UNIVERSIDAD
DE GRANADA

ANÁLISIS CAUSAL DE ESTUDIOS OBSERVACIONALES: APLICACIÓN INDUSTRIAL

DISERTACIÓN DOCTORAL

Presentado para obtener el título de

DOCTOR (*Doctor of Philosophy Degree*)

PROGRAMA DE DOCTORADO EN TECNOLOGÍA DE LA INFORMACIÓN Y COMUNICACIÓN

Presentado por

RICHARD ALBERTO AVILÉS LÓPEZ

Director

Juan de Dios Luna del Castillo

Granada, Septiembre 2024

A Lorena y María Faustina

Agradecimientos

Al concluir este trabajo de investigación quiero expresar mi profundo agradecimiento a mi amada esposa, *Lorena* y a mi querida hija *María Faustina* por todo el apoyo y confianza brindado a lo largo de estos años, por su paciencia y cariño que han sabido manifestarme en los momentos más difíciles de incertidumbre o desventura, y por haber soportado mi ausencia presencial inmerso en los entresijos de esta investigación.

Quiero empezar agradeciendo a mi querido director Don *Juan de Dios Luna del Castillo*, que tuvo la deferencia de dirigirme en un tema muy desafiante y de mucha actualidad. Su confianza, paciencia, esfuerzo, apoyo y disponibilidad fue fundamental para llegar a buen término, en especial su trato respetuoso, amable, cordial y profesional, en clave "filial". Sus comentarios, llenos de ciencia y experiencia por acercarme más a la verdad, desencadenaban consejos y enseñanzas de altísimo nivel que iluminaban diferentes perspectivas para comprender y explorar el análisis causal del fenómeno investigado.

También agradezco profundamente todo el soporte brindado por mi tutor Don *Juan Manuel Fernández Luna,* que siempre estuvo disponible para cualquier inquietud o comentario, con amabilidad y cordialidad, mostrándome su confianza y profesionalidad en todo momento.

Además, mi eterno agradecimiento a Don *José Balderas Cejudo*, Jefe de Servicio de la Escuela Internacional de Posgrado por comprender, orientar y brindar el soporte en todos los trámites que requería para continuar mis estudios doctorales. De igual manera, a quienes dirigen el Programa de Doctorado en Tecnología de la Información y Comunicación por su atención ágil y eficaz.

Agradezco de forma especial a mi cuñada *Tania Abadíe* por su ayuda desinteresada e incondicional, y por compartir momentos de alegría y felicidad con mi familia.

Son muchos mis amigos que tanto quiero y aprecio, pues me animaban constantemente a continuar y perseverar en este trabajo de investigación. A ellos les manifiesto mi profundo y sincero agradecimiento por todo su apoyo, su confianza alentadora y sus comentarios oportunos y acertados. Entre ellos destacan *Rafael Salazar Vera, Miguel Ángel Pardillo, Ricardo Gámez Valdez, Tomás Sola Martínez, Luis Pérez Loor y Tomás García Hernández*. También destaco la ayuda de *José Luis* por brindarme sus comentarios en la revisión final del presente trabajo.

Finalmente, extiendo mi sincero agradecimiento a todas las personas que me han orientado y animado en todo momento y que por motivos de espacio no son nombradas o que han preferido mantenerse en el anonimato. A todos los llevaré eternamente en mi corazón.

ANÁLISIS CAUSAL DE ESTUDIOS OBSERVACIONALES: APLICACIÓN INDUSTRIAL

Resumen

La creciente demanda de empaques de cartón, favorecida por al aumento del comercio electrónico y las exportaciones post pandemia, exigen un mayor control de los impactos ambientales, energéticos y económicos de sus procesos productivos. El sistema de producción de este tipo de industria se compone principalmente de dos subsistemas: corrugado e impresión. El primero elabora láminas de cartón y el segundo las imprime, recorta, engoma, dobla, apila y paletiza. La principal fuente de desperdicio se genera en el subsistema de impresión compuesto por varias líneas de producción (máquinas), y cada línea dispuesta en varias etapas secuenciales según los módulos concatenados que determinan sus características y potencialidades. Al procesar un pedido, la máquina tiene una parada inicial llamada *setup,* y después de iniciada la corrida de producción puede tener cero o más paradas programadas o no programadas. Las unidades de desperdicio de una orden de producción se cuentan al finalizar su elaboración, ya que no es práctico hacerlo en cada parada porque involucra tiempo de inactividad de la máquina y, por tanto, no se disponen de datos de desperdicio por parada.

La estrategia para disminuir el desperdicio se realiza generalmente con técnicas de Mantenimiento Productivo Total, aunque tendría mayor impacto si se determinara qué máquinas y en qué intervalos de tiempo de parada se generan; no obstante, esta es una tarea compleja por el nivel de ruido que existe en una industria. Sin embargo, si fuera significativo el desperdicio, es muy probable que la máquina se detenga y el tiempo de parada sea mayor. Por tanto, si medimos el tamaño del efecto del tiempo de paradas (posteriores al *setup*) sobre el desperdicio en cada máquina, podríamos examinar con el experto las intervenciones a realizar según los mecanismos que gobiernan su generación. Esto requiere investigar cuál es el tamaño del efecto causal del tiempo total de parada (posteriores al *setup*) sobre el desperdicio.

Para abordar este desafío, primero se estudian los conceptos fundamentales de Inferencia Causal, los Métodos de Evaluación de Efectos Causales, las Teorías Causales, y el Modelamiento Causal. Además, se sintetizan los aportes de los principales investigadores para elaborar una estructura del Diseño de Estudios Observacionales, desde la perspectiva de los métodos de matching, que servirá de guía para la presente investigación.

Tradicionalmente, la estrategia sugerida por los investigadores es generar varios modelos de matching hasta encontrar el que presente mejor equilibrio y, hasta donde sabemos, sin aportar criterios generales sino específico enfocados a cada problema. Por esta razón, se incorpora dentro del Diseño de Estudio Observacional la metodología *Exploratory Model Matching Search Algorithm* (EMMSA) para realizar de forma exploratoria una búsqueda de los mejores modelos de matching. EMMSA está compuesto por dos algoritmos: Exploratory Matching Algorithm (EMA) y Homologous Matching Search Algorithm (HMSA). El primero ejecuta los métodos de matching de forma exploratoria y el segundo realiza la búsqueda de los mejores modelos de matching aportando criterios generales. HMSA establece dos condiciones, la primera condición buscará los modelos con todas las variables equilibradas o hasta un 20% de variables desequilibradas; y la segunda condición buscará los modelos que descarten hasta un l0% de unidades. Además, establece una tercera condición para seleccionar los modelos con tamaño de efectos más grandes, por encima del 95% de la media de los modelos encontrados en la respectiva máquina-tratamiento.

Los resultados muestran que, de los 7691 modelos generados para las siete máquinas en diez tratamientos, solo el 34.38% (2644) de los modelos cumplían las dos primeras condiciones, y el 9,40% (723) cumplía también con la tercera condición, a los cuales se les llamó modelos homólogos por presentar características similares de equilibrio.

