa o política.

En general, la visión común entre los evaluadores es que no existe un único diseño que, independientemente de las circunstancias, debería aplicarse en todas las evaluaciones

de impacto (Rossi, Lipsey & Freeman, 2004). Obviamente, la disponibilidad de datos, tiempo, recursos, características del programa y otros será determinante para escoger entre los diversos diseños.

En la Figura 5-9 se muestran las diversas técnicas cuantitativas que pueden ser utilizadas en el contexto de una evaluación *ex post*.

Figura 5-9 técnicas cuantitativas



Fuente: Elaboración propia

6 | METODOLOGÍAS CUALITATIVAS DE EVALUACIÓN

Las metodologías cualitativas se aproximan a la evaluación de políticas públicas de forma distinta a la perspectiva cuantitativa habitual. Su objetivo es abordar preguntas que los métodos cuantitativos no son capaces de responder. En particular, permiten explorar supuestos que se asumen como evidentes, obtener diferentes visiones sobre una misma política o programa y/o entender cómo se produce el proceso de cambio una vez implementada la política o programa. Además, ayudan a complementar la interpretación de los resultados cuantitativos, formular hipótesis, cadenas de resultados, diseñar cuestionarios aplicados en evaluaciones cuantitativas, y examinar casos particulares de éxito o fracaso del programa.

En general bajo un contexto de evaluación de impacto *ex post*, la evaluación cualitativa de impacto puede contribuir a comprender lo que está ocurriendo con el programa, al dar algunas explicaciones sobre el por qué se observan ciertos resultados cuantitativos y abrir la "caja negra" del impacto generado por el programa (Bamberger, Rao & Woolcock, 2010).

La evaluación mediante métodos cualitativos se caracteriza por ser flexible, cambiante y circular, ya que no parte de hipótesis previas, lo cual asume diversas perspectivas de una misma realidad (Vásquez *et al.*, 2006).

El concepto de metodología cualitativa incluye diferentes técnicas para generar y analizar datos no numéricos que se caracteriza por los siguientes elementos:

- Se reconoce que no existe una realidad única y objetiva.
- Utiliza un proceso inductivo desde casos particulares a una teoría general.
- La recopilación de información combina diferentes técnicas, siendo intensiva más que extensiva.
- La muestra no es estadísticamente representativa.
- La investigación se desarrolla dentro del contexto natural del fenómeno social.
- Se reconoce que la visión del evaluador puede influir en el proceso de investigación.
- El diseño de la investigación tiene flexibilidad para modificarse en la medida que surgen nuevos datos.

La(s) técnica(s) seleccionada(s) finalmente dependerá(n) de la naturaleza del programa o política que se desea evaluar.

Algunas preguntas que podrían motivar la aplicación de una evaluación cualitativa incluyen las siguientes:

- ¿Es una evaluación exploratoria?
- ¿Se requiere precisar o redefinir los objetivos del programa?
- ¿Se requiere entender la cadena de resultados del programa?
- ¿Existe interés por conocer aspectos débiles y fuertes del proceso de implementación del programa?
- ¿No es factible obtener información cuantitativa a través de registros administrativos o encuestas?
- ¿Es necesario identificar y analizar casos exitosos o casos en los cuales el programa no obtuvo los resultados esperados?
- ¿Se requiere analizar la diversidad de algún aspecto especial del programa?
- ¿Se requiere información detallada de la implementación del programa?
- ¿Se requiere explicar cómo se generan ciertos resultados del programa?
- ¿Es necesario complementar la información cuantitativa?
- ¿Es posible que el programa afecte de forma imprevista al grupo de tratamiento?
- ¿Se requiere conocer externalidades a partir de la percepción o experiencia del programa?
- ¿Se requiere información inmediata a través de un análisis selectivo de casos o entrevistas?
- ¿Es posible interpretar en profundidad los resultados estadísticos?
- ¿El programa no cuenta con un sistema para el levantamiento de información cuantitativa?
- ¿No es posible aplicar encuestas o identificar grupos de control?
- ¿Se requiere enriquecer los indicadores a través de la profundidad y detalle que aporta el estudio de casos concretos?

Las técnicas cualitativas permiten obtener información intensiva sobre casos no representativos pero escogidos de manera estratégica, lo cual ayuda a comprender cómo funciona una política pero no pretende generalizar las conclusiones a toda la población.

No existe una única manera para definir las técnicas cualitativas a utilizar. En la práctica, cada política o programa puede ser mejor evaluada con técnicas específicas o bien puede requerir el uso de una combinación de diferentes técnicas cualitativas, algunas pueden generar una aproximación inicial mientras otras pueden profundizar los resultados de la política o programa.

Una de las características de un análisis cualitativo es que los datos recogidos principalmente corresponden a "textos", por ello se requiere transcribir literalmente las grabaciones de las entrevistas o comentarios específicos, los cuales deben ser clasificados en categorías, conceptos o tipologías de temas, es decir, se segmentan los datos textuales en códigos que agrupan fragmentos de texto asociados a una misma clasificación temática, reduciendo los datos de forma manual o mediante *software* de temas recurrentes o excepcionales que pueden ser vinculados a las preguntas iniciales de la política o programa. Esto ayuda a detectar patrones, tendencias, diferencias, realizar comparaciones o describir casos únicos que faciliten el desarrollo de propuestas teóricas que muestren la percepción de los grupos tratados y/o gestores del programa, ayuden a establecer relaciones, o bien, a explicar e interpretar los datos. Finalmente, de forma inductiva el evaluador construye su explicación sobre cómo perciben los informantes una política o programa.

De acuerdo a lo descrito previamente, es muy importante que las percepciones de los tratados o gestores del programa se distingan claramente de las interpretaciones que realizan los investigadores en la exposición de los resultados. Además, el análisis cualitativo se refuerza si se utilizan métodos de triangulación de investigadores y técnicas, es decir, distintos investigadores observando los mismos datos y utilizando distintas técnicas. Sin embargo, una garantía de calidad metodológica en la investigación cualitativa se consigue al verificar que los resultados obtenidos se acercan a la realidad del objeto bajo estudio.

Desde que los estudios cualitativos fueron introducidos, se transformaron en un complemento para el conjunto de técnicas cuantitativas experimentales y estadísticas. Aunque el debate sobre su legitimidad sigue abierto, hay cada vez más investigadores que prefieren diseñar evaluaciones que mezclan métodos cualitativos y cuantitativos ya que ambas aproximaciones metodológicas tienen puntos fuertes y débiles que se complementan entre sí (Tashakkori & Teddlie, 2010).

A continuación se describen las diversas técnicas cualitativas que pueden ser utilizadas en el contexto de una evaluación *ex post*.

6.1 ANÁLISIS DOCUMENTAL

6.1.1 Descripción

El análisis documental realiza un análisis de los registros formales de una política o programa sin interrumpir su funcionamiento. Estos registros pueden ser documentos y discursos oficiales, informes, leyes, normativas, estadísticas, materiales audiovisuales, reportajes, noticias, diarios, radios, internet o revistas, entre otros. Este análisis se caracteriza por una clasificación sistemática, descripción e interpretación de los contenidos narrativos de los registros, de acuerdo a los objetivos del estudio. Sin embargo, como los registros son heterogéneos (textos, audiovisuales, informes públicos o privados, etc.) no es sencillo establecer reglas generales para su análisis.

Figura 6-1. Ejemplos de registros en un análisis documental

Fuente: Elaboración propia

El levantamiento de los registros existentes permite delinear los aspectos centrales de una determinada política o programa, así como sus objetivos, presupuesto, acciones, responsabilidades, entre otros.

6.1.2 ¿Cuándo utilizar?

Cuando se requieren conocer aspectos históricos, contextuales, normativos, organizacionales, institucionales, opinión pública, entre otros, relacionados con la política o programa que se está evaluando.

Además, la información levantada con el análisis documental puede ser combinada con otras técnicas para enriquecer los resultados obtenidos de la evaluación cualitativa.

6.1.3 Ventajas/ Desventajas

Ventajas	Desventajas
• Es una técnica no intrusiva[23] que no interrumpe el funcionamiento de la política o programa evaluado. • Los datos registrados son permanentes en el tiempo y pueden obtenerse de forma longitudinal. • La información puede ser más confiable que la levantada mediante entrevistas o encuestas. • Algunos registros pueden obtenerse fácilmente y a bajo costo.	• El proceso de levantamiento de registros es lento. • El éxito depende de la obtención de la documentación que está en manos del ente ejecutor. • Los registros pueden estar incompletos, ser limitados o parciales. • Algunos registros pueden tener otros propósitos.

6.1.4 Ejemplo(s) de aplicación

Abdul-Manan *et al.* (2015) realizan un análisis histórico de la introducción y descripción de políticas energéticas para determinar la eficacia de las políticas que llevaron al actual mix energético de Malasia. Se revisaron documentos de políticas oficiales, textos legales, publicaciones científicas de diferentes campos, libros, información de sitios web de agencias gubernamentales y no gubernamentales, tanto locales como internacionales. Los autores concluyen que las políticas energéticas de los últimos años han llevado a una base energética más diversificada, que los precios de la energía no toman en cuenta los costos asociados con la degradación ambiental, y además, que los sistemas tecnológicos e institucionales han girado en torno a la industria petrolera haciendo extremadamente difícil y costosa una transición a un sistema energético más sustentable.

23. Ya que está diseñada para no interferir en el comportamiento observado, y así, evitar cambios indeseados en las actividades normales de las unidades bajo análisis.

6.2 OBSERVACIÓN DIRECTA

6.2.1 Descripción

La observación directa analiza sin manipulación la realidad social, lo cual permite que siga su curso de acción natural. La observación directa permite capturar lo que realmente hace la población respecto a lo que la población dice que hace en las entrevistas o encuestas. Entrega información sobre las rutinas de los actores sociales que incluso no son percibidas por ellos mismos. Esta observación y registro de comportamientos naturales permite entenderlos en profundidad. Además, permite contrastar las percepciones, comentarios o explicaciones de la población afectada por la política o programa con respecto a sus acciones reales.

La observación puede presentar diferentes grados de estructuración, nivel de conocimiento de la población observada (oculta[24] vs. abierta), y participación de los evaluadores en el proceso de observación.

24. En este caso las unidades no saben que están siendo observadas.

6.2.2 ¿Cuándo utilizar?

En general, se puede utilizar cuando se conoce poco de la política o programa bajo estudio. Se puede utilizar como técnica exploratoria, descriptiva y orientada a la interpretación teórica de la cadena de resultados. Además, permite a los evaluadores una mejor compresión del impacto de un programa y puede llegar a ser la única alternativa para observar comportamientos reales en situaciones ilegales (por ejemplo: botar escombros de forma ilegal o detección de patrones de comercialización de leña en el comercio informal).

6.2.3 Ventajas/ Desventajas

Ventajas	Desventajas
• Información sobre el funcionamiento real de una política o programa. • Es una técnica que permite una rápida recolección de datos.	• Puede ser difícil la interpretación de algunos comportamientos observados sin utilizar otras técnicas complementarias. • Puede ser difícil la categorización de las observaciones. • El conocimiento de la observación por parte de la población puede alterar su comportamiento. • El análisis de los resultados puede requerir de bastante tiempo.

6.2.4 Ejemplo(s) de aplicación

Aunque no se pudieron encontrar estudios medioambientales con esta técnica, a modo de ejemplo, se puede mencionar el trabajo de Morgan *et al*. (2009) quienes desarrollan una medida fiable de la frecuencia y la duración de la exposición al agua en playas de surf por parte de los bañistas según sexo y grupo de edad. Los bañistas seleccionados fueron observados sistemáticamente entrar al agua durante las horas del día a seis playas con o sin vigilancia por 10 días. Las variables medidas fueron: el clima y las condiciones del agua, entradas de agua, duración de la exposición del agua, la ubicación exposición al agua y las características de la persona. Los resultados sugieren que la sobrerrepresentación de los varones adolescentes y adultos en las estadísticas de ahogamiento en las playas de surf es en parte un producto de la mayor exposición total al agua, a exposición más frecuente a aguas profundas y a bañarse más lejos de la costa.

6.3 ENTREVISTAS EN PROFUNDIDAD

6.3.1 Descripción

Las entrevistas en profundidad permiten comprender las percepciones, ideas o valores que tienen los entrevistados sobre una política o programa determinado. La información se recoge en forma sistemática de acuerdo a objetivos preestablecidos, a diferencia de una conversación informal.

Las entrevistas en profundidad son claves para evaluar políticas o programas que requieren un análisis detallado de las percepciones, valores, actitudes y opiniones de los actores involucrados.

El grado de estructuración de las entrevistas en profundidad puede variar desde un nivel bajo hasta un nivel muy estructurado. Dependiendo del grado de estructuración las entrevistas pueden ser clasificadas en: entrevistas informales (no hay predeterminación de preguntas ni temas), entrevistas abiertas (existe un guion de grandes temas a tratar), entrevistas semi - estructuradas (las preguntas están ordenadas por un guion temático) y entrevistas estructuradas (las preguntas tienen respuestas categorizadas previamente).

No obstante lo anterior, las entrevistas en profundidad pueden mezclar diferentes grados de estructuración. Por ejemplo, al inicio las preguntas para caracterizar al entrevistado pueden ser estructuradas, para luego ir ampliando el grado de apertura e incluso llegar a la improvisación de temas.

Las entrevistas en profundidad requieren una preparación previa para elaborar el guion temático y un conocimiento acabado de la política o programa que se está evaluando.

La información levantada por medio de las entrevistas en profundidad se obtiene por medio de grabaciones consentidas por el entrevistado o bien notas del entrevistador cuando la grabación no es consentida, las cuales luego son transcritas en datos textuales. También es importante registrar observaciones respecto al lenguaje corporal del entrevistado o percepciones del entrevistador.

6.3.2 ¿Cuándo utilizar?

Es útil cuando se requiere conseguir información muy compleja, confidencial o delicada. También, cuando se requiere flexibilidad para explorar una política o programa. En general, se puede considerar una fase previa para la elaboración de cuestionarios que permitan identificar los contenidos a incluir en el levantamiento a través de encuestas.

6.3.3 Ventajas/ Desventajas

Ventajas	Desventajas
• Flexibilidad para la adaptación a diferentes contextos de evaluación a través de diferentes tipos de entrevistas. • En las entrevistas menos estructuradas es posible plantear preguntas y ordenarlas de forma flexible según el transcurso de la conversación. • Permite observar gestualidad corporal o expresiones de los entrevistados. • Facilidad de seguimiento. • Permite obtener información completa y detallada de los entrevistados. • Costo bajo respecto a cuestionarios de técnicas cuantitativas.	• La interpretación de las entrevistas personales exige un alto grado de competencia y requiere tiempo. • Cuando existe dificultad para acceder a los entrevistados puede tener un costo alto o involucrar un tiempo relativamente prolongado. • Se obtiene gran cantidad de datos detallados aunque referidos a muy pocas personas. • Se pueden producir las mismas exageraciones y distorsiones que caracterizan los intercambios verbales entre personas. • Los entrevistadores pueden malinterpretar el lenguaje de los entrevistados. • Puede existir discrepancia entre lo que dicen los entrevistados y lo que realmente hacen[25].