Después del Análisis de Sensibilidad se seleccionaron los 70 mejores modelos, uno por cada máquina-tratamiento. El 95.71% de los modelos tenían todas sus covariables equilibradas y ninguna unidad de tratamiento perdida. Además, el 82.86% de estos modelos eran más robustos ($\Gamma > 2$) a los cuales se les puede dar una interpretación causal de sus efectos; el 10% ($1.1 \leq \Gamma \leq 2$) de los modelos eran menos robustos y su interpretación causal es menos defendible, a excepción de los modelos cercanos o iguales a 2; y solo el 7.14% de modelos no equilibran los confusores ocultos ($\Gamma = 1$). Los métodos de matching que encontraron los mejores modelos son Nearest, Genetic y Optimal, en combinación con las distancias Mahalanobis, GBM, y Scaled Euclidean. Finalmente, se realizó un meta-análisis para los niveles de abstracción meso (máquina) y macro (subsistema de impresión) con el fin de estimar una medida del efecto causal que integre varios tratamientos por máquina, o máquinas para el subsistema de impresión, respectivamente.

Las implicaciones prácticas del presente estudio son: la identificación de las máquinas y tratamientos en donde se producen los mayores desperdicios, y el establecimiento de prioridades para su reducción.

Las principales aportaciones de esta Tesis son:

i. La síntesis de una metodología esquemática del Diseño de Estudios Observacionales en el contexto de los métodos de matching, a partir de los aportes de los principales investigadores en este campo.
ii. La concepción sistémica del estudio para facilitar la organización de las variables y la extracción de ocho criterios de selección de variables para el proceso de matching.
iii. La construcción de la metodología EMMSA con sus algoritmos EMA y HMSA, pudiendo utilizarse EMA como una plataforma de generación de modelos de matching y HMSA como un método de selección de modelos.
iv. El uso de meta-análisis para integrar los resultados de los diez tratamientos de una máquina y obtener la medida del efecto a nivel meso; e integrar los efectos de las siete máquinas para obtener una medida del efecto a nivel macro.

Las ventajas de esta metodología es que tiene una alta probabilidad de encontrar modelos equilibrados y sesgo reducido, para los datos reales de una industria cartonera. Además, al utilizarse criterios generales y restrictivos de selección de modelos, EMMSA puede aplicarse en otras industrias o en otras áreas de conocimiento. Además, EMA puede incluir nuevos métodos de matching o distancias para la generación de modelos. HMSA puede incorporar nuevas condiciones para mejorar la búsqueda exploratoria de modelos, así como considerar nuevos test de sensibilidad para la obtención de modelos más robustos. Las principales limitaciones de EMMSA es la necesidad de mejorar la parametrización de los métodos Fullmatch y CEM; y reducirse el tiempo de ejecución de EMA.

Esta tesis es una investigación seminal y puede derivarse hacia múltiples líneas de investigación:

i. Métodos de selección de intervalos de tratamiento para la variable causal;
ii. Funciones de transformación alternativas a Propensity Scores o Mahalanobis;
iii. Métodos de selección de covariables;
iv. Ampliación de la librería EMMSA para nuevos métodos de matching y técnicas de análisis de sensibilidad, así como la inclusión de otras estrategias de ajuste de confusores para su comparación.
v. Desarrollo de modelos paramétricos para la estimación del efecto causal en la región de soporte común.

Keywords

inferencia causal; análisis causal; causalidad; EMMSA; EMA; HMSA; algoritmo de búsqueda exploratoria de modelos de matching; algoritmo de matching exploratorio; algoritmo de búsqueda de modelos homólogos; matching

Contenido

Lista de Figuras

Lista de Tablas

Lista de Abreviaturas

Capítulo 2

Sección 2.1

ATE = Average Treatment Effect.
ATT = Average Treatment Effect on the Treated group.
CATE = Conditional Average Treatment Effect.
SUTVA = Stable Unit Treatment Value Assumption.

Sección 2.2

BART = Bayesian Additive Regression Tree.
CART = Classification And Regression Tree.
CTAM = Conditional Treatment-Adversarial learning based Matching.
i.i.d = independent and identically distributed.
IPTW = Inverse Probability of Treatment Weighting.
IPW = Inverse Propensity Weighting.
KM = Kernel Matching.
LLM = Linear Local Matching.
PS = Propensity score.
RF = Random Forest.
SITE = Similarity preserved Individual Treatment Effect.

Sección 2.3

ADTE = Active Dynamic Transfer Entropy.
AIC = Akai Information Criteria.
BN = Bayesian Network.
D0TE = Direct Zero Transfer Entropy.
DBN = Dynamic Bayesian Network.
DCI = Dynamic Causality Index.
DISSIM = Dissimilarity Index.
DPCA = Dynamic PCA.
DTE = Direct Transfer Entropy.
DTW = Dynamic Time Warping.
EWMA = Exponential Weighted Moving Average.
FDD = Fault Detection and Diagnostics.
FTA = Faul Tree Analysis.
FTRCI = Fault Traversal and Root Cause identification.
GC = Granger Causality.
KECA = Kernel Entropy Component Analysis.
LASSO = Least Absolute Shrinkage and Selection Operator.
MFM = Multilevel Flow Models.
NNG = NoNegative Garrote.
OSAVA = Orthogonal Self-Attentive Variational Autoencoder.
P&ID = Piping and Instrumentation Diagrams.
PCA = Principal Component Analysis.
PDF = Probability Density Function.
RBC = Reconstruction Based Contribution.
RCI = Root Cause Identification.
SOR = Sum of Redundancy.
TE = Transference Entropy.
TTE = Trend Transfer Entropy.
VR = Variance Ratio.
VSAE = variational self-Attentive Autoencoder.
XML = eXtensible Markup Language.

Capítulo 3

Sección 3.1

INUS = Insufficient but Non-redundant part of an Unnecessary but Sufficient condition.

Sección 3.2

AGFI = Adjusted Goodness-of-fit statistic.
AIC = Akaike Information Criterion.
CAIC = The Consistent Version of AIC.
CFI = Comparative Fit Index.
CPD = Conditional Probability Distribution.
DAG = Direct Acyclic Graphs.
NFI = Normed-Fit Index.
PDAG = Partial Direct Acyclic Graphs.
PLS = Partial Least Square.
RMSEA = Root Mean Square Error of Approximation.
SEM = Structural Equation Models.
SRMR = Standardized Root mean square residual.

Sección 3.3

BART = Bayesian Additive Regression Tree.
CEM = Coarsened Exact Matching.
GAM = Generalized Additive Models.
GBM = Gradient Boosting Machine.
GenMatch = Genetic Matching.
GLM = Generalized Linear Models.
NNET = Single-hidden-layer neural network.
NNM = Nearest Neighbor Matching.
SMD = Standardized Mean Difference.
VR = Variance Ratio.

Capítulo 4

Sección 4.1

JIT = Just in Time.

Sección 4.3

ACP = Alteraciones en la cola de producción.
AP = Aceleración de la producción.
CD = Complejidad del diseño.
CMP = Calidad de mantenimientos preventivos.
EEHT = Eficiencia del equipo humano de trabajo.
EOM = Estado operativo de la máquina.
PP = Planificación de la producción.

Sección 4.6

CSV = Comma Separated Value.
EMA = Exploratory Matching Algorithm.
EMMSA = Exploratory matching model search algorithm.
HMSA = Homologous Model Selection Algorithm.

Sección 4.8

AvgEfMed = Valor medio de Efectos medios.
AvgEfMin = Valor mínimo de Efectos medios.
AvgS = Desviación estándar de Efectos medios.
AvgvMax = Valor máximo de Efectos medios.
MxEfMax = Valor máximo de Efectos máximos.
MxEfMed = Valor medio de Efectos máximos.
MxEfMin = Valor mínimo de Efectos máximos.
MxS = Desviación estándar de Efectos máximos.
ZSMax = Valor máximo de ZScore.
ZSMed = Valor medio de ZScore.
ZSMin = Valor Mínimo ZScore.
ZsS = Desviación estándar de ZScore.

Abreviaturas en orden alfabético

ACP = Alteraciones en la cola de producción.
ADTE = Active Dynamic Transfer Entropy.
AGFI = Adjusted Goodness-of-fit statistic.
AIC = Akai Information Criteria.
AIC = Akaike Information Criterion.
AP = Aceleración de la producción.
ATE = Average Treatment Effect.
ATT = Average Treatment Effect on the Treated group.
AvgEfMed = Valor medio de Efectos medios.
AvgEfMin = Valor mínimo de Efectos medios.
AvgS = Desviación estándar de Efectos medios.
AvgvMax = Valor máximo de Efectos medios.
BART = Bayesian Additive Regression Tree.
BN = Bayesian Network.
CAIC = The Consistent Version of AIC.
CART = Classification And Regression Tree.
CATE = Conditional Average Treatment Effect.
CD = Complejidad del diseño.
CEM = Coarsened Exact Matching.
CFI = Comparative Fit Index.
CMP = Calidad de mantenimientos preventivos.
CPD = Conditional Probability Distribution.
CSV = Comma Separated Value.
CTAM = Conditional Treatment-Adversarial learning based Matching.
D0TE = Direct Zero Transfer Entropy.
DAG = Direct Acyclic Graphs.
DBN = Dynamic Bayesian Network.
DCI = Dynamic Causality Index.
DISSIM = Dissimilarity Index.
DPCA = Dynamic PCA.
DTE = Direct Transfer Entropy.
DTW = Dynamic Time Warping.
EEHT = Eficiencia del equipo humano de trabajo.
EMA = Exploratory Matching Algorithm.
EMMSA = Exploratory matching model search algorithm.
EOM = Estado operativo de la máquina.
ERP = Enterprise Resource Planning
EWMA = Exponential Weighted Moving Average.
FDD = Fault Detection and Diagnostics.
FTA = Faul Tree Analysis.
FTRCI = Fault Traversal and Root Cause identification.
GAM = Generalized Additive Models.
GBM = Gradient Boosting Machine.
GC = Granger Causality.
GenMatch = Genetic Matching.
GLM = Generalized Linear Models.
HMSA = Homologous Model Selection Algorithm.
i.i.d = independent and identically distributed.
INUS = Insufficient but Non-redundant part of an Unnecessary but Sufficient condition.
IPTW = Inverse Probability of Treatment Weighting.
IPW = Inverse Propensity Weighting.
JIT = Just in Time.
KECA = Kernel Entropy Component Analysis.
KM = Kernel Matching.
LASSO = Least Absolute Shrinkage and Selection Operator.
LLM = Linear Local Matching.
MFM = Multilevel Flow Models.
MxEfMax = Valor máximo de Efectos máximos.
MxEfMed = Valor medio de Efectos máximos.
MxEfMin = Valor mínimo de Efectos máximos.
MxS = Desviación estándar de Efectos máximos.
NFI = Normed-Fit Index.
NNET = Single-hidden-layer neural network.
NNG = NoNegative Garrote.
NNM = Nearest Neighbor Matching.
OSAVA = Orthogonal Self-Attentive Variational Autoencoder.
P&ID = Piping and Instrumentation Diagrams.
PCA = Principal Component Analysis.
PDAG = Partial Direct Acyclic Graphs.
PDF = Probability Density Function.
PLS = Partial Least Square.
PP = Planificación de la producción.
PS = Propensity score.
RBC = Reconstruction Based Contribution.
RCI = Root Cause Identification.
RF = Random Forest.
RMSEA = Root Mean Square Error of Approximation.
SEM = Structural Equation Models.
SITE = Similarity preserved Individual Treatment Effect.
SMD = Standardized Mean Difference.
SOR = Sum of Redundancy.
SRMR = Standardized Root mean square residual.
SUTVA = Stable Unit Treatment Value Assumption.
TE = Transference Entropy.
TTE = Trend Transfer Entropy.
VR = Variance Ratio.
VSAE = Variational Self-Attentive Autoencoder.
XML = eXtensible Markup Language.
ZSMax = Valor máximo de ZScore.
ZSMed = Valor medio de ZScore.
ZSMin = Valor Mínimo ZScore.
ZsS = Desviación estándar de ZScore

1. INTRODUCCIÓN

1.1. Antecedentes

Con el aumento del comercio electrónico y las exportaciones post pandemia, la industria cartonera ha crecido vertiginosamente en los últimos años debido a las necesidades de empaquetamiento. Esto se debe principalmente a las ventajas que ofrece el cartón por su versatilidad para proteger, almacenar y transportar productos. Se estima que el volumen de producción de cartón para el 2040 alcanzará los 700 millones de toneladas métricas, y el tamaño de mercado estaría alrededor de los 220 mil millones de dólares para el 2028 (Research and Markets, 2023). El crecimiento de la demanda generará más desperdicios, lo que repercutirá en un mayor control de los impactos ambientales, energéticos y económicos de sus procesos productivos.

El cartón utilizado en el embalaje está formado por dos láminas de papel externas, y entre ellas se incorpora un lámina de papel corrugado para darle la capacidad de absorción de energía que se requiera. En una industria cartonera, el sistema de producción consta de dos subsistemas: el subsistema de corrugado y el subsistema de impresión. El primero elabora láminas de cartón a partir de bobinas de papel, cuyas dimensiones y tipo de corrugado dependerán de las características del diseño del producto. El segundo, imprime las láminas de cartón, recorta las diferentes unidades que se diseñen en cada lámina y luego procede con el doblado, pegado, apilado, atado y puesta en pallets para su embalaje y despacho.

Antes de que una orden de producción sea procesada, se abastece previamente de los recursos requeridos desde otras áreas departamentales de la organización. Por ejemplo, el planificador de la planta de producción asigna los recursos humanos previstos y orquesta la provisión de recursos materiales; el subsistema de corrugado provisiona las láminas de cartón; el departamento de mantenimiento verifica la operatividad de la máquina y el desgaste de consumibles al inicio o fin de turno; el departamento de clisé proporciona los diseños de impresión del pedido (clisés y troquel), revisa su conformidad y verifica el desgaste; y el departamento de aprovisionamiento provee las tintas, gomas, consumibles, repuestos, etc.