6.3.4 Ejemplo(s) de aplicación

Schroeder (2012) trató de identificar los factores que influencian la disposición a unirse a esquemas agroambientales. Para lo anterior, se desarrollaron entrevistas con cuestionarios estandarizados a 32 granjeros en la región de Yorkshire y The Humber en el noreste de Inglaterra durante el año 2010 que ya habían participado en esquemas agroambientales. Con los datos obtenidos se construyeron indicadores sobre las creencias y la evaluación individual de cada granjero respecto a las preguntas de interés. Los resultados mostraron que la actitud general y aceptación del esquema fue alta. En particular, se percibió que el esquema era valioso y permitía una mejora en la biodiversidad, paisaje y recursos naturales. Un resultado percibido, pero no deseado, fue un incremento en las malezas.

Technopolis (2009) buscó identificar el grado del cumplimiento del proyecto FP6 *sub-priority "Global Change and Ecosystems"* de la Unión Europea en términos de impactos científicos, económicos, sociales y políticas. El proyecto mencionado financió 280 proyectos entre 2002 y 2006, de los cuales se escogió una muestra de 94 proyectos. La justificación fue que algunos proyectos no habían sido completados al momento de la evaluación. La metodología utilizó seis métodos complementarios, entre ellos entrevistas en profundidad para recoger ejemplos de los impactos generados y entrevistas

25. Este tema fue abordado por un paper seminal de LaPiere (1934) sobre la diferencia entre actitudes y conducta. El autor viajó durante dos años en Estados Unidos acompañando a una pareja de chinos visitando hoteles y restaurantes. Del total de 251 establecimientos visitados solamente uno se rehusó a atenderlos. Seis meses más tarde LaPiere envió un cuestionario a cada establecimiento preguntando si aceptarían como huéspedes a personas de raza china. De los 128 establecimientos que contestaron solo uno respondió que atendería a personas chinas.

con *stakeholders* para conseguir la perspectiva de los usuarios. El estudio determinó que los impactos ambientales, económicos y sociales fueron indirectos. Además, la mayoría de los proyectos asociados a políticas ambientales contribuía a políticas nacionales e internacionales para incrementar el conocimiento de base y/o el desarrollo de métodos y herramientas.

6.4 GRUPOS FOCALES O FOCUS GROUPS

6.4.1 Descripción

Es una entrevista grupal moderada por un entrevistador que orienta la conversación a partir de una guía temática previamente diseñada. La dinámica grupal generada por la interacción de los participantes (típicamente de seis a ocho personas) permite obtener la opinión de los miembros del grupo, así como también, la opinión compartida por los miembros del grupo. El objetivo principal es identificar actitudes, creencias, experiencias y reacciones de un grupo respecto a una determinada política o programa.

Existen distintos tipos de dinámicas grupales como: grupos de consenso (para ratificar o priorizar criterios), grupos de discusión (pueden ser seleccionados en base a criterios muestrales), grupos naturales (grupo que existe previo a la investigación) y grupos de participación (abierto a distintos miembros de la comunidad). Estos distintos tipos de dinámicas pueden combinarse para alcanzar los objetivos de la evaluación.

6.4.2 ¿Cuándo utilizar?

En una evaluación de impacto, los *focus groups* se pueden utilizar para reunir a distintos grupos de actores implicados en la intervención y analizar sus puntos de vista sobre la política o programa.

También, se puede utilizar cuando se requiere comprender el fundamento de las opiniones expresadas por los participantes en el programa.

6.4.3 Ventajas/ Desventajas

Ventajas	Desventajas
• Forma rápida de obtener impresiones de una política o programa. • Es útil para recopilar información sobre el impacto percibido por los beneficiarios. • Se obtiene información profunda y detallada en un tiempo corto. • Permite obtener diferentes visiones de una misma política o programa. • Puede entregar información confiable con costos mucho menores que los levantamientos masivos de información. • Sirve para extender los resultados obtenidos a través de encuestas u otros datos cuantitativos.	• La necesidad de personal muy entrenado para el manejo del grupo y el análisis de los resultados. • Requiere destreza para generar la dinámica grupal que favorezca la aparición de las diferentes visiones sobre la política o programa. • Es habitual que los participantes se dejen llevar por la presión del grupo. • El registro y análisis de los datos es altamente complejo porque depende de los estilos de comunicación y de las reacciones no verbales de los participantes. • Es difícil generalizar los resultados.

6.4.4 Ejemplo(s) de aplicación

Aunque no se encontraron estudios medioambientales con esta técnica, a modo de ejemplo se puede mencionar el estudio desarrollado para la DIPRES el año 2012 por el Centro de Políticas Públicas de la Pontificia Universidad Católica titulado "Análisis del estado de implementación y el diseño de evaluación del programa Bono Trabajador Activo". En el estudio se utilizaron los resultados de *focus groups* de beneficiarios y encargados regionales del programa en tres regiones del país, como información complementaria de la operatoria del programa. Los resultados de los *focus groups* pusieron en duda la efectividad del programa, por ejemplo respecto a la infraestructura efectivamente ofrecida, específicamente sobre los relatores y técnicas de los cursos.

6.5 PANEL DE EXPERTOS O MÉTODO DELPHI

6.5.1 Descripción

Es una técnica exploratoria que pretende generar un consenso fiable entre las opiniones de un grupo de expertos, a través de una serie de cuestionarios que se responden anónimamente. El primer cuestionario se acompaña de una carta de invitación, en la cual se explica la importancia de su participación, se agradece la colaboración y se indica el plazo de respuesta. Al recibir los cuestionarios se analizan las soluciones y comentarios, agrupándolas en ideas generadas por el grupo. Luego, se envía a los participantes un nuevo cuestionario en el que se incluyen todas las ideas y comentarios generados por el panel, de acuerdo a esta lista se le pide a los participantes que seleccionen y ordenen las ideas más importantes, así como también, que incluyan nuevos comentarios que consideren relevantes. Posteriormente, se envía un nuevo cuestionario en el cual se les comunica a los participantes los resultados del segundo cuestionario y se les pide que realicen un nuevo ordenamiento de las ideas y comentarios. Finalmente, se analizan todas las respuestas del tercer cuestionario y se realiza un informe final en el cual se describen los resultados propuestos por el panel.

Para poder llevarlo a cabo se requiere seleccionar un panel de expertos con diferentes visiones y obtener su compromiso para colaborar, explicándoles cual es el objetivo del estudio. Para seleccionar a los expertos se considera el grado de conocimiento sobre el problema planteado y el interés en solucionar dicho problema.

6.5.2 ¿Cuándo utilizar?

Es una técnica útil cuando se desea realizar pronósticos, identificar problemas, establecer prioridades, establecer metas, resolver problemas y establecer diferencias entre los grupos de referencia.

6.5.3 Ventajas/ Desventajas

Ventajas	Desventajas
• Los expertos expresan libremente sus opiniones. • Crea un clima que favorece la creatividad. • El consenso logrado es confiable. • Evita conflictos entre los expertos por su carácter anónimo. • Los expertos se sienten involucrados en la solución del problema. • No se requiere reunir físicamente a los expertos. • La necesidad de lograr un consenso obliga a un proceso de priorización de las soluciones.	• Para disminuir la carga subjetiva se requiere de varias iteraciones. • El proceso iterativo se puede volver tedioso generando el abandono de algunos expertos. • La necesidad de consenso puede eliminar ideas novedosas. • Para lograr consenso se requiere tiempo e involucra mayores recursos que otras técnicas. • Requiere tener buenos contactos y redes de expertos que disminuyan los tiempos de búsqueda. • Se requiere tener una buena capacidad de síntesis de las respuestas recibidas.

6.5.4 Ejemplo(s) de aplicación

A modo de ejemplo se puede mencionar que Technopolis (2009) entre otras técnicas cualitativas utiliza el juicio de dos expertos para evaluar los resultados de proyectos e identificar el grado del cumplimiento del programa FP6 *sub-priority "Global Change and Ecosystems"* de la Unión Europea, en términos de impactos científicos, económicos, sociales y de políticas ambientales[26].

6.6 ESTUDIO DE CASOS

6.6.1 Descripción

El estudio de casos permite conocer la implementación real de una política o programa mediante una selección estratégica de casos relevantes para la evaluación. El éxito depende de una buena selección y comparación de los casos estudiados. Para esto, es muy importante definir los criterios de inclusión o exclusión de casos a partir de los objetivos de la evaluación.

Se utilizan diferentes métodos para la recolección y análisis de la información como observación, grabaciones o notas de campo.

Los casos exploratorios buscan investigar, comprender y detallar en profundidad todos sus aspectos. Estos casos se guían generalmente en base a preguntas relacionadas con la comprensión de un determinado fenómeno. Por su parte, los casos analíticos buscan analizar el funcionamiento del fenómeno y posibles relaciones de causalidad. Son altamente utilizados para casos en salud (enfermedades) y hacen uso de grupos de control, que no necesariamente son utilizados en los casos exploratorios.

26. Aunque en la revisión bibliográfica no se encontraron estudios medioambientales con esta técnica.

6.6.2 ¿Cuándo utilizar?

Es útil cuando se busca que los casos ilustren un rasgo o problema particular. El estudio de casos también puede ayudar a refinar una teoría o cadena de resultados que permita comprender como se generan los impactos. Varios casos conjuntos permiten establecer condiciones o patrones generales.

También sirve para analizar una política o programa de forma imparcial, al abstraerse de opiniones de los demás o ideas preconcebidas del evaluador que tienden a sesgar la visión del funcionamiento de la política o programa.

6.6.3 Ventajas/ Desventajas

Ventajas	Desventajas
• Sirve para comprender casos reales de una política o programa. • Permite entender relaciones causales complejas y analizar una cadena de resultados. • Permite elaborar teorías o explicaciones integrales de una política o programa. • Permiten reflejar el funcionamiento de una política o programa. • Muestra una experiencia para un pequeño subgrupo de tratados o actores involucrados.	• Los resultados no pueden ser generalizados a la población objetivo. • Los resultados no se pueden extrapolar a otras políticas o programas. • Existe riesgo de sesgo en la selección de los casos. • Se genera demasiada información cualitativa y cuantitativa que puede dificultar la sistematización de la información. • Requiere bastante tiempo. • Tiene un costo relativamente alto comparado con otras técnicas cualitativas.

6.6.4 Ejemplo(s) de aplicación

Shopley & Brasseur (1996) estiman los efectos de subsidios a la energía en cinco empresas grandes y dos empresas pequeñas. Sus resultados afirman que el efecto del subsidio no es claro, ya que seis de las siete empresas redujeron su consumo energético en 20%, pero las reducciones de CO_2 que eran el foco principal sólo fue reducido fuertemente en dos empresas, las otras cinco sólo tuvieron reducciones moderadas.

Malaska, Luukkanen, Vehmas & Kaivo-oja (1997) analizan los costos generados por los impuestos cargados en distintos sectores industriales de Suecia, Dinamarca, Noruega y Finlandia. A partir de estudios de casos se simulan las tasas de impuestos que tendrían las empresas en otros países. Los resultados muestran que algunas empresas finlandesas tendrían una mayor carga tributaria en otros países, lo cual contradice la percepción de los altos impuestos pagados en Finlandia.

6.7 ANÁLISIS MULTICRITERIO

6.7.1 Descripción

El análisis multicriterio es un método que intenta ponderar simultáneamente diferentes criterios para la evaluación de políticas que a menudo están en conflicto. El método permite ordenar los puntajes de los criterios los cuales típicamente son medidos en diferentes unidades. Así, puede ser útil para considerar explícitamente los juicios de valor que están implícitos en los análisis individuales. Permite evaluar un rango más amplio de políticas porque los evaluadores están menos restringidos a considerar sólo los criterios que pueden ser fácilmente cuantificados en términos monetarios. Dado que las preferencias y prioridades de los evaluadores son incorporadas dentro del modelo de decisión, este análisis típicamente implica una evaluación subjetiva.

El método se puede resumir de la siguiente forma:

- Escoger los criterios de evaluación.
- Obtener medidas del funcionamiento para cada criterio.
- Transformar estas unidades a escalas, con el objetivo de combinarlas dentro de una función ponderadora (función de utilidad).
- Ponderar los criterios.
- Ordenar los puntajes que se obtienen.
- Realizar un análisis de sensibilidad.
- Tomar una decisión.

El objetivo central de esta metodología es proveer un mayor rigor analítico para acotar los debates que se dan en los métodos más deliberativos para la evaluación de las políticas. Además, ayuda a abrir el análisis a través de la exploración de las complejidades de algunos temas.

6.7.2 ¿Cuándo utilizar?

Su utilidad depende de que los objetivos puedan ser cuantificados y que los objetivos puedan ser formulados de tal modo que se logre información relevante y significativa sobre el programa o política.

Su aplicación es principalmente en la fase de formulación de las políticas, ya que ayuda a combinar diferentes objetivos complejos que permiten llevar a decisiones mejor fundamentadas. En este contexto el análisis multicriterio es un método de evaluación *ex ante* (Crabbé & Leroy, 2008). Sin embargo, algunas aplicaciones recientes han utilizado el método para realizar evaluaciones *ex post* de políticas ambientales (Soukopová & Bakos, 2013).

6.7.3 Ventajas/ Desventajas

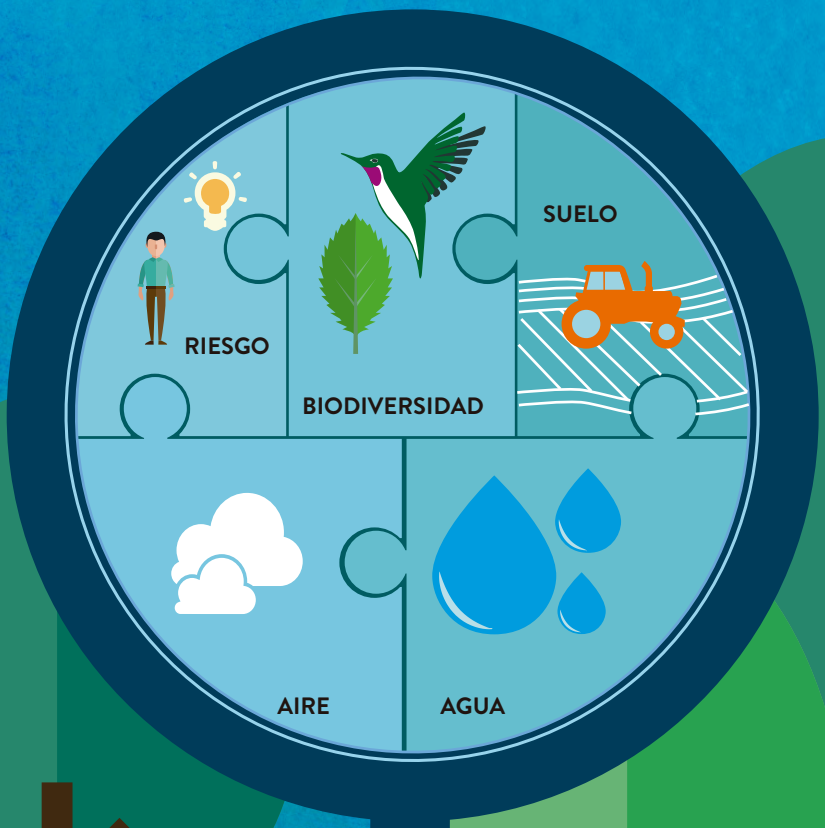
Ventajas	Desventajas
• Es una alternativa útil en situaciones muy complejas con objetivos en conflicto que dificultan otro tipo de evaluación. • Útil cuando la información disponible asociada a múltiples indicadores no es comparable.	• Los resultados son muy sensibles a las ponderaciones que los evaluadores le asignan a criterios específicos. • Debido a la complejidad y naturaleza de la técnica el método no es accesible para personal inexperto. • Los puntajes asociados a cada criterio no siempre están basados en medidas bien definidas, y por lo tanto, son debatibles. • Los puntajes asociados a cada criterio a veces carecen de transparencia.