La máquina empieza a procesar una orden con una parada de inicio llamada *setup* en la que se configura la orden de producción, se carga el troquel y se alinean los clisés. Una vez comprobadas la exactitud de las alineaciones de los colores, cortes y tonalidades en las impresiones de prueba, se imprime la orden de producción. Durante el proceso de impresión puede haber cero o más paradas de la máquina. Las paradas después del *setup* pueden ser programadas o no programadas. Todas las paradas se registran en tiempo real por medio de un software de seguimiento de alarmas que almacena la fecha, el tipo de parada, el código de la parada, y la duración de cada parada.

Por otro lado, los pedidos se registran en el sistema de información ERP (*Enterprise Resource Planing*) de la industria, pero no está integrado con el *Sistema de Gestión de Alarmas*. Cada sistema proporciona un solo fichero de datos que se utilizarán en el presente estudio.

Los desperdicios pueden surgir por motivos intrínsecos al diseño del producto, que se pueden calcular multiplicando el total de unidades producidas por el área no útil. Este desperdicio es inevitable, sobre él cual no podemos intervenir, ya que normalmente los diseños son optimizados para minimizarlo. Estamos interesados en el desperdicio generado por el subsistema de impresión, especialmente de las unidades de desperdicio (o no conformes) que se generan. Las unidades de desperdicio son contadas al finalizar la producción del pedido, ya que no es práctico hacerlo en cada parada porque involucra tiempo de inactividad de la máquina. Por esta razón,

no se disponen de datos de desperdicio por cada parada. Existen alrededor de ciento veinte tipos de paradas correspondientes a las diferentes etapas de la máquina.

Estas paradas pueden originarse por las condiciones de operatividad de la máquina, tonalidad incorrecta, cuchillas del troquel averiadas o sin filo, restos de cartón que se introducen en la etapa de impresión, atascos, fallos inesperados, inadecuada velocidad de la máquina, entre otros.

En el periodo de estudio (2019-2021), el número de registros de paradas de la planta se aproximan a medio millón de registros, y el número de pedidos totales se aproximan a ciento veinte mil pedidos, distribuidos en pedidos "*just in time*" (JIT) y no JIT. Nos interesa el desperdicio generado por pedidos JIT, especialmente por las repercusiones en la calidad del servicio y en la satisfacción del cliente. La planta opera en tres turnos rotativos, 24 horas al día, 7 días a la semana.

1.2. Descripción del problema a resolver

Reducir el desperdicio por paradas es relevante por la creciente demanda y por su impacto en el medio ambiente. Además, interesa el control del desperdicio por las nuevas regulaciones de producción verde que progresivamente se implantan en el espacio europeo con la agenda 2030. La estrategia seguida ha sido abordar todas las fuentes de desperdicio, por ejemplo, con "Mantenimiento Productivo Total" (Kot & Grondys, 2013), pero sería de valiosa ayuda y tendría mayor impacto si previo a cualquier estrategia respondemos las siguientes preguntas:

- *¿qué máquinas generan la mayor cantidad de desperdicio?*
- *¿en qué intervalos del tiempo de parada ocurren los mayores desperdicios?*
- *¿cuáles son las máquinas e intervalos de paradas con mayor desperdicio?*
- *¿por dónde se debería empezar a reducir el desperdicio?*

Es complejo determinar qué paradas son más importantes si no se disponen de datos de desperdicio por paradas. Sin embargo, si existiera un desperdicio significativo, es muy probable que la máquina se detenga y el tiempo de parada sea mayor. Por tanto, si medimos el tamaño del efecto del tiempo de paradas sobre el desperdicio en cada máquina, podríamos responder estas preguntas. Esto requiere investigar ¿cuál es el tamaño del efecto causal del tiempo total de parada (posteriores al *setup*) sobre el desperdicio?

1.3. Objetivos de la Investigación

1.3.1. Objetivo General

Desarrollar una metodología para medir el tamaño del efecto causal del tiempo total de paradas sobre el desperdicio.

1.3.2. Objetivos Específicos

I. Revisar la literatura de los conceptos fundamentales de inferencia causal, los métodos estadísticos de evaluación de efectos causales, y las principales aplicaciones en la industria.

II. Elaborar el marco teórico de las teorías causales, el modelamiento causal, el diseño de estudios observacionales, los métodos de control y ajuste de variables confusoras y los métodos de integración de la información.

III. Describir el Estudio Observacional Causal que incluya la descripción de la unidad de análisis en sus niveles macro y meso, el planteamiento del problema y la formulación de la pregunta causal, la descripción del conjunto de datos originales, su preprocesamiento, preselección y roles de las variables, el estudio de los mecanismos causales y factores de confusión, y el análisis de identificación.

IV. Presentar la metodología propuesta, los algoritmos utilizados, sus resultados, el meta-análisis de los niveles meso y macro, el esquema del código utilizado, la discusión de los resultados, las ventajas y desventajas de la metodología, su reproducibilidad y delinear las futuras líneas de investigación.

1.4. Estructura de la Tesis

Para abordar los objetivos específicos, en el Capítulo 2 se realizará un repaso de la literatura sobre los conceptos fundamentales de inferencia causal, los métodos de evaluación de efectos causales y los principales métodos causales utilizados en la industria.

En el Capítulo 3 se desarrollará el Marco Teórico. La Sección 3.1 incluirá las principales Teorías de Causalidad como la Teoría de la Regularidad, la Teoría de la Causalidad Probabilística y la Teoría Contrafactual. Además, en la Sección 3.2 se introducirán los principales marcos de trabajo (*framework*) de modelamiento causal, como son los Modelos de Ecuaciones Estructurales, Modelo de Grafos Causales y el Modelo Estructural Causal.

En la Sección 3.3 se describirán los métodos de ajuste y control de variables confusoras, empezando con el Diseño de Estudios Observacionales, los métodos de matching, las medidas de proximidad, la evaluación de la calidad de los grupos emparejados, el tamaño del efecto y el análisis de sensibilidad. Finalmente, se revisan los métodos de integración de la información (meta-análisis), usando el modelo de efectos comunes y el modelo de efectos aleatorios.

En el Capítulo 4 se expondrán los componentes del Estudio Observacional realizado. La Sección 4.1 formulará el estudio con la descripción de la unidad de análisis macro y meso, el nivel de abstracción, la descripción del problema, la población de estudio y el periodo de seguimiento. En la Sección 4.2 se describirán las fuentes de datos, el preprocesamiento, la selección de variables, sus roles y la descripción del conjunto de datos del estudio.

En las Secciones 4.3 y 4.4, se revisarán con el experto los factores de confusión presentes en la generación de desperdicio y se realizará el análisis de identificación del efecto causal. Luego, en la Sección 4.5 se introducirá la metodología propuesta EMMSA (*Exploratory matching model search algorithm*) describiendo las diferentes etapas que la conforman.

En la Sección 4.6 se presentan el Algoritmo de Matching Exploratorio (*Exploratory Matching Algorithm*, EMA) y el Algoritmo de Selección de Modelos Homólogos (*Homologous Model Selection Algorithm*, HMSA). En la Sección 4.7 se muestran los resultados de la metodología EMMSA.