6.7.4 Ejemplo(s) de aplicación

Finn *et al.* (2009) ilustran una evaluación *ex post* sobre el funcionamiento de esquemas agroambientales en tres regiones de la Unión Europea. La metodología utilizada generó una conexión del proceso de evaluación con los parámetros de diseño del programa, lo cual ayudó a identificar las causas de la poca efectividad del esquema agroambiental.

Soukopová & Bako (2013) implementan un diseño metodológico para evaluar la eficiencia del gasto municipal asociado a la protección del medioambiente. Para ello, definen indicadores basándose en información disponible generada por las autoridades regionales. Según los autores la metodología propuesta permite obtener una visión global de la efectividad y eficiencia en la asignación de los recursos en todas las áreas de gasto a nivel local asociadas a la protección ambiental.

7 | REVISIÓN DE EVALUACIONES *EX POST* EN UN CONTEXTO AMBIENTAL



A continuación se presenta un análisis de experiencias internacionales y nacionales de evaluaciones de impacto *ex post* de políticas o programas con componentes ambientales.

7.1 ESTUDIOS INTERNACIONALES

Para tener la primera aproximación de trabajos que abordaran metologías de evaluación *ex post* en un contexto ambiental, se revisaron diferentes documentos que realizaron revisiones de literatura asociada al tema. A partir de esta búsqueda se pueden mencionar los siguientes estudios.

Feser (2013) a partir del registro de citaciones de Google Scholar ensambladas con el software Zotero (que incluyó 104 artículos en revistas científicas, nueve libros, 12 capítulos de libros y 15 tesis) muestra que de todas las evaluaciones *ex post* con métodos cuasi-experimentales en el área de economía urbana y regional, solamente 4 tratan temáticas ambientales.

Martin *et al.* (2012) realizan una revisión de evaluaciones asociadas a la efectividad del sistema de transacción de emisiones de la Unión Europea. Ellos señalan que la literatura de evaluación *ex post* es más bien pequeña relativo a la literatura *ex ante*. En su revisión utilizan palabras clave en idioma inglés, francés, español y alemán, definiendo un total de 179 publicaciones; luego, por criterios de calidad, incluyen 56 en su análisis, pero de éstos sólo 6 tienen un análisis de causalidad avanzado o de alto nivel, lo cual es un criterio prioritario para definir una evaluación *ex post*.

Agnolucci (2004) realiza una revisión de evaluaciones *ex post* asociadas a impuestos al CO2, este trabajo incluye nueve publicaciones relevantes para la presente revisión.

Otra revisión exhaustiva (en un folleto informativo) que no se limita a evaluaciones *ex post*, incluye diversas experiencias para la promoción de tecnologías de baja emisión de CO2 y gestión del agua. El documento incluye 226 estudios, de los cuales sólo 55 estuvieron enfocados al análisis de programas, y en particular, ocho fueron de interés preliminar para la presente revisión.

Blackman & Woodward (2010) realizan una introducción y guía práctica para la evaluación *ex post* de políticas de conservación forestal, la cual aportó 10 estudios para la presente revisión.

Finalmente, Görlach *et al.* (2005) realizan una guía y revisión de políticas ambientales enfocándose en el análisis de costo – efectividad, estos autores identifican 28 casos de estudio, de los cuales 15 corresponden a análisis de costo – efectividad en un contexto *ex post*, por lo cual fueron incluidos en el presente análisis.

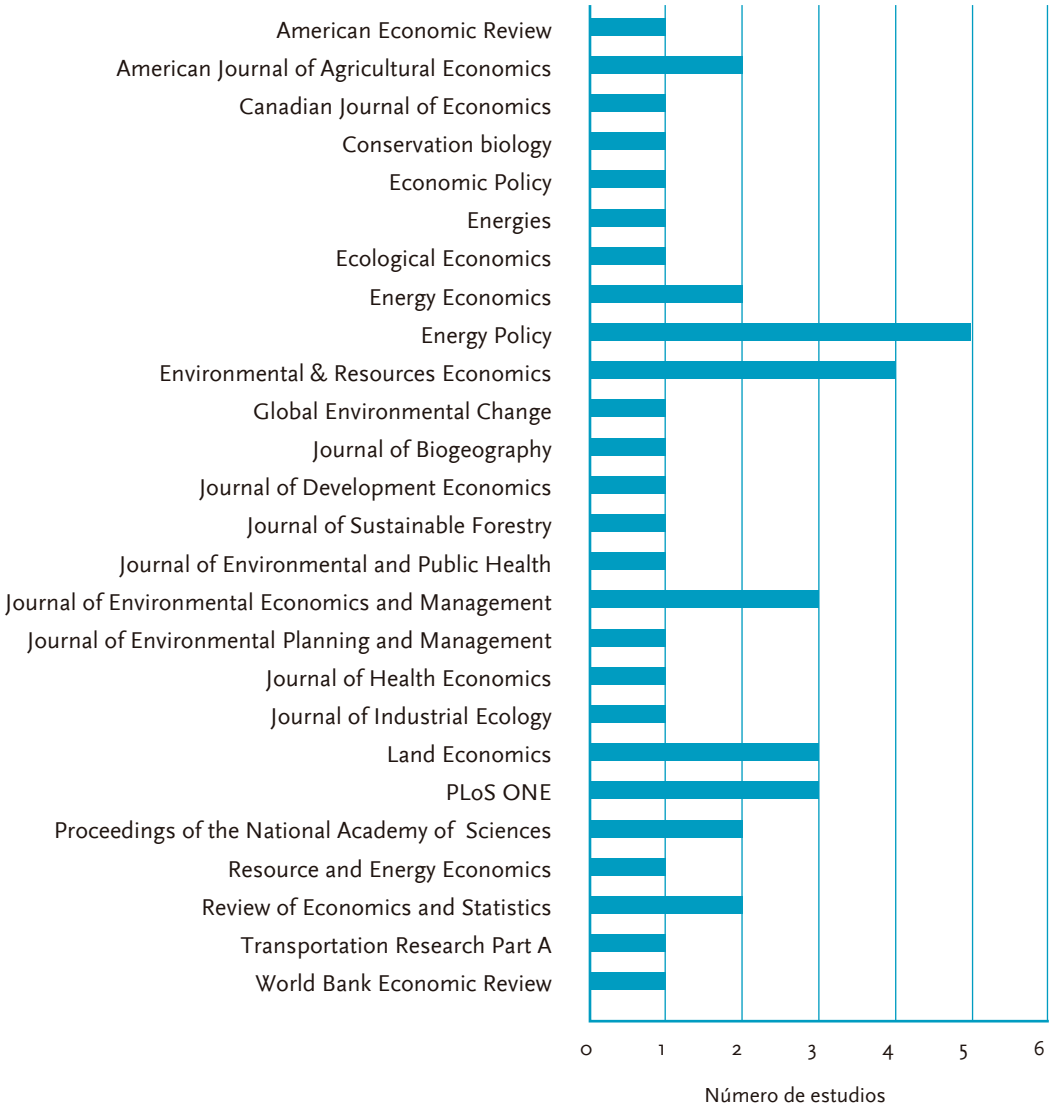
Posterior a esta revisión preliminar, nuevos artículos científicos con metodologías *ex post* asociadas a temáticas ambientales fueron identificados en Science Direct y también se obtuvieron artículos científicos, estudios o guías desde Google Scholar.

Del total de trabajos analizados durante la confección de esta Guía sólo se incluyeron aquellos que pasaron un filtro metodológico, los cuales incluyen evaluaciones cualitativas y cuantitativas. De los 67 estudios finales, 43 están publicados en prestigiosas revistas científicas. Como se observa en la Figura 7-1, la lista de revistas es variada, pero la mayoría

destaca por poseer un alto factor de impacto. Las revistas que más publicaciones han realizado en estas temáticas son Energy Policy y Environmental & Resources Economics.

Figura 7-1. Revistas científicas asociadas a los estudios seleccionados

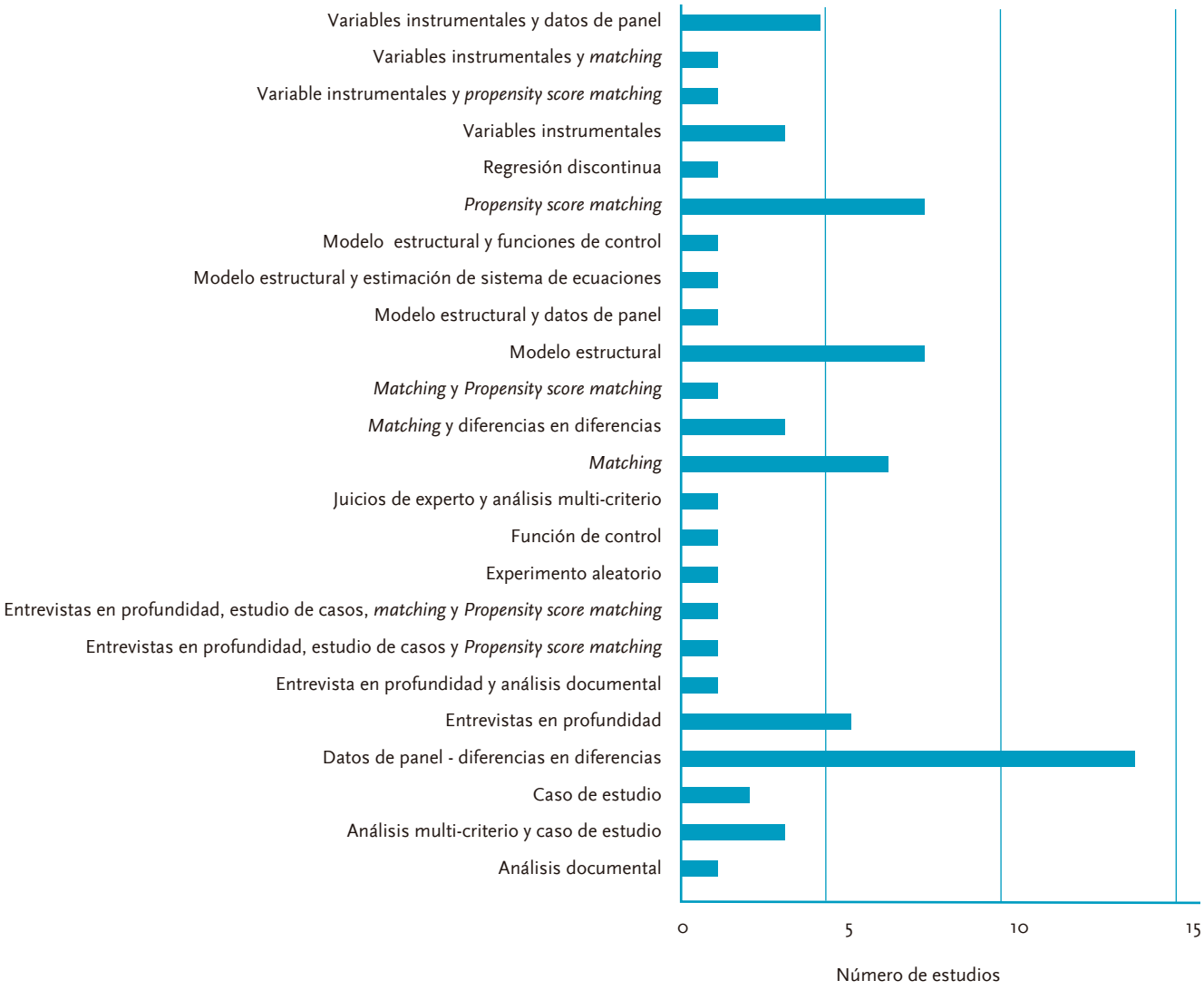
Fuente: Elaboración propia



Luego, de esta exhaustiva revisión bibliográfica se puede concluir de los estudios disponibles que no existe una metodología unificada para abordar la realización de evaluaciones *ex post* en temáticas ambientales. Algunos estudios involucran una simple evaluación cualitativa mientras otros incluyen una metodología de cuantificación más sofisticada.

Además, las técnicas utilizadas son muy variadas predominando el análisis con datos de panel también conocido como diferencias en diferencias, los análisis de *matching* o *propensity score matching* y los modelos estructurales. También, se aprecia en la Figura 7-2 que usualmente existe una combinación de técnicas, tanto si la evaluación es estrictamente cualitativa o cuantitativa.

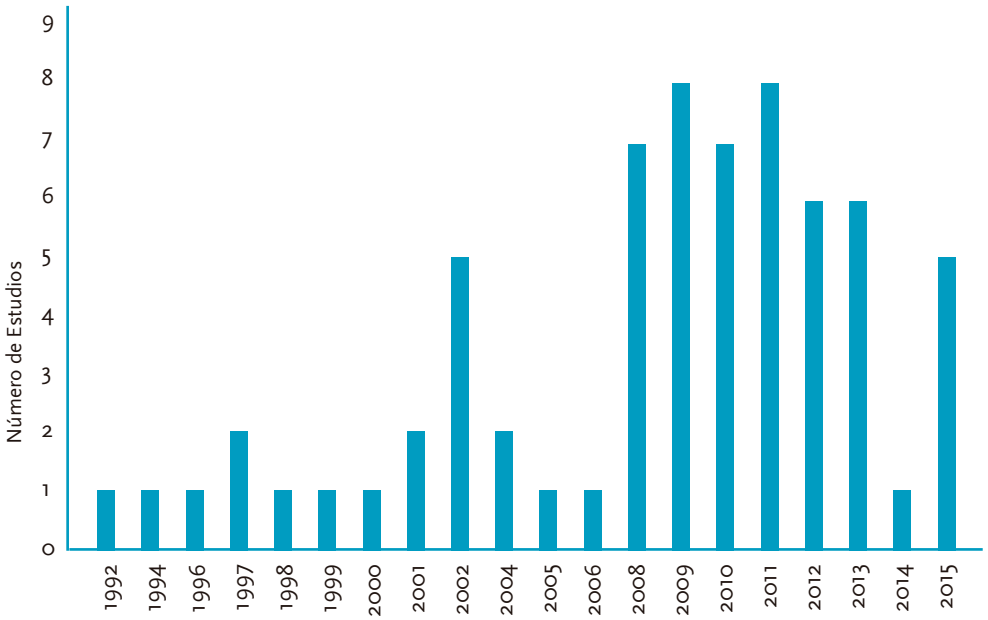
Figura 7-2. Técnicas utilizadas en los estudios seleccionados



Fuente: Elaboración propia

Además, en la Figura 7-3 se observa que casi todos los estudios han sido realizados en las últimas dos décadas, aunque la mayoría de los estudios se concentra a partir del año 2008.

Figura 7-3. Proceso de selección de la revisión bibliográfica



Fuente: Elaboración propia

7.2 ESTUDIOS NACIONALES

Calfucura & Montero (2013) analizan el impacto *ex post* del programa de descontaminación sobre la calidad del aire en Santiago de Chile. Para ello, utilizan técnicas de regresión discontinua y comparan los resultados con los obtenidos por mínimos cuadrados ordinarios. Los resultados muestran que en sólo la estación de monitoreo de Pudahuel se observaría una disminución significativa con el enfoque de regresión discontinua. Sin embargo, estos resultados podrían reflejar el hecho que el conjunto de medidas o al menos muchas medidas incorporadas en el PDA tienen un efecto gradual y no se hacen efectivas justo en la fecha oficial, lo cual pone en entredicho la capacidad de utilizar regresión discontinua en una pequeña ventana temporal para estimar los impactos de este plan.

Mullins & Bharadwaj (2015) examinan los efectos de corto plazo de un conjunto de políticas iniciadas en 1997 en Santiago para reducir la contaminación durante episodios ambientales. La dificultad para la evaluación *ex post* surge porque los días con episodios ambientales están por sobre el promedio de las concentraciones, por lo cual se debería esperar que los niveles de contaminación se reduzcan en los días siguientes independiente de si declara alerta ambiental o no. Para abordar este problema, se utiliza el enfoque de diferencias en diferencias en conjunto con *propensity score matching* para emparejar cada día de episodio ambiental con días similares en los cuales no hubo episodio. Los resultados muestran que decretar episodios ambientales para reducir la contaminación en el corto plazo son altamente efectivos, ya que las concentraciones de PM10 se reducen cerca de 20%. Además, muestran que tres días después de la implementación de un episodio se genera una reducción de 15 muertes en la población sobre 64 años.

Mardones (2016) realiza una evaluación *ex post* del Programa Fondo de Protección Ambiental con datos a nivel comunal utilizando una metodología de variables instrumentales. Este programa de carácter anual tiene por objetivo financiar proyectos

que mejoren el medioambiente a partir de diversas iniciativas comunitarias que promueven la educación y la participación ciudadana. Los resultados muestran que la efectividad del programa medido como la mejora en la percepción comunitaria sólo se logra en uno de diez componentes ambientales, por lo cual se concluye que es imprescindible una reasignación de recursos para que mejore el impacto y la costo-efectividad del programa.

Fuera del ámbito más académico, se puede mencionar que el Consejo de Producción Limpia ha llevado a cabo algunas evaluaciones que se presentan en el documento titulado "La Experiencia de los APL: 1999-2005" en el cual se menciona que los APL han tenido impacto en el ámbito económico productivo, el ámbito normativo y de control, y el ámbito ambiental y social. De acuerdo a este informe 14 de los 28 APL están orientados a reducir emisiones atmosféricas. Además, se realiza un análisis detallado de casos. Específicamente, se detalla el APL de fundiciones el cual habría reducido 244 ton/año de emisiones de material particulado, el APL de ladrillos que habría reducido 20,7 ton/año y el APL de la construcción que habría reducido 246 ton/año. También, se menciona que 16 de los 28 APL están orientados a reducir los volúmenes de Residuos Industriales Líquidos (RILes); específicamente se detalla el APL de Cerdos el cual habría reducido 50% los RILes y aumentado la superficie de riego hasta 20%, mientras el APL de Salmones habría reducido en 48% la disminución de aceites y grasas, 29% la DBO5, 46% los sólidos suspendidos y 85% el poder espumógeno. Finalmente, 23 de los 28 APL están orientados a la minimización, reutilización y reciclaje de residuos, específicamente el APL de Salmones habría aumentado el reciclaje de desechos de pescado en un 40% y reciclaje de plástico en 172%. No obstante lo anterior, resulta necesario aclarar que estas estimaciones se basan en un análisis antes – después, por lo cual, técnicamente no corresponderían a una evaluación *ex post* cuantitativa que haya identificado adecuadamente el efecto causal de los APL.

El informe "Uso de Tecnologías Limpias: Experiencias Prácticas en Chile" describe tecnologías aplicadas en diversos sectores industriales. A pesar que se reportan mejoras ambientales, también se basan en un análisis antes – después que, tal como se mencionó previamente, no es una técnica válida de evaluación *ex post*. Por último, se presenta un análisis de caso asociado al APL del sector productor de huevos. En este documento se mencionan y cuantifican beneficios ambientales basados en un análisis antes – después.

A pesar que los diferentes estudios relacionados a los APL no siguen una metodología cuantitativa válida en términos de evaluación de impacto, es de interés mencionar estos esfuerzos porque pueden servir como información para futuras evaluaciones *ex post*.

Por otra parte, de acuerdo a la información oficial de la DIPRES que detalla los diferentes programas evaluados por cada Ministerio en la categoría "Evaluación de Impacto de Programas (IE)" aparecen un par de evaluaciones de impacto *ex post* que podrían enmarcarse en el contexto de un programa ambiental[27].

"Consultorías Profesionales Agraria" evaluaron el programa de Bonificación Forestal D.L. N°701 de 1974 del Ministerio de Agricultura[28]. En este estudio el impacto se estimó con la técnica del *Propensity Score Matching* y Diferencias en Diferencias, concluyendo que el programa no tiene impacto en el ingreso de los propietarios que se acogieron a la bonificación forestal. Tampoco tiene impacto sobre la superficie forestada, por lo cual las externalidades positivas obtenidas del control de la erosión en la superficie forestada

27. http://www.dipres.gob.cl/595/w3-propertyvalue-23076.html

28. http://www.dipres.gob.cl/574/articles-141063_informe_final.pdf

con bonificación no pueden ser atribuidas al programa.

También, existe un estudio para el diseño de la evaluación de impacto del programa "Sistema de Incentivos para la Sustentabilidad Agroambiental de los Suelos Agropecuarios" en la categoría "Evaluación de Proyectos Nuevos" de la DIPRES. Este estudio realizado por el Centro de Microdatos de la Universidad de Chile detalla estrategias empíricas de evaluación, ventajas y desventajas, así como los resultados posibles a evaluar y las limitaciones para proyectar los resultados encontrados a la población objetivo. Además, incluye una descripción de la información que se requiere levantar y la información disponible para una futura evaluación de impacto *ex post*

Este mismo programa fue evaluado por Caro, Melo & Forester (2006) quienes analizan su impacto en usuarios de INDAP con la técnica de función de control, específicamente utilizan el modelo de selección de Heckman (Heckit). No obstante, los autores no encuentran que exista un problema de autoselección en los datos y por ello, finalmente, terminan utilizando los resultados estimados por mínimos cuadrados ordinarios. Sus resultados sugieren que hay efectos directos positivos y significativos atribuibles al programa, asociados específicamente a los componentes de fertilización fosfatada y recuperación de praderas.

Otro estudio que aborda la evaluación del mismo programa es EMG (2002), al cual no se tuvo acceso ya que no está disponible en Internet pero que fue analizado en el estudio realizado por el centro de Microdatos de la Universidad de Chile. En el citado estudio se afirma que comparte la metodología desarrollada por Caro, Melo & Forester (2006) y, además, que en ambos estudios la selección del grupo de control es considerada "poco rigurosa".

En la DIPRES también existe la categoría "Evaluaciones de Programas Gubernamentales (EPG)" que incluye algunos programas que abordan aspectos ambientales. El objetivo de la EPG es evaluar el diseño, gestión y resultados de los programas públicos para apoyar su gestión, basada en la metodología del marco lógico utilizada por el Banco Mundial y el BID. Los programas evaluados en esta categoría son los siguientes:

- Subsidio Protección del Patrimonio Familiar, Ministerio de Vivienda y Urbanismo, SERVIUs (año de evaluación 2011).
- Programa Nacional de Eficiencia Energética, Ministerio de Energía, CNE (año de evaluación 2008).
- Programa de Prevención y Control de la Contaminación, Ministerio Secretaría General de la Presidencia y CONAMA (año de evaluación 2008).
- Programa de Recursos Naturales y Biodiversidad, Ministerio Secretaría General de la Presidencia y CONAMA (año de evaluación 2008).
- Aplicación Limpieza de Calles, Ministerio del Interior, Gobierno Regional R.M. de Santiago (año de evaluación 2005).
- Sistema Nacional de Áreas Silvestres Protegidas, Ministerio de Agricultura, CONAF (año de evaluación 2005).
- Programa Sendero de Chile, Ministerio Secretaría General de la Presidencia y CONAMA (año de evaluación 2005).
- Programa Borde Costero, Ministerio de Defensa, Subsecretaría de Marina (año de evaluación 2004).
- Programa Fondo Protección Ambiental, Ministerio Secretaría General de la Presidencia CONAMA (año de evaluación 2001).

- Participación Ciudadana en Instrumentos Gestión, Ministerio Secretaría General de la Presidencia CONAMA (año de evaluación 2001).
- Recuperación de Suelos Degradados, Ministerio de Agricultura, INDAP y SAG (año de evaluación 2000).
- Proyecto de Protección de la Capa de Ozono, Ministerio Secretaría General de la Presidencia y CONAMA (año de evaluación 2000).
- Manejo y Diversificación Forestal, Ministerio de Agricultura, CONAF (año de evaluación 1999).

Estas evaluaciones EPG analizan el programa a través de un panel de evaluadores quienes realizan una verificación del marco lógico, para luego hacer recomendaciones sobre el diseño, analizar aspectos de la organización y gestión del programa y, también generar conclusiones sobre la eficacia y eficiencia del programa a través de algunos indicadores. Sin embargo, estos estudios no encajan en el concepto de evaluación de impacto *ex post* cuantitativa analizado en esta guía metodológica.

Por otra parte, según información del Banco Integrado de Programas Sociales o BIPS (plataforma pública que centraliza todas las evaluaciones que ha realizado el Estado para cada programa social) existen los registros de dos programas asociados al Ministerio de Medioambiente.

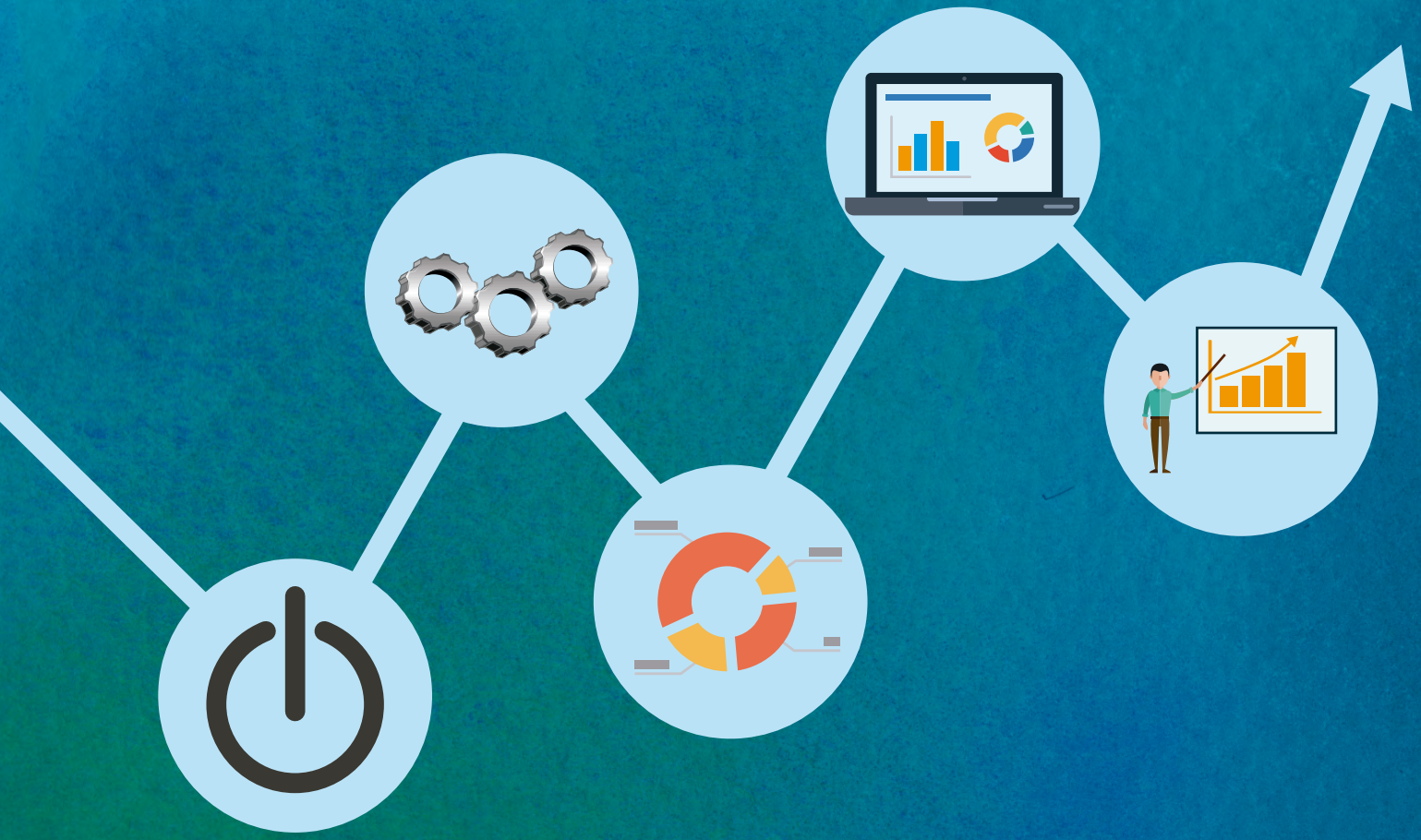
- Fondo de Protección Ambiental (Presupuesto: $1.077 millones el año 2015, $1.056 millones el año 2014, $1.056 millones el año 2013, $1.028 millones el año 2012).
- Recambio de Calefactores a leña (Presupuesto $3.347 millones el año 2015, $1.781 millones el año 2014, $1.712 millones el año 2013, $1.776 millones el año 2012).

Y además, otros programas con componente ambiental ejecutados por diversos Ministerios:

- Ley de bosque nativo, Ministerio de Agricultura, CONAF (Presupuesto $6.754 millones el año 2015, $6.139 millones el año 2014).
- Protección Contra Incendios Forestales, Ministerio de Agricultura, CONAF (Presupuesto $17.797 millones el año 2015, $16.879 millones el año 2014).
- Áreas Silvestres Protegidas, Ministerio de Agricultura, CONAF (Presupuesto $14.129 millones el año 2015, $12.546 millones el año 2014).
- Protección del Medio Ambiente y Recursos Naturales, CONADI (Presupuesto $53 millones el año 2013).

De acuerdo al sitio web de BIPS el FPA fue evaluado por la Universidad Diego Portales (2010). Este fondo apoya iniciativas ciudadanas de recuperación, fomento o protección del medio ambiente. El FPA existe hace más de una década y al momento de su evaluación había involucrado a más de 4.650 organizaciones sociales, financiando alrededor de 1.450 proyectos con una cobertura que alcanzó a más del 75% de las comunas del país. Los resultados de la evaluación señalan que el nivel de conciencia ambiental de la organización ejecutora del proyecto FPA es significativamente superior al grupo de control, y que el grupo de control posee un índice de conciencia ambiental superior al exhibido por la comunidad. Por lo anterior el estudio puede ser criticado porque no construye un grupo de control válido.

8 | IMPLEMENTANDO UNA EVALUACIÓN *EX POST*



En esta sección se abordan los criterios que se pueden utilizar para priorizar las políticas o los programas a evaluar, los requerimientos de información bajo un contexto experimental o no experimental, la definición de la población objetivo, los requerimientos de muestreo, la determinación del tamaño muestral, el proceso de levantamiento de información, el diseño del cuestionario, la opción de realizar una evaluación cualitativa para complementar la evaluación cuantitativa y la estructura tentativa del informe para una evaluación *ex post* en un contexto ambiental.

8.1 CRITERIOS PARA REALIZAR UNA EVALUACIÓN *EX POST*

A pesar de la importancia de conocer las evidencias sobre las políticas públicas, no todos los programas ameritan realizar una evaluación de impacto *ex post*, ya que estas evaluaciones al ser costosas requieren una priorización de los recursos.