En la Sección 4.8 se realiza el Análisis Causal que incluye la medición del tamaño del efecto y el meta-análisis de los niveles meso y macro, antes y después del análisis de sensibilidad. En la Sección 4.9 se resumirán las principales aportaciones de esta tesis. Finalmente, en la Sección 4.10 se revisarán los aspectos de reproducibilidad del estudio.

En el Capítulo 5 se extraerán las conclusiones del estudio, empezando con un resumen de los resultados, su aplicación práctica, las limitaciones del estudio, los datos de la publicación asociada a la tesis y las futuras líneas de investigación.

El Capítulo 6 organiza el material complementario tales como: repositorios de datos, programas y salidas; librerías utilizadas; resultados distribuidos por métodos y distancias; modelos seleccionados con mayor ZScore;

modelos seleccionados después del análisis de sensibilidad; resultados del meta-análisis para los niveles meso y macro; la inclusión de la publicación asociada; el código de EMA, el código de HMSA; y finalmente la visualización de las primeras líneas del conjunto de datos.

2. REVISIÓN DE LA LITERATURA

Desde la antigüedad, el hombre ha estado preocupado por conocer las causas de los fenómenos naturales, personales, sociales y trascendentales. Su interés fue madurando y con el pasar de los siglos se fueron desarrollando diversas teorías con aciertos, enfoques revolucionarios y limitaciones, desde Aristóteles (Falcon, 2023) hasta Imbens, Rubin (G. W. Imbens & Rubin, 2015) y Pearl (Pearl, 2009b). Desde sus orígenes, el interés por las relaciones causales se convirtió en esencial para cualquier tipo de conocimiento que pretendía acercarse a la verdad y así, capturar los procesos de generación de nuevas realidades posibles.

El conocimiento causal es uno de los más buscados por los investigadores, juristas y tomadores de decisiones, porque modelan cómo se estructuran y relacionan los objetos, acontecimientos o eventos que interactúan en una unidad de análisis. Evidentemente, con el surgimiento de nuevas metodologías causales, impulsado por el avance tecnológico del nuevo milenio, estamos viviendo un cambio de paradigma en la investigación disciplinar para pasar de estudios empíricos asociacionales a estudios observacionales causales, lo cual es muy probable que se vea más evidenciado con el paso de los años.

En la actualidad, la causalidad es un campo emergente de conocimiento interdisciplinar que ha adquirido un vertiginoso interés, y se nutre de disciplinas de áreas de conocimiento tan diversas como las ciencias filosóficas, matemáticas, económicas, políticas, jurídicas, psicológicas, médicas, computación, entre otras.

Dentro del contexto científico, existen dos tipos de estudios causales. Primero, el estándar de oro son los *ensayos aleatorizado*s, en donde el investigador diseña experimentos y controla a qué individuos o unidades se les aplica un determinado tratamiento, usando esencialmente métodos de asignación aleatoria (Angrist & Pischke, 2009). Segundo, los *estudios observacionales*, en los cuales el investigador no puede controlar el mecanismo de asignación de tratamientos, porque los datos se generan según las regularidades o singularidades de un sistema, como consecuencia de sus interacciones factuales (P. R. Rosenbaum, 2010).

En el campo de los estudios observacionales, ha habido un interés creciente e intenso en los últimos años debido a que estos métodos son vistos como herramientas más económicas frente a los ensayos aleatorizados, no intrusivos y permiten mejorar la comprensión de los fenómenos. Para esto, se requiere estimar el contrafactual, es decir, el resultado que se hubiese obtenido si no se hubiera aplicado un tratamiento.

En los estudios observacionales causales, existen diferentes subcampos de conocimiento como la inferencia causal y el descubrimiento causal. La *inferencia causal* se preocupa de responder preguntas causales a partir de datos empíricos provenientes de estudios observacionales o experimentales. El *descubrimiento causal,* a partir de datos observacionales, procura modelar o inferir estructuras causales subyacentes de las variables de un sistema, representándolas con grafos dirigidos acíclicos, conocidos como DAG (*Direct Acyclic Graphs*) (Spirtes et al., 2000).

Así, contestar preguntas causales a partir de estudios observacionales resulta muy difícil porque requiere comprender cómo se comporta el flujo de información en la estructura subyacente, lo cual requiere inferencias y triangulaciones más allá de los datos. Es decir, pretendemos derivar conclusiones causales a partir de datos factuales generados por complejos procesos estocásticos (Shalizi, 2018).

2.1. Conceptos Fundamentales de Inferencia Causal

A continuación, empezaremos a abordar el núcleo conceptual de la inferencia causal desde la perspectiva de los estudios observacionales.

2.1.1. Causalidad

La causalidad estudia causas y efectos, y determina de qué manera una causa actual es responsable de su efecto y el efecto dependiente de su causa (Hernán & Robins, 2020).

La causalidad ha sido abordada desde distintas disciplinas como la (bio) estadística (Han & Zhou, 2022; Hullsiek & Louis, 2002), econometría (Granger, 1969; G. W. Imbens & Rubin, 2015), epidemiología (Rothman & Greenland, 2005; Vandenbroucke et al., 2016), psiquiatría (Ohlsson & Kendler, 2020), filosofía (Falcon, 2023; Pearl, 2009a), política (Box-Steffensmeier et al., 2009), aprendizaje automático (Singh et al., 2017), jurisprudencia (Ho & Rubin, 2011; Knobe & Shapiro, 2021), entre otras.

Los conceptos fundamentales han sido provistos desde las teorías filosófica de la causalidad en medio de un intenso debate. Entre estas tenemos: la *Teoría de la Regularidad* (Andreas & Guenther, 2021; Mill, 1865), la *Teoría Contrafactual* (Lewis, 1986b), la *Teoría Probabilística* (Suppes, 1984), la *Teoría de Modelos de Ecuaciones Estructurales* (K. Bollen, 1989; Jöreskog & Sörbom, 1980), la *Teoría del Modelo Estructural Causal* (Pearl, 2009b), la *Teoría de los Modelos Causales* (Hitchcock, 2009; Pearl, 2009b; Spirtes et al., 2000; Wright, 1925), entre otras. Estas teorías causales no podemos abordarlas con profundidad por la inviabilidad de incluir este vasto conocimiento en el presente trabajo.

En un sentido amplio, la causalidad está relacionada con *una acción aplicada a una unidad* (Imbens & Rubin, 2015, p.4), en el contexto de una *unidad de análisis*. La *acción* puede ser la manipulación de una variable, la aplicación de un tratamiento o la realización de una intervención. En todos estos casos a la acción la llamaremos simplemente *tratamiento*, por ser un estándar en las disciplinas epidemiológicas y econométricas.

2.1.2. Unidad de observación y unidad de análisis

La unidad de análisis puede ser cualquier ente susceptible de ser alterado (interna o externamente) y estudiado en momentos diferentes. Por ejemplo, un objeto físico, una persona, un grupo de personas, un aula de clases, un área departamental, un negocio, una comunidad, un mercado (Imbens & Rubin, 2015, p.4), una máquina industrial o una industria. La *unidad observada* es un sujeto u objeto del cual se miden o recopilan datos generados por (o en) la unidad de análisis (Sedgwick, 2014).