Para decidir si es conveniente realizar una evaluación de impacto se pueden analizar ciertos criterios básicos como el presupuesto asignado al programa, la magnitud del resultado esperado, evidencias del éxito de programas similares nacionales o extranjeros, novedad del programa, replicabilidad en otros contextos, posibilidad de expansión, pertinencia estratégica, efectividad no comprobada, entre otros.

Rossi *et al.* (2004) señalan que existen tres factores claves:

- El tiempo en el que se espera que ocurran los efectos, ya que si se decide evaluar demasiado pronto el impacto puede ser nulo.
- La incertidumbre sobre la magnitud del impacto, que prioriza las evaluaciones de programas en los cuales se desconoce el efecto causal debido a la falta de experiencias previas.
- El costo de recolección de la información.

Otros criterios relevantes mencionados por Blasco & Casado (2009) para realizar una evaluación de impacto se relacionan con que haya pasado el tiempo suficiente para haber alcanzado una estabilidad en el impacto del programa, haber identificado adecuadamente en el diseño del programa los impactos que deben estimarse, contrastar la planificación con el proceso de implementación del programa para interpretar adecuadamente los resultados de la evaluación y escoger el momento adecuado para medir el impacto, ya que algunos efectos pueden tardar en producirse, acumularse o desaparecer con el tiempo.

Por otra parte, las evaluaciones de impacto pueden planificarse en conjunto con el diseño del programa (evaluación prospectiva) o alternativamente después de la implementación del programa (evaluación retrospectiva). En el primer caso es posible recolectar datos de línea base para los grupos de tratamiento y control los cuales son identificados antes de aplicar el programa, lo cual permite obtener resultados más válidos y creíbles. En el segundo caso, no existe información sobre la línea base para tratados y controles, lo cual lleva a utilizar bases de datos existentes que limitan la posibilidad de encontrar estimaciones contrafactuales válidas porque se basan en más supuestos para la utilización de métodos cuasi - experimentales.

Para establecer la priorización de programas ambientales a evaluar en Chile se generó una lista de criterios que abordan las temáticas básicas sugeridas en la literatura de evaluación *ex post*, así como también algunas características específicas de acuerdo a la realidad nacional. Los criterios se detallan a continuación.

- Oportunidad: que el programa esté actualmente en operación o esté en la etapa de diseño para su inicio futuro.
- Frecuencia: que el programa se realice de forma habitual o permanente.
- Tiempo de efectos: que el programa ya iniciado lleve tiempo en operación para que sus resultados estén asentados.
- Recursos involucrados: que el programa involucre un desembolso importante de recursos públicos.
- Impacto ambiental: que el programa de acuerdo a su diseño o en base a las evaluaciones *ex ante* tenga un impacto significativo en términos ambientales.
- N° de Beneficiarios: que el programa tenga un número significativo de beneficiarios o agentes regulados.
- Análisis Técnico: que el programa tenga la posibilidad real de construir cuantitativamente un escenario contrafactual válido.
- Extrapolación: que los resultados del programa identificados a través de una muestra puedan ser extrapolados a toda la población objetivo.
- Información disponible: que el programa tenga estudios cualitativos previos, información útil de línea base o seguimiento que facilite su evaluación o la generación de un grupo contrafactual.
- Evaluación previa: que el programa no tenga evaluaciones cuantitativas previas por lo cual exista la necesidad de generar una evaluación en el corto plazo.
- Multiplicidad de técnicas: que el programa o política pueda ser abordado por diversas técnicas cuantitativas.
- Alcance geográfico: que el programa o política tenga impacto nacional o en grandes zonas geográficas.
- Relevancia de la política: que el programa forme parte de compromisos internacionales a nivel país (OCDE, Objetivos del Milenio, entre otros), metas programa de gobierno, cruce con otros programas (que su éxito influya en el desempeño de otros programas).

8.2 REQUERIMIENTOS DE INFORMACIÓN EN UNA EVALUACIÓN *EX POST*

Los requerimientos de información en una evaluación *ex post* dependen de la propuesta de diseño que se desea utilizar. A continuación se analiza una alternativa de diseño experimental y otra no experimental.

8.2.1 Propuesta de diseño experimental

Un experimento aleatorio es una técnica considerada robusta para estimar el impacto causal mientras se pueda cumplir con el diseño de asignación.

La asignación aleatoria consiste en una invitación, para las unidades a participar en el programa, en la cual se les entrega la información de cómo inscribirse y los procesos que deben seguir. Esta asignación se podría basar en un concurso especial que tenga los mismos requisitos que los programas normalmente ejecutados cada año, salvo el hecho que estará focalizado sólo en la muestra aleatoria escogida. Otra opción, más fácil podría ser generar una asignación aleatoria sobre los postulantes al programa.

Para facilitar la recolección de información en las unidades del grupo de control, se podría entregar un incentivo monetario.

En el caso que no todas las unidades invitadas decidan participar, de todas formas la asignación aleatoria del programa estará altamente correlacionada con la participación, aunque existan otras variables no observables que afecten la decisión de participar (no utilizar el incentivo). Así, la asignación aleatoria se puede interpretar como una variable instrumental para identificar el impacto del tratamiento, ya que las unidades invitadas son más propensas a participar en el programa y, además, esta variable instrumental no está correlacionada con las variables no observables que afectan el resultado.

Bajo estas propuestas se requeriría levantar al menos una encuesta de seguimiento, aunque también sería útil contar con una encuesta de línea base.

8.2.2 Propuesta de diseño no experimental

En el caso que la asignación aleatoria decida no ser realizada, la evaluación *ex post* se puede ejecutar con otras técnicas alternativas como diferencias en diferencias o *matching* con diferencias en diferencias, entre otras.

Dado que los incentivos del programa no se asignan aleatoriamente entre el grupo de postulantes, sino que a partir de puntaje, postulación voluntaria (autoselección), criterios de elegibilidad, u otros, la asignación sobre los beneficiarios no es aleatoria sino que se basa en características observables y no observables.

Por lo anterior, una alternativa de evaluación consiste en tomar como grupo de tratamiento a los beneficiarios en el año inicial del programa, a los cuales se les podría aplicar un cuestionario para identificar las variables de resultado y características en la actualidad (encuesta de seguimiento) y, además, se les consultaría por la misma información pero en el año previo a la asignación del programa (línea base), es decir, información retrospectiva. Finalmente, se requeriría aplicar estos mismos cuestionarios a un grupo de control que no haya sido beneficiario.

Otra alternativa de evaluación consiste en tomar como grupo de tratamiento a los beneficiarios en el año previo al inicio del programa, a los cuales se les podría aplicar un cuestionario para identificar las variables de resultado y características en la actualidad (línea base) y, luego, en un periodo posterior a la asignación del programa se les consultaría por la misma información (encuesta de seguimiento). Finalmente, se requeriría aplicar estos mismos cuestionarios a un grupo de control que no haya sido beneficiario.

Para desarrollar un diseño cuasi-experimental o no experimental, se requiere que la base de datos contenga información equivalente para el grupo de tratamiento y grupo de control y debería contener variables asociadas a los resultados y características de las unidades. La información levantada se podría complementar con la información obtenida en las postulaciones al programa.

Cabe señalar que la información retrospectiva es muy importante para asegurarse que previo al tratamiento las variables son estadísticamente iguales, en las unidades escogidas para realizar la comparación.

En todo caso la información levantada bajo este diseño debe incluir las mismas temáticas que las propuestas bajo el diseño experimental, la única diferencia es que bajo la alternativa que requiere información retrospectiva se deben consultar los datos para el año en el cual se asigna el tratamiento como para un periodo previo.

Al igual que en la metodología experimental, bajo esta propuesta se realizaría una evaluación de impacto justo después de levantar la encuesta de seguimiento, aunque también se podría realizar otra evaluación posterior si se incluye una segunda encuesta de seguimiento.

8.3 DEFINICIÓN DE POBLACIÓN OBJETIVO

Para definir la población objetivo se requiere determinar el alcance geográfico y características específicas de las unidades sobre las cuales se medirán los indicadores de resultados (por ejemplo: hogares urbanos del quintil 1 y quintil 2 de la zona central del país). Para que los resultados tengan validez externa se requiere que la muestra sea representativa de la población de interés, ya sea definida como todas las unidades potenciales o bien como las unidades postulantes preseleccionadas. Para esto se utilizan generalmente muestreos aleatorios como el muestreo aleatorio simple, el muestreo aleatorio estratificado y el muestreo por conglomerados.

Además, una vez que se tiene la base de datos completa con la información para un grupo de tratamiento y el grupo de control es posible realizar las estimaciones de impacto a través de indicadores para toda la muestra, por región o ciudad, o para otros subgrupos de la muestra que sean de interés.

8.4 REQUERIMIENTOS DE MUESTREO

Un tema importante para la evaluación *ex post* es determinar la muestra que se requiere para determinar la diferencia en los indicadores de resultados entre el grupo de tratamiento y de control. Por ello, luego de definir cuál será la población objetivo, se debe determinar cuáles serán los indicadores de resultados a medir y en qué periodos se levantará la información.

Se desean medir se debe tener conciencia que los indicadores que presentan mucha variabilidad requerirán muestras más grandes, o alternativamente, los indicadores que se ven poco afectados por la intervención, podrían no ser útiles para la evaluación de impacto.

El levantamiento debe ser planificado de forma adecuada, ya que en la mayoría de las técnicas cuantitativas se requiere contar con datos de línea base y al menos una encuesta de seguimiento.

La elección de la técnica cuantitativa también influye en los requerimientos de información, bajo un experimento aleatorio existen requerimientos precisos de tamaños mínimos muestrales de acuerdo a la variabilidad del indicador, el nivel de confianza y la potencia requerida. Por otra parte, las técnicas que controlan por variables observables como el *matching* requieren gran cantidad datos sobre las características de los grupos de tratamiento y grupo de control, mientras en la técnica de diferencias en diferencias además se requiere de variables o factores exógenos que evolucionen en el tiempo para ambos grupos.

La disponibilidad de datos existentes puede contribuir considerablemente a la evaluación de impacto en términos de tiempo y costo. No obstante, no es usual que estos datos sean suficientes, ya que en general las bases de datos existentes (por ejemplo: Encuesta CASEN) o registros administrativos no incluyen a un número importante de unidades del grupo de tratamiento y control, pueden contener un conjunto reducido de variables caracterizadoras, poseer datos desactualizados o bien disponer de datos que no coinciden con los periodos de tiempo requeridos.

En general, los datos requieren contar con suficientes observaciones en el grupo de tratamiento y control que permitan detectar los cambios en los indicadores de resultado con una potencia estadística suficiente, que el marco muestral sea representativo de la población objetivo y que sean recolectados con la periodicidad requerida para la evaluación.

8.5 TAMAÑO MUESTRAL

El tamaño mínimo de la muestra requerido para realizar una evaluación de impacto se determina a través de un cálculo de potencia (la probabilidad de no cometer un error del tipo II[29]).

Esta muestra debe permitir identificar diferencias significativas en los resultados entre el grupo de tratamiento y grupo de control. Por lo tanto, es un aspecto crucial para determinar el éxito o fracaso de un programa a través de una evaluación *ex post*.

Los cálculos de potencia son distintos si el programa asigna un tratamiento aleatoriamente entre conglomerados (ciudades, barrios, escuelas, etc.) cuyo efecto se mide a nivel individual (hogares o individuos), o si el programa asigna un tratamiento aleatoriamente entre todas las unidades de una población (hogares o individuos). Si los indicadores de resultados se van a comparar entre subgrupos (por ejemplo: según quintil de ingreso) se requiere un cálculo distinto de la potencia y finalmente una muestra más grande.

El tamaño muestral obviamente será mayor con supuestos más conservadores (como un menor impacto previsto, una mayor varianza del indicador del resultado o un nivel de potencia superior). Se pueden estimar cálculos de potencia para varios indicadores de resultado, ya que los tamaños serán distintos si algunos indicadores de resultado tienen mayor variabilidad que otros. Además, es recomendable disponer de un tamaño algo más grande al mínimo estimado (10% a 20%) por posibles factores que afecten la estimación como cuestionarios incompletos o la atrición de la muestra generada porque hay unidades que abandonan el tratamiento, emigran o ya no quieren seguir contestando las encuestas de seguimiento.

Para especificar el tamaño del efecto que se pretende detectar, se podría considerar el menor efecto detectable que permita considerar que el programa fue un éxito. Para otros contextos, Cohen (1988) señala que un efecto de 0,2 desviaciones estándar se considera pequeño, 0,5 desviaciones estándar se considera un efecto medio y un efecto de 0,8 desviaciones estándar se considera grande.

29. Se comete error de tipo I si se determina que un programa ha tenido impacto cuando en realidad no lo ha tenido. Se comete un error de tipo II si se determina que un programa no ha tenido impacto cuando en realidad lo ha tenido. Para reducir el error de tipo I, se puede establecer un nivel de confianza de 95% o 99%, mientras para reducir el error del tipo II se puede definir un tamaño de muestra más grande. La mayoría de los investigadores utilizan una potencia de 0,8, lo cual significa que se observará un impacto en el 80% de los casos en que se haya producido.

8.6 CÁLCULO DE TAMAÑO MUESTRAL PARA EL DISEÑO NO EXPERIMENTAL

Bajo el diseño cuasi – experimental o no experimental se sugiere utilizar un tamaño igual o mayor a los tamaños muestrales definidos con un enfoque experimental. No obstante, el número de unidades tratadas podría estar ya definido cuando se aplique la primera encuesta, y por lo tanto, en este caso sería útil contar con un número similar o superior de unidades no tratadas en cada zona geográfica.

Si existe disponibilidad de recursos para el levantamiento de información, se podría considerar un cálculo de tamaño muestral para cada zona geográfica seleccionada, lo cual elevaría considerablemente el costo de las mediciones.

8.7 LEVANTAMIENTO DE INFORMACIÓN

Se requiere que los procedimientos e instrumentos para levantar la información en el grupo de tratamiento y grupo de control sean iguales, incluyendo los mismos periodos de tiempo para que el comportamiento no sea distinto (ejemplo: temporada de frío o calor en el consumo de energéticos en los hogares) y los mismos métodos de recolección de datos (ejemplo: entrevistas personales o entrevistas telefónicas). No obstante, a veces esto es complejo, ya que existe una mejor relación o disposición para la entrega de información en el grupo de tratamiento. Además, el levantamiento de información debería ser idealmente realizado por una institución independiente.

El objetivo general de las encuestas de línea base es construir una base de datos de las variables de resultados y de características de las unidades preseleccionadas.

Los objetivos específicos de las encuestas de línea base son los siguientes:

- Levantamiento de encuestas en terreno para las unidades preseleccionadas.
- Cuantificar el indicador de resultado en el grupo de tratamiento y grupo de control.
- Obtener las variables caracterizadoras de las unidades preseleccionadas en ambos grupos.

Los datos de línea de base son útiles aun cuando la técnica cuantitativa no lo requiera (por ejemplo: experimento aleatorio, regresión discontinua, variables instrumentales o función de control). Por ejemplo contar con datos de línea base permite realizar un análisis de diferencias en diferencias si es que falla un experimento aleatorio, si la variable instrumental es débil, si se demuestra que las características entre tratados y no tratados saltan en el punto de corte de la regresión discontinua, o bien, si no existe continuidad en el puntaje de elegibilidad para la regresión discontinua.

El objetivo general de las encuestas de seguimiento es complementar la base de datos existente de línea base con los cambios en las variables de resultados y características de las unidades preseleccionadas en los periodos siguientes a la implementación del programa. Esta base de datos permitirá posteriormente analizar los resultados o

impactos atribuibles al programa, y evaluar futuras mejoras tanto en diseño como en implementación.

Los objetivos específicos de las encuestas de seguimiento son los siguientes:

- Generar un segundo levantamiento de encuestas en terreno para las unidades preseleccionadas.
- Cuantificar el indicador de resultado en el grupo de tratamiento y grupo de control utilizando la misma encuesta utilizada para la línea base.
- Obtener las variables caracterizadoras de las unidades preseleccionadas en ambos grupos que permita la aplicación de técnicas cuantitativas de evaluación de impacto.

8.8 DESARROLLO DEL CUESTIONARIO

El instrumento para la recolección de los datos debe ser formulado de tal forma que incluya toda la información necesaria para responder a las preguntas de la evaluación de impacto. Esto requiere tanto la determinación de los indicadores de resultado de interés como también las variables caracterizadoras de cada unidad y factores exógenos.

En los cuestionarios a veces se incluyen demasiadas preguntas lo cual alarga innecesariamente la duración de las entrevistas, por ello una forma de reducir su extensión es determinar si una pregunta incluida en el cuestionario es útil, al preguntarse qué tipo de análisis pueden requerir de esa información. Así, si una pregunta no va a ser realmente utilizada es mejor no incluirla y priorizar aquellas que son necesarias.

Además, los cuestionarios deben ser validados en pruebas pilotos para chequear la claridad de las preguntas, la continuidad de las preguntas, la comprensión del lenguaje y determinar su extensión real.

Por ejemplo, una encuesta destinada a levantar información para evaluar el impacto de un programa de recambio de calefactores podría caracterizar los siguientes aspectos:

- Dirección de la vivienda.
- El tipo de inmueble.
- El número de integrantes del hogar.
- El volumen de leña u otro combustible utilizado para calefacción.
- La fecha y canales de adquisición de la leña u otros combustibles.
- El precio que pagan por la energía (gas, electricidad, combustibles de biomasa, u otro).
- Los tipos de artefactos de combustión utilizados.
- La temporalidad y estacionalidad del consumo.
- La forma de operación de los artefactos.
- El nivel socioeconómico.
- Los materiales de construcción y aislamiento térmico de la vivienda.
- La percepción de confort térmico.

- Humedad de la leña (medición con xilohigrómetro).
- Preguntas situacionales.
- Otras variables de interés.

8.9 DISEÑO Y APLICACIÓN DEL CUESTIONARIO DEFINITIVO

Antes de la aplicación del cuestionario para la línea base y encuestas de seguimiento, se debería realizar una validación del instrumento, para lo cual se propone revisar encuestas previas sobre estas temáticas, *focus groups* con beneficiarios previos y expertos con experiencia en evaluación *ex post*. El instrumento propuesto debería ser diagramado en un formato profesional para permitir una fácil administración de las respuestas. Luego, debería ser testeado en terreno aplicándolo a un conjunto de unidades para verificar su redacción, calidad y fluidez. Finalmente, se deben realizar los ajustes que surjan de la etapa previa para obtener el cuestionario definitivo.

Además, se debe contar con estrictos protocolos para la capacitación de encuestadores, trabajo en terreno, digitación, codificación, validación y control de calidad.

8.10 EVALUACIÓN CUALITATIVA

Los estudios cualitativos permiten generar hipótesis, entender el efecto causal del programa y complementar e interpretar los resultados cuantitativos. Por esto se propone desarrollar un estudio cualitativo a los beneficiarios tanto en el periodo de línea base, como en la ronda de seguimiento luego de la ejecución del programa.

El objetivo general de la evaluación cualitativa inicial, antes del levantamiento de línea base, es ayudar al diseño definitivo del cuestionario que se aplicará para el levantamiento de información cuantitativa.

El objetivo general de la evaluación cualitativa en la ronda de seguimiento es conocer la percepción de los actores involucrados sobre las dificultades y beneficios generados por la participación en el programa.

En este sentido los objetivos específicos podrían incluir:

- Obtener la percepción de los actores involucrados respecto a los aspectos administrativos del funcionamiento del programa.
- Definir potenciales mejoras en el diseño, planificación, coordinación e implementación para facilitar la aplicación del programa.
- Establecer el nivel de satisfacción de los beneficiarios en conjunto con los beneficios generados por la participación en el programa.
- Determinar las barreras o dificultades que limitan una mayor satisfacción entre los beneficiarios.
- Identificar soluciones respecto a los problemas enfrentados.

En términos de diseño metodológico se propone realizar *focus groups*, los cuales deberían ser guiados por pautas de conversación y realizados con los beneficiarios de cada zona geográfica representativa.

Cada *focus group* debería estar integrado por 6 a 8 personas guiadas por un moderador. La aplicación y el análisis de los resultados de estos instrumentos requieren personal altamente capacitado en técnicas cualitativas como sociólogos, psicólogos o antropólogos. Además, se podría usar algún *software* para hacer los análisis e interpretar los resultados.

La primera ronda de *focus groups* se debería realizar previo al levantamiento de la línea base y previo a la encuesta de seguimiento. El análisis, la interpretación de resultados y conclusiones deberían ser incorporados en el informe de evaluación.

8.11 EVALUACIÓN DE IMPACTO CUANTITATIVA

La evaluación de impacto se debe realizar una vez que se haya levantado al menos una encuesta de seguimiento, tanto para el grupo de tratamiento como para el grupo de control. Como las diferentes encuestas serán iguales en los diferentes momentos del tiempo para ambos grupos, es posible construir una base de datos con estructura de panel que permita utilizar una técnica de experimento aleatorio, variables instrumentales, diferencias en diferencias o técnicas de *matching* con diferencias en diferencias, entre otras.

La encuesta de seguimiento permitirá identificar los impactos finales del programa.

8.12 ESTRUCTURA DEL INFORME FINAL PARA UNA EVALUACIÓN *EX POST*

De acuerdo a Gertler *et al.* (2011) la estructura de un informe de evaluación de impacto debería ser la siguiente:

- Introducción
- Descripción de la intervención
- Objetivos de la evaluación
- Diseño de la evaluación
- Muestreo y datos
- Validación del diseño de la evaluación
- Resultados
- Pruebas de robustez
- Conclusión y recomendaciones

A partir de la estructura anterior, se recomienda utilizar la siguiente propuesta para las evaluaciones *ex post* de políticas o programas ambientales realizados en Chile.

- Introducción
- Descripción del programa.
- Cuantificación de la cobertura del programa considerando la población potencial y beneficiarios efectivos.
- Descripción de la intervención en términos de diseño e implementación, incluyendo una caracterización de las unidades preseleccionadas y elegidas.
- Descripción de los objetivos de la evaluación.
- Desarrollo de hipótesis, teoría del cambio o cadena de resultados.
- Principales indicadores de resultados a evaluar.
- Descripción del diseño de la evaluación en términos teóricos y prácticos.
- Descripción de las estrategias de muestreo y cálculos de potencia a partir de los datos disponibles.
- Evaluar el impacto de los resultados generados por el programa en cada ronda de seguimiento respecto a la línea base.
- Pruebas de robustez de los resultados.
- Recomendar o no la continuación o modificación del programa sustentado en los diversos análisis previos y la medición global de su desempeño.

9 | CASOS DE ESTUDIO SELECCIONADOS

En esta sección se describen tres casos de estudio para ejemplificar la evaluación de impacto a través de *matching*, regresión discontinua y diferencia en diferencia.

9.1 CASO 1: *MATCHING*

Una forma reciente de incentivar la conservación es a través de los llamados "pagos por servicios ecosistémicos", los cuales se han expandido en países en vías de desarrollo fomentados por los acuerdos internacionales que buscan aumentar la captura de CO_2 a través del fomento a la forestación.

En estos casos los "pagos por servicios ecosistémicos" fomentan la forestación o evitan la deforestación al elevar los retornos económicos de la tierra forestada. No obstante, han surgido algunas preocupaciones sobre su efectividad, ya que se podría estar pagando dinero a propietarios de predios que hubiesen mantenido la tierra forestada incluso en ausencia del programa, o bien, por el surgimiento de externalidades negativas al producirse deforestación en otras zonas.

En este contexto, Alix-García *et al.* (2012) analizaron la efectividad y externalidades para el programa nacional de pago por servicios hidrológicos (PSAH) en México. Este programa establece contratos renovables por cinco años que son firmados por los propietarios de los predios para mantener la cobertura forestal a cambio de un pago de 36US$/ha o 27US$/ha dependiendo del tipo de bosque (bosque tropical o bosque semideciduo). El pago se realiza cada año luego de chequear imágenes satelitales de cobertura vegetal. Si se detecta que los predios tienen menor cobertura forestal el pago no se efectúa y son eliminadas del programa. Aproximadamente 2,3 millones de hectáreas entraron al programa entre 2003 y 2009, transformándolo en uno de los programas más grandes del mundo.

Para evaluar el impacto de este programa se requería construir un contrafactual plausible, específicamente para los beneficiarios del año 2004. Por ello, se utilizó como grupo de control a los postulantes del año 2004 que fueron rechazados por aspectos geográficos (cobertura forestal mínima de 80%) o problemas administrativos y también a los beneficiarios del programa pero que postularon el año 2006. Al extraer controles de esta forma se garantizaba que los propietarios de los predios poseían un costo de oportunidad suficientemente bajo para desear postular al programa.

Sin embargo, al comparar las características de los predios participantes del año 2004 respecto a todos los potenciales predios utilizados como controles, se observaron diferencias significativas entre ambos grupos. Específicamente, los predios participantes estaban en áreas con mayor pendiente, elevación, densidad de caminos, densidad población, pobreza y menor proporción de bosque tropical (ver Tabla 2).

Tabla 9-1. Estimaciones del impacto del programa sobre la deforestación

Variable	Participantes	No participantes	Test para diferencia de medias
Área (km^2)	7,04	9,35	2,26
Ln(pendiente)	2,44	2,33	2,12
Elevación promedio (km)	2,09	1,87	3,40
Proporción de área con bosque semi-deciduo	0,20	0,32	4,03
Proporción de área con bosque tropical	0,33	0,26	2,39
Ln (densidad de caminos)	6,64	6,48	3,36
Ln (densidad de población)	3,52	3,16	3,55
Índice de marginalidad municipal	-0,14	-0,26	1,79
Proporción con deforestación	0,22	0,23	0,17
Porcentaje de deforestación	1,41	2,36	1,99
Porcentaje de deforestación| deforestación > 0	6,30	10,28	2,27

Fuente: Alix-García, Shapiro & Sims (2012)

Por lo anterior, en vez de utilizar todos potenciales controles, se escogieron los predios con técnicas de *matching* condicionando por región, tipo de propiedad, pendiente del terreno, tipo de base forestal, tasas de deforestación previa, densidad poblacional, grado de marginalidad y acceso a mercados. Así, se restringió la muestra de predios tratados sólo a aquellos que presentaban razonablemente buenos contrafactuales (90% del total).

Una limitación de los datos de deforestación utilizados en el estudio fue la resolución de las imágenes satelitales, ya que no eran capaces de detectar el lugar exacto de la deforestación dentro de cada pixel (equivalente a seis hectáreas). No obstante, el área mínima para participar en el programa el año 2004 fue 50 hectáreas y el tamaño promedio de los predios que participaron fue 700 hectáreas, por lo cual el error de medición no debería afectar los resultados.

A diferencia de estudios previos que encuentran impactos pequeños o no encuentran efectos para este tipo de programas, el PSAH redujo la probabilidad de deforestación entre 10% a 11%, lo cual representa una disminución de entre 33% a 37% en la probabilidad de tener algún grado de deforestación. Además, el programa redujo el porcentaje del área deforestada entre 1,1% a 1,6%. Esto se puede explicar porque a diferencia de otros países, México aún presenta una tasa significativa de deforestación.

Tabla 9-2. Estimaciones del impacto del programa PSAH sobre la deforestación

	Métrica Mahalonobis		Inverso del error estándar muestral	
Variable dependiente	% deforestado	Deforestación	% deforestado	Deforestación
Efecto del tratamiento	-1,10 ***	-0,10 ***	-1,57 ***	-0,11 ***
Error estándar	(0,35)	(0,03)	(0,44)	(0,03)
Observaciones	633	633	633	633

significancia: *** p < 0,01

Fuente: Alix-García, Shapiro & Sims (2012)

Además, para evaluar las externalidades negativas del programa se compararon las tasas de deforestación para el grupo de tratados y grupo de controles respecto a otros predios no postulantes de los mismos dueños y que eran cercanos al predio que participó en el programa. Los resultados reflejaron un efecto indirecto sobre la deforestación en las localidades más pobres, y además, que la magnitud era relevante al menos en algunas áreas.

Finalmente, a partir de la evaluación *ex post* se observó una heterogenidad importante en los efectos sobre la deforestación evitada, ya que el programa pareció ser más efectivo en las localidades donde la pobreza era menor, así como también, en las regiones del sur y el noreste de México.

9.2 CASO 2: REGRESIÓN DISCONTINUA

Para diseñar políticas que reduzcan efectivamente las emisiones de gases de efecto invernadero es útil conocer los factores que determinan el comportamiento de las empresas. Por ello, el estudio de Martin *et al.* (2012) analizó los "procesos de innovación limpia" que ayudan a reducir las emisiones.

Para ello se recolectaron datos de aproximadamente 800 empresas industriales seleccionadas de forma aleatoria en seis países europeos (Bélgica, Francia, Alemania, Hungría, Polonia y Reino Unido). Las entrevistas con los gerentes fueron vía telefónica, las cuales incluyeron preguntas abiertas y preguntas más específicas, que posteriormente fueron clasificadas por el entrevistador en diversas categorías. Específicamente, se utilizó una escala ordinal de 1 a 5 para medir diversas prácticas relacionadas al cambio climático. El cuestionario aplicado a las empresas estaba dividido en cuatro secciones:

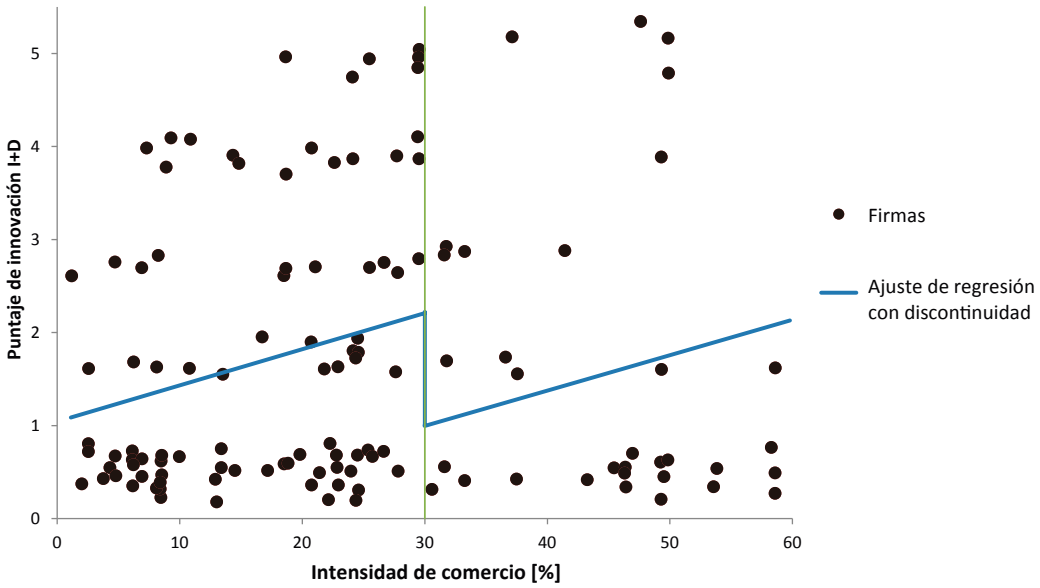
- Efectos actuales y expectativas futuras de los efectos regulatorios del sistema de emisiones transables de la Unión Europea.
- Precios para la energía y el CO2, grado de competencia y otros factores que afectan las prácticas de gestión relacionadas con el cambio climático.
- Medidas específicas que fueron adoptadas y/o descartadas por las empresas.
- Características económicas de las empresas.