La unidad de análisis es el ente del que se analiza la información de sus unidades observadas y del que se extraen conclusiones al finalizar el análisis (Sedgwick, 2014). En el Capítulo 4, a la unidad de análisis la enfocaremos como sistema (Bertalanffy, 1950) para facilitar la comprensión del estudio, debido a que existen diferentes tipos de relaciones causales embebidos en diferentes tipos de sistemas (Cartwright, 2004).

Cada unidad observada proporciona un solo registro de datos, y como hay diferentes medidas que se pueden tomar de cada unidad, éstas pueden variar de una unidad a otra, comportándose como variables aleatorias (Udny Yule, 1897). El registro de cada unidad consta de variables aleatorias que, dependiendo del rol que desempeñan, pueden ser categorizadas como *covariable o pretratamiento, tratamiento*, *postratamiento* y *resultado*.

2.1.3. Definición de inferencia causal

La inferencia causal es el proceso de extraer conclusiones a partir de las relaciones de dependencia de un sistema, determinar las condiciones para que una causa produzca un efecto (Yao et al., 2021) *y analizar el cambio del efecto cuando la causa es manipulada* (Pearl, 2009b). Para esto, se formulan preguntas de inferencia causal y se establecen *supuestos críticos y explícitos que requieren ser justificados* para hacer viable la inferencia causal.

Las preguntas de inferencia causal se pueden realizar en dos direcciones: hacia adelante o hacia atrás. Es decir, los efectos de las causas - hacia adelante -, y la causa de los efectos - hacia atrás (Gelman, 2010; Mill, 1843). La inferencia causal hacia adelante se puede modelar usando la notación del Modelo de Resultado Potencial (Neyman, 1923; Rubin, 1974), Contrafactual (Lewis, 1986a, 2000), o su equivalente usando Modelos Estructural Causal (Pearl, 2009b). La inferencia causal hacia atrás requiere estudiar trayectorias causales en términos de inferencia causal hacia adelante (Gelman, 2010).

Sin embargo, esta tarea es desafiante, agotadora y desalentadora por su extrema complejidad (Hernán & Robins, 2020; Yao et al., 2021). Esta complejidad radica en la necesidad de triangular evidencias de múltiples fuentes y en la aplicación de una variedad de metodologías de diferentes ciencias (Hernán & Robins, 2020). Su naturaleza interdisciplinar (Szostak, 2013) ha generado confusión y ralentizado su desarrollo.

Para evitar esta confusión y desorientación, un estudio causal debe partir de un diseño cuidadoso. Esto implica partir de una pregunta causal explícita sobre el comportamiento de una unidad de análisis en un intervalo de tiempo, seleccionar las fuentes de datos, identificar el rol de las variables del conjunto de datos seleccionado, estudiar los mecanismos causales y los factores de confusión presentes (Sección 2.1.7); establecer la estrategia de identificación seleccionando un modelo causal, establecer supuestos verificables que soporten la interpretabilidad de los hallazgos, y seleccionar un estimador estadístico apropiado para medir el efecto del tratamiento (Dahabreh & Bibbins-Domingo, 2024). Además, debe discutir los resultados y concluir con una interpretación causal defendible de los hallazgos que expliquen el comportamiento de la unidad de análisis (Hernán & Robins, 2020), usando una estructura condicional con juicios subjetivos (Dahabreh & Bibbins-Domingo, 2024).

2.1.4. Tratamiento, Pretratamiento, Postratamiento y Proxy

Se concibe como tratamiento (o manipulación) a los distintos estados causales que pueden ser aplicados a las unidades de una población de interés pertenecientes a una unidad de análisis. En cada unidad que se aplica un tratamiento, se va a medir *a posteriori* el cambio producido en la *variable resultado* (VanderWeele & Hernan, 2013).

Los estados causales del tratamiento pueden ser binario o múltiples, y estos estados causales alternativos deben ser mutuamente excluyentes (Rubin, 1974, 1990). Cada estado causal es conocido como "*nivel del tratamiento*" o "*versión del tratamiento*" (VanderWeele & Hernan, 2013). *A priori*, el valor de la variable resultado de una unidad será observable cuando se aplique algún nivel del tratamiento. Un tratamiento puede ser más activo - con mayor incidencia, y otros más pasivo - con menor o nula incidencia (Morgan & Winship, 2007).

Cuando la variable causal es dicotómica, se asigna a los estados causales los valores de *0* o *1*, para indicar la ausencia o presencia del tratamiento, respectivamente. Tradicionalmente, si la incidencia es nula o ausente, se llama *tratamiento de control.* Si es activo se llama simplemente *tratamiento* (Imbens & Rubin, 2015, p.4). Además, las unidades observadas que no reciben el tratamiento activo se conocen como *grupo de control* y las que sí reciben se denomina *grupo de tratamiento* (Morgan & Winship, 2015).

Cuando la variable causal tiene múltiples tratamientos, su número dependerá de la cantidad de datos disponibles de cada tratamiento que aseguren la posibilidad de medir su respectivo efecto. Para cada nivel de tratamiento es posible obtener valores diferentes en la variable resultado, pero también es posible que para algunos niveles de tratamiento se obtengan valores iguales o aproximadamente iguales en la variable resultado. Esto es conocido como "*variación irrelevante del tratamiento*" (VanderWeele, 2009).

Las variables *pretratamiento,* también llamadas *variables del contexto,* son aquellas que no son afectadas por el tratamiento (Yao et al., 2021). Generalmente, son características preexistentes, permanentes y conocidas de la unidad observada previo a la aplicación del tratamiento. Al ser variables relacionadas con el tratamiento también se las conoce como *covariables*. Estas variables se utilizan para estimar de forma precisa el resultado potencial mediante la creación de subgrupos homogéneos de covariables (Imbens & Rubin, 2015, p.16).

Se consideran *variables postratamiento* aquellas que están afectadas por el tratamiento (Yao et al., 2021). Estas variables también podrían afectar al resultado, en cuyo caso se siguen diferentes estrategias para aislar la incidencia de la variable postratamiento y así poder medir el efecto causal (Frangakis & Rubin, 2002; Joffe, 2011).

Las *variables proxy* son variables que están relacionadas con algún factor de confusión no observado y se las suele incluir en el conjunto de datos para reducir el sesgo de confusión oculto (VanderWeele, 2019).

2.1.5. Resultado potencial

Cuando analizamos un tratamiento aplicado sobre una unidad (Sección 2.1.2), el valor de la variable resultado, o *resultado observado*, puede ser temporalmente determinado después de la aplicación del tratamiento; estableciéndose así una dirección temporal causal (Lewis, 1986a).

El *efecto causal individual* de un tratamiento con respecto a otro, aplicados sobre la misma unidad en el mismo momento postratamiento, es la *comparación de sus resultados observados*, ya sea como diferencia o cociente (Imbens & Rubin, 2015, p.6).

Sin embargo, en una unidad se puede observar un solo tratamiento a la vez, debido a que después de aplicarlo, la unidad ya ha cambiado o se ha contaminado de este tratamiento y, por tanto, no se podría aplicar otro. Por esta razón, a este valor de la variable resultado, dado la aplicación del único tratamiento que puede recibir, se lo conoce como *resultado potencial* (Imbens & Rubin, 2015, p.4) o *resultado factual*. "El *resultado contrafactual* es el resultado si la unidad no recibió el tratamiento" (Yao et al., 2021), refiriéndose a un resultado potencial no observado.