Los datos revelaron que el 70% de las empresas estaban comprometidas con "procesos de innovación limpia" cuyo objetivo era reducir las emisiones y/o el consumo de energía. Mientras el 40% de las empresas también desarrollaron "innovación de productos limpios" cuyo objetivo era elaborar productos que pueden ayudar a los consumidores a reducir sus emisiones. Además, se observaron diferencias significativas entre países.

Para investigar la relación entre el sistema de emisiones transables e innovación se aprovechó el hecho que la regulación ambiental había cambiado en diversos periodos. Específicamente, la Fase III del sistema de transacción de emisiones de la Unión Europea generó cambios y reducciones drásticas en el número de permisos que se entregaban gratuitamente a las empresas. Además, que bajo las nuevas propuestas, la Unión Europea dejaría exento de regulación a ciertos sectores industriales basado en indicadores de intensidad del comercio y la intensidad del CO_2. Por ello, los autores notaron que existían empresas que recibieron o no recibieron permisos de emisión de forma gratuita sólo por el valor que tenía su indicador de intensidad de comercio.

En la Figura 9-1 se observa que la regresión para estimar la innovación en las empresas sufre una caída de 30% en el umbral del criterio de intensidad de comercio.

Figura 9-1. Regresión con salto de discontinuidad



Fuente: Martin et al. (2012)

De esta forma realizan estimaciones con regresión discontinua restringiendo la muestra de empresas sólo a aquellas que están localizadas en un intervalo de 10 puntos porcentuales a cada lado del umbral que les permitiría o no recibir permisos de emisión de forma gratuita.

Los resultados muestran que las expectativas sobre exigencias regulatorias más estrictas se explican en parte por los criterios de elegibilidad para obtener permisos de emisiones gratuitos, pero estos mismos factores no son relevantes cuando se trata de explicar la exigencia regulatoria actual. También se detectan impactos negativos y significativos de la exención de regulación ambiental sobre la innovación en productos e innovación promedio. Por lo cual, se concluye que la regulación ambiental tiene un impacto económicamente relevante sobre la innovación.

Tabla 9-3. Estimaciones con regresión discontinua

Ítem	Exigencia Esperada	Exigencia Actual	Innovación en Productos	Innovación en procesos	Innovación promedio
Exención regulatoria	-0,41 **	-0,21 **	-1,14 ***	-1,13	-0,63 **
Intensidad sectorial de CO_2	5,24 ***	0,08	7,75 **	4,57 *	6,16 **
Intensidad de comercio sectorial	1,00	1,09	2,93 **	-1,44	0,75
Intensidad de CO_2 de la empresa	0,10	0,04	-0,01	0,07	0,03
Multinacional	0,25	-0,12	0,39	0,48 *	0,44 **
I+D	0,11	-0,02	0,16	0,22	0,19
Empleo	0,16 **	0,03	0,15 **	0,23 ***	0,19 ***
Competidores fuera de la UE	-0,07	0,02	-0,22	0,47 *	0,12
Observaciones	236	236	236	236	236
R2	0,28	0,17	0,29	0,36	0,35

significancia: *** p =10% / ** p =5% / * p =1%

Fuente: Martin et al. (2012)

Estos resultados son robustos a diferentes especificaciones de la regresión discontinua (cuadrática y ancho de banda). Así, se puede afirmar que la innovación relacionada al cambio climático responde a las exigencias regulatorias del sistema de transacción de emisiones de la Unión Europea. Esto se explica porque tener que pagar por todos los permisos de emisiones requeridos le genera a las empresas un incentivo para comprometerse en estrategias que desarrollen líneas de productos que reduzcan las emisiones asociadas al cambio climático.

9.3 CASO 3: DIFERENCIAS EN DIFERENCIAS

El sistema de transacción de emisiones de la Unión Europea (EU ETS) específica como meta una cantidad acumulada máxima de emisiones de efecto invernadero y permite que las empresas reguladas puedan transar permisos de emisión en el mercado. Así, el precio de estos permisos entrega una señal sobre qué tan valiosas son las actividades de mitigación de emisiones.

A la fecha del estudio el EU ETS estaba divido en dos fases. La fase I (2005-2007) y la fase II (2008-2012). La asignación inicial de permisos y las reglas de comercio fueron significativamente distintas en ambos periodos, por ejemplo, en la segunda etapa fue permitido el préstamo y el almacenamiento de emisiones previamente abatidas (banking). Además, en la segunda fase la cantidad de permisos fue reducido en 11% y las emisiones verificadas excedieron los permisos asignados en 2,9%. Por esta razón, la falta de permisos subió sus precios a aproximadamente 20 euros, pero luego debido a la crisis económica de 2009 los precios cayeron a 15 euros.

Así, el desafío del estudio fue identificar la reducción de emisiones asociadas al EU ETS, al comparar las emisiones observadas respecto a las emisiones que hubiera generado cada empresa en ausencia del programa, porque este contrafactual no es observable. Sin embargo, los autores utilizan una base de datos de tipo panel para estimar el cambio entre la primera y la segunda fase del EU ETS.

El estudio utiliza datos para las emisiones de las empresas europeas reguladas (base de datos CITL) y a partir de la dirección de la empresa se combina con otros datos (base de datos AMADEUS) que incluyen empleo, margen de comercialización, valor agregado, trabajo y costos fijos del capital de cada firma. Así, se pudo construir una base de datos consolidada que incluye 2101 empresas que representan el 59% de las emisiones totales.

Con los datos a nivel de empresa se estimó un factor de asignación (AF), definido como la razón entre la asignación de emisiones y las emisiones verificadas. Un AF > 1 refleja que la empresa recibió un exceso de permisos de emisiones pudiendo vender una parte y un AF < 1 significa que recibió muy pocos permisos por lo cual debe comprar más para cumplir con el EU ETS. También, se incluyeron otros factores que pudiesen haber inducido cambios en las emisiones como el ambiente económico que pudo llevar una reducción en la producción de las empresas.

Para averiguar si existió una aceleración en la reducción de las emisiones en la segunda fase de EU ETS los autores parten del siguiente modelo:

$$y_{it} = \alpha_0 + \alpha_1 \cdot d_{it} + \alpha_2 \cdot cv^1_{it} + \alpha_3 \cdot cv^2_{it} + e_{it}$$

Donde, i representa a la empresa, t es el año (2005, 2006, 2007 ó 2008), y_{it} es el logaritmo natural de las emisiones verificadas, d_{it} es una variable *dummy* del tiempo, cv^1_{it} es un conjunto de variables de control que incluye el logaritmo natural del volumen de negocios y el empleo, cv^2_{it} es un segundo conjunto de variables de control que incluye el sector económico y país, y e_{it} es un error aleatorio.

Los resultados muestran una relación positiva significativa entre cambios en la producción y cambios en las emisiones. Además, la reducción en las emisiones es significativa al cambio de fase, luego de controlar por factores económicos (volumen de producción, número de empleados, sector y país), por lo cual es posible concluir que la reducción que se observó entre 2007 y 2008 se debió al cambio desde la fase I a la fase II del EU ETS.

Tabla 9-4. Diferencial en la tasa de crecimiento de las emisiones 2005-2006 vs 2007-2008

Ítem	Todas las empresas	Empresas con poca asignación inicial (AF<1,15)	Empresas con alta asignación inicial (AF>1,15)
Cambio de fase	-0,04 **	-0,03**	0,02
Cambio en volumen producido	0,19 ***	0,19***	0,21***
Cambio en N° de empleados	0,00	-0,03	0,07
R^2 ajustado	0,17	0,21	0,23

significancia: *** p =10% / ** p =5% / * p =1%

Fuente: Abrell, Ndove-Faye & Zachmann (2011)

Las empresas que obtuvieron menos permisos en relación a sus emisiones reales mostraron una conducta de mitigación, mientras que las empresas que recibieron relativamente más permisos iniciales no incrementaron sus esfuerzos por reducir emisiones entre las dos fases del EU ETS. También, la respuesta al cambio de fase fue diferente entre sectores económicos. Algunos sectores como metálica básica y minerales no metálicos, incrementaron significativamente sus esfuerzos de reducción de emisiones, y otros sectores como electricidad y calefacción no lo hicieron. Finalmente, los autores encuentran que la asignación inicial de permisos y las emisiones *ex post* estaban correlacionadas, lo cual permitiría concluir que el mercado de permisos se desvía de las condiciones competitivas ideales asumidas en el teorema de Coase.

10 | REFERENCIAS BIBLIOGRÁFICAS

Abadie, A. (2002). Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. Journal of the American Statistical Association, American Statistical Association, 97, 284-292.

Abdul-Manan, A., Baharuddin, A., & Chang, L. (2015). Ex-Post Critical Evaluations of Energy Policies in Malaysia from 1970 to 2010: A Historical Institutionalism Perspective. Energies, 8, 1936-1957.

Abrell, J., Ndoye-Faye, A., & Zachmann, G. (2011). Assessing the impact of the EU ETS using firm level data. Bruegel Working Paper 2011/08.

Agnolucci, P. (2004). Ex post evaluations of CO2 –based taxes: a survey,. Tyndall Centre for Climate Change Research Working Paper 52.

Aichele, R., & Felbermayr, G. (2011). What a Difference Kyoto Made: Evidence from Instrumental Variables Estimation. Ifo Working Paper No. 102 Institue for Economic Research University of Munich.

Alix-Garcia, J., N., S. E., & Sims, K. R. (2010). The environmental effectiveness of payments for ecosystem services in Mexico: preliminary lessons for REDD. First paper draft.

Alix-Garcia, J., Shapiro, E. N., & Sims, K. R. (2012). Forest Conservation and Slippage: Evidence from Mexico's National Payments for Ecosystem Services Program. Land Economics, 88, 613-638.

Andam, K., Ferraro, P., Pfaff, A., Sanchez-Azofeifa, G., & Robalino, J. A. (2008). Measuring the effectiveness of protected area networks in reducing deforestation. Proceedings of the National Academy of Sciences, 105(42), 16089–16094.

Andam, K., Ferraro, P., Sims, K., Healy, A., & Holland, M. (2010). Protected areas reduced poverty in Costa Rica and Thailand. Proceedings of the National Academy of Sciences of the USA, 107(22), 9996-10001.

Anger, N., & Oberndorfer, U. (2008). Firm performance and employment in the EU emissions trading scheme: An empirical assessment for Germany . Energy Policy, 36(1), 12-22.

Angrist, J. (2004). Treatment effect heterogeneity in theory and practice. The Economic Journal, 114, C52–C83.

Angrist, J., & Krueger, A. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? Quarterly Journal of Economics, 106(4), 979-1014.

Antweiler, W., Copeland, B., & Taylor, M. S. (2001). Is Free Trade Good for the Environment? The American Economic Review, 91(4). 877-908.

Arano, K., & Blair, B. (2008). An ex-post welfare analysis of natural gas regulation in the industrial sector. Energy Economics, 30, 789-806.

Arriagada, R. (2008). Private Provision of Public Goods: Applying Matching Methods to Evaluate Payments for Ecosystem Services in Costa Rica. Ph. D. dissertation, Graduate Faculty of North Carolina State University.

Arriagada, R., E. Sills, E., Pattanayak, S. K., & Ferraro, P. J. (2008). Private landowners, public payments, and forest cover in Costa Rica: evaluating the impact of payments for ecosystem services. Working paper presented at the annual meetings of the European Association of Environmental and Resource Economics, Gotenburg, Sweden.

Arriagada, R., Ferraro, P., Sills, E., Pattanayak, S., & Cordero, S. (2012). Do payments for environmental services reduce deforestation? A farm level evaluation from Costa Rica. Land Economics, 88(2), 382-399.

Arriagada, R., Sills, E., Pattanayak, S., & Ferraro, P. (2009). Combining qualitative and quantitative methods to evaluate participation in Costa Rica's Program of Payments for Environmental Services. Journal of Sustainable Forestry, 28(3-5), 343-367.

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. Review of Economics and Statistics, 60, 47-57.

Athey, S., & Imbens, G. (2006). Identification and Inference in Nonlinear Difference-in-Differences Models. Econometrica, 74(2), 431-497.

Bamberger, M., Rao, V., & Woolcock, M. (2010). Using mixed methods in monitoring and evaluation: experiences from international development. Policy Research Working Paper Series 5245, The World Bank.

Banerjee, A., Duflo, E., Cole, S., & Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. Quarterly Journal of Economics.

Bennett, J. (2011). Ex-post environmental impact assessment: lessons from four CGIAR case studies. CGIAR Independent Science and Partnership Council, BRIEF NUMBER 39.

Berkhouta, P., Ferrer-i-Carbonell, A., & Muskens, A. (2004). The ex post impact of an energy tax on household energy demand. Energy Economics, 26, 297-317.

Bjorner, T., & Jensen, H. (2002). Energy Taxes, Voluntary Agreements and Investment Subsidies – a Micro-panel Analysis of the Effect on Danish Industrial Companies' Energy Demand. Resource and Energy Economics, 24, 229-249.

Blackman, A. (2012). Ex Post Evaluation of Forest Conservation Policies Using Remote Sensing Data. An Introduction and Practical Guide. Environment for Development, Discussion Paper Series EfD DP 12-05.

Blackman, A., & R. Woodward, R. (2010). User financing in a national payments for environmental services program: Costa Rican hydropower. Ecological Economics, 69(8), 1626-1638.

Blackman, A., Pfaff, A., & Robalino, J. (2015). Paper Park Performance: Mexico's Natural Protected Areas in the 1990s. Global Environmental Change, 31, 50-61.

Blasco, J., & Casado, D. (2009). Evaluación del impacto. Barcelona: Ivàlua, (Guías prácticas sobre evaluación de políticas públicas; 5).

Blundell, R., & Costa-Dias, M. (2009). Alternative approaches to evaluation in empirical microeconomics. Journal of Human Resources, 44, 565-640.

Brouhle, K., & Ramirez-Harrington, D. (2010). GHG Registries: Participation and Performance Under the Canadian Voluntary Climate Challenge Program. Environmental & Resources Economics, 47, 521–548.

Bushnell, J., Chong, H., & Mansur, E. (2013). Profiting from Regulation: Evidence from the European Carbon Market. Economic Policy, 5(4), 78-106.

Calfucura, E., & Montero, R. (2013). Programas de Descontaminación y Calidad del Aire en América Latina: Una Evaluación Ex-Post para el Caso de Santiago de Chile. Working Papers 47, Facultad de Economía y Empresa, Universidad Diego Portales.

Caliendo, M. & Kopeinig, S. (2008). Some practical guidance for the implementation of *propensity score matching*. Journal of Economic Surveys, 22(1), 31-72.

Carley, S. (2009). State renewable energy electricity policies: An empirical evaluation of effectiveness. Energy Policy, 37(8), 3071-3081.

Carneiro, P., Heckman, J., & Vytlacil, E. (2010). "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin. Econometrica, 78(1), 377-394.