Esto exige *estimar el efecto causal* considerando múltiples unidades, entre las cuales se establece un grupo expuesto a un tratamiento activo (grupo de tratamiento) y otro expuesto al tratamiento alternativo (grupo de control) (Imbens & Rubin, 2015, p.6).

Así, el problema de la inferencia causal radica en que sólo podemos observar una parte de los resultados potenciales por cada unidad de observación (Holland, 1986, p.947; Imbens & Rubin, 2015, p.21). En otras palabras, el problema fundamental de la inferencia causal es un problema de datos faltantes (Rubin, 1974), lo que conduce al siguiente problema: ¿cómo se selecciona adecuadamente las unidades que se les asignará o no un tratamiento? Este proceso es conocido como *mecanismo de asignación del tratamiento* (Imbens & Rubin, 2015, p.6), utilizado para evitar la introducción de sesgos por una inadecuada selección de unidades, lo cual está relacionado con el *diseño del estudio*.

En experimentos aleatorizados, el mecanismo que asigna el tratamiento a cada unidad observada es conocido y controlado por el investigador, pero en estudios observacionales el investigador no tiene control, ni conoce cómo se realiza la asignación del tratamiento (Cochran & Chambers, 1965).

Por tanto, para emprender la tarea causal, se requieren de un cuidadoso diseño de estudio y de supuestos: el primero es un factor determinante (Imbens & Rubin, 2015, p.12.14) y el segundo, es necesario y exigible para hacer viables e interpretables las afirmaciones causales. Además, para establecer adecuadamente un problema de análisis causal, es conveniente construir una *pregunta o afirmación causal* lo más clara y precisa posible, que esté articulada con la variable tratamiento y con su resultado potencial (Imbens & Rubin, 2015, p.5).

2.1.6. Modelo de Resultado Potencial de Neyman y Rubin

El Modelo de Resultado Potencial tuvo sus orígenes con una serie de trabajos experimentales agrícolas elaborados por Neyman (1923), quien ideó la comparación de resultados potenciales. Su desarrollo fue avanzando en esta área de estudio hasta los trabajos de Cochran (1954) y Cox (1958). También tuvo sus raíces en las ciencias económicas con los trabajos de Quandt (1972) y Roy (1951). Finalmente, este modelo fue definido por Rubin (1974) y, tras varios trabajos lo formalizó en 1990 (Rubin, 1990). El Modelo de Resultado Potencial es equivalente al Modelo Estructural Causal (Pearl, 2012a) y sigue un enfoque contrafactual similar al de Lewis (Lewis, 1973, 1986a, 2000) definiendo distribuciones equilibradas.

En este modelo, se utilizan los conceptos definidos anteriormente. Utilizamos la variable aleatoria T para indicar los estados causales (o tratamientos) que pueden ser $\{0, 1, 2, \ldots, N_T\}$, donde N_T es el número máximo de tratamientos activos. Solamente un tratamiento a la vez podrá ser aplicado a cada unidad, lo que conocemos como *estado factual (F)*, y el estado que no se aplicó se lo conoce como *estado contrafactual (CF)*. El resultado contrafactual es fundamental para establecer efectos causales y permite explorar lo que podría ocurrir en diferentes escenarios posibles mejorando la comprensión del efecto (Hernán & Robins, 2020, p.4; Morgan & Winship, 2015, p.37).

Para tratamientos dicotómicos, T=1 denota los miembros de la población que son expuestos al tratamiento activo (F) pero no al tratamiento de control (CF); y T=0 denota los miembros de la población que son expuestos a las condiciones de control (F) pero no al tratamiento activo (CF). *T* denota el estado factual y *(1-T)* el estado contrafactual.

Debido a la imposibilidad de calcular el efecto causal individual, se estima el Efecto Causal Agregado (*Average Treatment Effect – ATE*) como los promedios de los efectos individuales de la población como un todo. Las variables aleatorias de los resultados potenciales de los grupos de tratamiento y control se las representa por Y^1 o Y^0, respectivamente, siendo posible observar solo uno de ellos a nivel individual. *E[.]* es el operador del valor esperado de la teoría de probabilidades.

$$ATE = E[Y^1 - Y^0] = E[Y^1] - E[Y^0] \quad (1)$$

Sin embargo, si los efectos varían entre subgrupos, ATE no sería una medida representativa para cada uno de ellos. Por esta razón, se introduce el efecto del tratamiento sobre el grupo de los tratados y se conoce como *Average Treatment Effect on the Treated group* (ATT). Se centra en las unidades que son destinadas a uno de los dos tratamientos. Se define como:

$$ATT = E[Y^1 - Y^0 | T=1] = E[Y^1 | T=1] - E[Y^0 | T=1] \quad (2)$$

Siendo $E[Y^1|T=1]$ y $E[Y^0|T=1]$, el valor esperado del resultado potencial de tratamiento y control sobre el grupo de los tratados (que han recibido alguno de los dos tratamientos), respectivamente.

Cuando se intenta estimar el efecto de un tratamiento en subpoblaciones definidas, condicionadas en ciertas covariables, se utiliza "*Conditional Average Treatment Effect*" (CATE).

$$CATE = E[Y^1|X=x] - E[Y^0|X=x] \quad (3)$$

Siguiendo a Yao et al. (2021), el conjunto de datos observacional lo describimos como el conjunto $\{X_i, T_i, P_i, Y_i^F\}_{i=1}^N$, donde N representa el número total de unidades, X_i representa el vector de variables pretratamiento o covariables de la unidad i, T_i el tratamiento aplicado a la unidad i, P_i representa el vector de variables postratamiento de la unidad i, y Y_i^F el resultado potencial factual de la unidad i.

Imbens & Wooldridge (2009) consideran cinco ventajas del Modelo de Resultado Potencial:

i. Permite definir el efecto causal de interés sin considerar propiedades probabilísticas, ni definir previamente el mecanismo de asignación (endogeneidad o exogeneidad), ni asumir si los efectos son constantes o varían a lo largo de la población, ni especificar antes los supuestos de la forma funcional o distribucional (e.g. modelos de regresión); separando estos aspectos.
ii. Vincula el análisis del efecto causal para manipulaciones explícitas, obligando al investigador a pensar en los escenarios en los cuales se podría identificar los efectos causales, facilitando su interpretación.
iii. Separa la modelización (resultado potencial) del mecanismo de asignación, permitiendo múltiples fuentes de datos o la triangulación de información.
iv. Permite formular supuestos probabilísticos en términos de variables potencialmente observables y evaluar la validez considerando la estructura de dependencia si todos los resultados fueran observados, evitando así las críticas a los términos de error por factores ocultos en los modelos de regresión.
v. Clarifica de dónde viene la incertidumbre en los estimadores.

2.1.7. Supuestos fundamentales en Inferencia Causal

Existen tres supuestos fundamentales en la literatura sobre inferencia causal para la estimación del efecto de un tratamiento.