Caro, J., Melo, O., & Forester, W. (2006). Participación e Impacto del Programa de Recuperación de Suelos Degradados en Usuarios de INDAP. Economía Agraria, 10, 11-24.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New Jersey: Lawrence Erlbaum.

CONAMA. (1998). Una política ambiental para el desarrollo sostenible. Santiago.

Consultorias Profesionales Agraria. (2005). Programa Bonificación Forestal DL 701. Informe final. Ministerio de Agricultura, CONAF.

Crabbé, A., & Leroy, P. (2008). The Handbook of Environmental Policy Evaluation. London: Earthscan.

Dean, J. (2002). Does trade liberalization harm the environment? A new test. Canadian Journal of Economics, 35(4), 819-842.

Dehejia, R., & Wahba, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. Journal of the American Statistical Association, 94(448), 1053-1062.

Duflo, E., & Hanna, R. (2006). Monitoring Works: Getting Teachers to Come to School. NBER Working Paper No. 11880.

Edmonds, E. (2002). Government Initiated Community Resource Management and Local Resource Extraction from Nepal's Forests. Journal of Development Economics, 68(1), 89-115.

EMG (2002). Estudio del Impacto del Sistema de Incentivos para la Recuperación de Suelos Degradados. Informe Final.

Fafchamps, M., & Minten, B. (2012). Impact of SMS-Based Agricultural Information on Indian Farmers. The World Bank Economic Review.

Fahrenkrog, G., Tubke, A., Polt, W., & Rojo, J. (2002). Avenues for RTD Evaluation in the future policy context. *En* Fahrenkrog, G., Polt, W., Rojo, J., Tubke, A. & Znöker, K., RTD Evaluation Toolbox. Assessing the Socio- Economic Impact of RTD-Policies (p. 243-247). Seville, European Commission-Join.

Ferraro, P., & Hanauer, M. (2011). Protecting Ecosystems and Alleviating Poverty with Parks and Reserves: 'Win-Win' or Tradeoffs? Environmental & Resources Economics, 48, 269-286.

Feser, E. (2013). Isserman's impact: quasi-experimental comparison group designs in regional research. International Regional Science Review, 36(1), 44-68.

Finn, J., Bartolini, F., Bourke D., Kurz, I. & Viaggi, D. (2009). *Ex post* environmental evaluation of agri-environment schemes using experts' judgements and mulicriteria analysis. Journal of Environmental Planning and Management, 52, 717-737.

Frankel, J., & Rose, A. K. (2005). Is Trade Good or Bad for the Environment? Sorting Out the Causality. Review of Economics and Statistics, 87(1), 85-91.

Frondel, M., & Vance, C. (2013). Fuel Taxes versus Efficiency Standards – An Instrumental Variable Approach. Ruhr Economic Papers #445. Ruhr-Universität Bochum (RUB), Department of Economics.

Gaveau, D., Epting, J., Lyne, O., Linkie, M., Kumara, I., Kanninen, M., y otros. (2009). Evaluating whether protected areas reduce tropical deforestation in Sumatra. Journal of Biogeography, 36 , 2165-2175.

Gelo, D., Koch, S., & Muchapondwa, E. (2013). Do the Poor Benefit from Devolution Policies? Evidences from Quantile Treatment Effect Evaluation of Joint Forest Management. Working Papers 400, Economic Research Southern Africa.

Gertler, P., Martinez, S., Premand, P., Rawlings, L., & Vermeersch, C. (2011). La Evaluación de Impacto en la Práctica. Banco Mundial.

Görlach, B. I., Newcombe, J., & Johns, H. (2005). Cost-effectiveness of environmental policies. Informe Final.

Green Alliance. (2002). Next Steps for Energy Taxation – a Survey of Business Views. Green Alliance, London.

Greenstone, M. (2004). Did the Clean Air Act cause the remarkable decline in sulfur dioxide concentrations? Journal of Environmental Economics and Management, 47(3), 585-611.

Hammitt, J. (2000). Are the Costs of Proposed Environmental Regulations Overestimated? Evidence from the CFC Phaseout. Environmental and Resource Economics, 16, 281–301.

Harbaugh, W., Levinson, A., & Wilson, D. M. (2002). Reexamining the Empirical Evidence for an Environmental Kuznets Curve. Review of Economics and Statistics, 84 (3), 541-551.

Heckman, J. (1979). Sample selection bias as a specification error. Econometrica 47, 153-161.

Heckman, J., & Navarro-Lozano, S. (2004). Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models. Review of Economics and Statistics, 86, 30-57.

Heckman, J., & Vytlacil, E. (1999). Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects. Proceedings of the National Academy of Sciences, 96(8), 4730-4734.

Heckman, J., & Vytlacil, E. (2001). Local Instrumental Variables. *En* C. Hsiao, K. Morimune, & J. L. Powell, Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya (p. 1-46). New York: Cambridge University Press.

Heckman, J., & Vytlacil, E. (2007). Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Economic Estimators to Evaluate Social Programs and to Forecast Their Effects in New Environments. *En* J. Heckman, & E. Leamer, Handbook of Econometrics (p. 4875-5144). Amsterdam: Elsevier.

Heckman, J., Ichimura, H., & Todd, P. (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program. Review of Economic Studies, 64, 605-654.

Hirano, K., & Imbens, G. (2004). The propensity score with continuous treatments. En A. Gelman, & X. Meng, Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives (p. 73-84). New York : Wiley.

Honey-Roses, J., Baylis, K., & Ramirez, I. (2011). Do our conservation programs work? A spatially explicit estimator of avoided forest loss. Conservation Biology, 25(5), 1032-1043.

Ichino, A., Mealli, F., & Nannicini, T. (2006). From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity. Discussion Paper No. 2149, IZA, Bonn.

Imbens, G. (2000). The role of the propensity score in estimating dose-response functions. Biometrika, 87(3), 706-710.

Imbens, G., & Angrist, J. (1994). Identification and estimation of local average treatment effects. Econometrica, 61, 467-476.

Imbens, G., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. Journal of Econometrics, 142(2), 615-635.

Imbens, G., & Wooldridge, J. (2009). Recent Developments in the Econometrics of Program Evaluation. Journal of Economic Literature, 47(1), 5-86.

Jeffords, C., & Minkler, L. (2014). Do Constitutions Matter? The Effects of Constitutional Environmental Rights Provisions on Environmental Outcomes. Working papers 2014-16, University of Connecticut, Department of Economics.

Johannsen, K., & Togeby, M. (1998). The Danish CO2 Tax on Trade and Industry. Inter-SEE: Interdisciplinary Analysis of Successful Implementation of Energy Efficiency in the Industrial, Commercial and Service Sector, VOL. II., Contract JOS3-CT95-0009. JOULE III.

Joppa, L., & Pfaff, A. (2009). High and Far: Biases in the Location of Protected Areas. PLoS ONE, 4(12): e8273.

Kelly, A., Lumbreras, J., Maas, R., Pignatelli, T., Ferreira, F., & Engleryd, A. (2010). Setting national emission ceilings for air pollutants: policy lessons from an ex-post evaluation of the Gothenburg Protocol. Environmental Science & Policy, 13, 28-41.

King, A., & Lenox, M. (2001). Does It Really Pay to Be Green? An Empirical Study of Firm Environmental and Financial Performance. Journal of Industrial Ecology, 5(1), 105-116.

Kotchen, M., Moore, M., Lup, F., & Rutherford, E. (2006). Environmental Constraints on Hydropower: An Ex Post Benefit-Cost Analysis of Dam Relicensing in Michigan. Land Economics, 82 (3), 384-403.

Larsen, B., & Nesbakken, R. (1997). Norwegian Emissions of CO2 1997 - 1994. Environmental and Resource Economics, 9, 275-290.

Lee, D. (2008). Randomized experiments from non-random selection in U.S. House elections. Journal of Econometrics, 142, 675-697.

Leidner, A. (2014). Estimating the Effectiveness of Health-Risk Communications with Propensity-Score Matching: Application to Arsenic Groundwater Contamination in Four US Locations. Journal of Environmental and Public Health, Volume 2014.

Lin, B., & Li, X. (2011). The effect of carbon tax on per capita CO2 emissions, . Energy Policy, 39, 5137-5146.

Lin, C., & Liscow, Z. D. (2013). Endogeneity in the Environmental Kuznets Curve: An Instrumental Variables Approach. American Journal of Agricultural Economics, 95 (2), 268-274.

Lucas, R., Wheele, D., & Hettige, H. (1992). Economic Development, Environmental Regulation and the International Migration of Toxic Industrial Pollution: 1960-88, International Trade and the Environment. World Bank Discussion Paper No. l59.

Malaska, P., Luukkanen, J., Vehmas, J., & Kaivo-oja, J. (1997). Environment-based Energy Taxation in the Nordic Countries. Ministry of the Environment, Helsinki.

Mardones, C. (2016). Ex - post Evaluation of Green Community Initiatives. Working Paper, Department of Industrial Engineering, University of Concepción.

Martin, R., Muûls, M. d., & Wagner, U. (2012). Anatomy of a Paradox: Management Practices, Organisational Structure and Energy Efficiency. Journal of Environmental Economics and Management, 63(2), 208-223.

Martin, R., Muûls, M., & Wagner, U. (2012). An evidence review of the eu emissions trading system, focussing on effectiveness of the system in driving industrial abatement. Informe final. Departamento para Energía y Cambio Climático de la Unión Europea.

Martin, R., Muûls, M., & Wagner, U. (2012). Carbon Markets, Carbon Prices and Innovation: Evidence from Interviews with Managers. Draft preliminary.

Miguel, E., & Kremer, M. (2003). Networks, Social Learning, and Technology Adoption: The Case of Deworming Drugs in Kenya. Working Paper 61.

Ministerio de Desarrollo Social. (2015). Recuperado el Diciembre de 2015, de http://sni.ministeriodesarrollosocial.gob.cl/evaluacion/ex-post/

Morgan, D., Ozanne-Smith, J., & Triggs, T. (2009). Self-reported water and drowning risk exposure at surf beaches. Australian New Zealand Journal of Public Health, 33(2), 180-188.

Morley, B. (2010). Empirical Evidence on the Effectiveness of Environmental Taxes. Department of Economics University of Bath.

Mullins, J., & Bharadwaj, P. (2015). Effects of Short-Term Measures to Curb Air Pollution: Evidence from Santiago, Chile. American Journal of Agricultural Economics, 97, 1107-1134.

Naciones Unidas. (1987). Brundland Report. Oxford University Press.

Nelson, A., & Chomitz, K. (2011). Effectiveness of Strict vs. Multiple Use Protected Areas in Reducing Tropical Forest Fires: A Global Analysis Using Matching Methods. PLoS ONE, 6(8): e22722.

NUTEK. (1994). Utvardering av Styrmedel och Stod for Begransning av Koldioxidutslapp i Sverige. Informe.

OECD. (1997). Evaluating Economic Instruments for Environmental Policy. Paris.

Olken, B. (2007). Monitoring Corruption: Evidence from a Field Experiment in Indonesia. Journal of Political Economy, 115(2), 200-249.

Pfaff, A., Robalino, J., & Sanchez-Azofiefa, G. (2008). Payments for environmental services: Empirical analysis for Costa Rica. Terry Sanford Institute of Public Policy Working Paper Series SAN08-05.

Pilavachi, P., Dalamagaa, T., Rossetti di Valdalberob, D., & Guilmot, J. (2008). Ex-post evaluation of European energy models. Energy Policy, 36, 1726-1735.

Robalino, J., Pfaff, A., Sanchez-Azofeifa, G., Alpízar, F., León, C., & Rodríguez, C. (2009). Changing the deforestation impacts of ecopayments: Evaluation (2000–2005) in Costa Rica's PSA program. IOP Conference Series Earth and Environmental Sci.

Rogan, F., Dennehy, E., Daly, H., Howley, ,. M., & Gallachóir, B. (2011). Impacts of an emission based private car taxation policy – First year ex-post analysis. Transportation Research Part A, 45, 583-597.

Rosenbaum, R., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41-55.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). Evaluation: A systematic approach. Thousand Oaks, CA: Sage.

Schroeder, L. (2012). Assessing farmers' acceptance and perception of agri-environment schemes by ex-post application of the 'Theory of Planned Behaviour' - A case study in England. Paper prepared for the 126th EAAE Seminar.

Schultz, P. (2001). School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program. Working Papers 834, Economic Growth Center, Yale University.

Shah, P., & Baylis, K. (2015). Evaluating Heterogeneous Conservation Effects of Forest Protection in Indonesia. PLoS ONE, 10(6): e0124872.

Shopley, & Brasseur. (1996). Competitiveness and Employment: Summary Report on Two Microeconomic Case Studies in the Field of Business and the Environment. citado en Agnolucci (2004).

Shopley, & Brasseur. (1996). Competitiveness and Employment: Summary Report on Two Microeconomic Case Studies in the Field of Business and the Environment. citado en Agnolucci (2004).

Sims, K. (2010). Conservation and development: Evidence from Thai protected areas. Journal of Environmental Economics and Management, 60, 94–114.

Soukopová, J., & Bakos, E. (2013). Environmental protection expenditure: Ex–post evaluation. Masaryk University. Working paper WP KVE 08/2013.

Standard & Poor's, D., & Leuven, K. (1999). The Auto-Oil II Cost-Effectiveness Study. European Commission.

Stock, J., & Watson, M. (2003). Introduction to Econometrics. Pearson Education, International Edition.

Stock, J., & Yogo, M. (2005). Testing for Weak Instruments in IV Regression. *En* D. W. Stock, Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas Rothenberg (p. 80-108). Cambridge University Press.

Tanaka, S. (2015). Environmental regulations on air pollution in China and their impacton infant mortality. Journal of Health Economics, 42, 90–103.

Tashakkori, A., & Teddlie, C. (2010). Current developments and emerging trends in integrated research methodology. *En* A. Tashakkori, & C. Teddlie, Handbook of mixed methods in social and behavioral research (2nd Edition). Thousand Oaks, CA: Sage.

Technopolis. (2009). Ex-post Impact Assessment. FP6 sub-priority "Global Change and Ecosystems". Informe Final para la Comisión Europea.

Uchida, E., Rozelle, S., & Xu, J. (2009). Conservation Payments, Liquidity Constraints and Off-Farm Labor: Impact of the Grain for Green Program on Rural Households in China. American Journal of Agricultural Economics, 91(1), 70-86.

Universidad Diego Portales. (2010). Evaluación de Impacto Iniciativas Ambientales Ciudadanas Ministerio del Medio Ambiente. Informe Final.

Vásquez, M., Ferreira, M., Mogollón, A., Fernández, M., Delgado, M. & Vargas, I. (2006). Introducción a las técnicas cualitativas en salud. Cursos GRAAL5. Universitat Autónoma de Barcelona. Servei de Publicacions, Bellaterra 2006.

Webber, P., Gouldson, A., & Kerr, N. (2015). The impacts of house hold retrofit and domestic energy efficiency schemes: A large scale, ex post evaluation. Energy Policy, 84, 35-43.

Guía Metodológica para la evaluación *ex post* de programas y normativa ambiental

Santiago, 2017

EQUIPO DE TRABAJO

ASESORÍA EXPERTA

Cristian Mardones

MINISTERIO DEL MEDIO AMBIENTE

Francisco Donoso
Isabel Rojas
Ixsy Valdés
Sandra Briceño
Rodrigo Pizarro

DISEÑO Y DIAGRAMACIÓN

Francisca Villalón