2.1.7.1. Supuesto "Stable Unit Treatment Value Assumption" - SUTVA

"Los resultados potenciales de cualquier unidad no varían con el tratamiento asignado a otras unidades y, para cada unidad, no existen diferentes formas o versiones de cada nivel de tratamiento, que conduzcan a resultados potenciales diferentes" (Rubin, 1980).

Este supuesto indica que ninguna unidad *interfiere* con otra, es decir, no hay *interacción entre unidades* (Cox, 1958). En otras palabras, las unidades son independientes e idénticamente distribuidas (*i.i.d.*). Además, indica que cada unidad recibe una sola versión del tratamiento. Este supuesto está presente en la mayoría de los métodos de inferencia causal.

Sin embargo, preocupa cuando existe dependencia de datos y tratamientos continuos. Para solucionar la falta de cumplimiento de este supuesto se puede redefinir la unidad de análisis de interés, cambiar el nivel de abstracción, o modelar directamente la interacción de las unidades de observación (Imbens & Wooldridge, 2009). En la Sección 2.2.10.1 se presentan algunos métodos para relajar este supuesto.

2.1.7.2. Supuesto de Ignorabilidad (Ignorability) o ausencia de confusión

"*Dadas las variables pretratamiento observadas, el tratamiento asignado T es independiente de los resultados potenciales, es decir,* $T \perp (Y^0, Y^1) \mid X$" (P. Rosenbaum, 2002; Rubin, 1978).

Esto indica que, si dos unidades tienen las mismas variables pretratamiento (observadas) del contexto X, su asignación al tratamiento puede considerarse aleatoria y sus resultados potenciales deberían ser los mismos, sea cual sea el tratamiento asignado (Yao et al., 2021). En otras palabras, el mecanismo de asignación del tratamiento es *ignorable* (Rosenbaum & Rubin, 1983) cuando el resultado potencial es independiente del tratamiento, dentro del estrato definido para las combinaciones del vector de covariables que determinan sistemáticamente la asignación del tratamiento (Morgan & Winship, 2015, p.120).

Este supuesto también es conocido como *Supuesto de Independencia Condicional* (Lechner, 1999). Tiene una estrecha conexión con el *criterio Back-Door* de Pearl (2009a), aunque presenta algunas diferencias (Morgan & Winship, 2015, p.120). El Supuesto de Ignorabilidad también es conocido como *Ausencia de confusión* (G. W. Imbens, 2004) y está presente en la mayoría de los estudios de inferencia causal existentes.

Sin embargo, si hay confusores ocultos la asignación ya no podría considerarse aleatoria, siendo este supuesto muy difícil de cumplir porque es imposible identificar y medir todos las confusores que intervienen en la asignación del tratamiento. Además, la presencia de confusores ocultos tiene fuertes implicaciones en el aprendizaje de efectos causales porque puede derivar en conclusiones erradas y estimaciones imprecisas o sesgadas. En estos casos, controlar el sesgo de confusión es fundamental y un desafío a la hora de estimar efectos causales por su misma naturaleza no observable.

En un estudio observacional, Morgan & Winship (2015, p.120) ofrece dos criterios que un investigador podría evaluar si la asignación del tratamiento es ignorable:

i. *Teórico*: Determinar qué son las covariables, a partir de estudios pasados y supuestos fundamentados en la teoría
ii. *Práctico*: Medir cada una de las covariables y colectar suficientes datos que permitan estimar consistentemente el resultado potencial en cada estrato definido en X.

En la Sección 2.2.10.2 se detallan los diferentes métodos que actualmente se utilizan para relajar este supuesto.

2.1.7.3. Supuesto de Positividad

"Para cualquier valor de X, la asignación del tratamiento es no determinístico":

$$0 < P(T=t \mid X=x) < 1 \; ; \; \forall \, t, x \tag{4}$$

Este supuesto también es conocido como *superposición de covariables* o *soporte común* en la cobertura de las covariables entre los grupos de tratamiento y control, lo que permite una comparación entre ambos grupos.

Rosenbaum y Rubin (1983) denominan a los supuestos de ignorabilidad y positividad como *fuerte ignorabilidad*, también llamado ausencia de confusión (*unconfoundedness*) o exogeneidad (*exogeneity*) en la literatura econométrica (G. W. Imbens, 2004).

En la Sección 2.2.10.3 se señala las estrategias que se utilizan para relajar este supuesto.

2.1.8. Mecanismos de asignación

El mecanismo de asignación del tratamiento es un componente fundamental en el análisis causal y consiste en el proceso por el cual se escogen qué unidades reciben qué nivel de tratamiento y, por tanto, qué resultados potenciales se observan y cuáles no. Este mecanismo describe la probabilidad de cualquier vector de asignación como una función de todas las covariables y los resultados potenciales (Imbens & Rubin, 2015, p.31).

Existen tres restricciones o *propiedades de los mecanismos de asignación*:

i. la *asignación individualista (nivel unitario),* en función de sus covariables y resultados potenciales.

ii. la *asignación probabilística*, cuando la probabilidad de asignación unitaria está entre cero y uno, para cada unidad.

iii. la *asignación sin confusión*, cuando el mecanismo de asignación no depende de los resultados potenciales y solo depende de sus covariables, $T \perp (Y^0, Y^1) | X$". La asignación probabilística y sin confusión es también llamada *asignación del tratamiento fuertemente ignorable* (Rosenbaum & Rubin, 1983).

Existen tres tipos de mecanismos de asignación según Imbens & Rubin (2015, p.41):

i. Experimentos aleatorizados clásicos. El mecanismo de asignación satisface las tres restricciones anteriores para el diseño de un experimento. El investigador conoce y controla el mecanismo de asignación, y el efecto causal se estima de manera directa y sencilla. Sin embargo, estos estudios tienen la desventaja que son costosos, consumen tiempo y por ello no pueden involucrar muchas unidades observadas, siendo posible que éstas no sean representativas de la población y por ello, sus conclusiones son con respecto a la muestra. Además, en ocasiones es inviable realizarlo por razones prácticas o éticas.

ii. Estudios observacionales con Mecanismos de asignación regular. El investigador no tiene por qué conocer ni controlar el mecanismo de asignación de unidades al tratamiento, solo observa unidades y registra datos. También cumplen las tres restricciones anteriores, pero ahora se convierten en supuestos que requieren evaluar su cumplimiento, especialmente cuando existen covariables no observadas.

> "*Un mecanismo de asignación corresponde a un estudio observacional si la forma funcional del mecanismo de asignación es desconocida". "Un mecanismo de asignación es regular si el mecanismo de asignación es individualista, probabilística y sin confusión*" (Imbens & Rubin, 2015, p.41).

Además, se pueden usar los métodos simples de los experimentos aleatorizados clásicos para medir el efecto causal si, tras un diseño del estudio observacional (métodos para mejorar el equilibrio), las distribuciones de las covariables entre los grupos de tratamiento y control se hacen aproximadamente similares (o equilibradas), tal que permitan evaluar el contrafactual. De esta manera, se haría más plausible la ausencia de confusión y, mediante el análisis de sensibilidad podríamos evaluar las implicaciones del incumplimiento de la asignación sin confusión; caso contrario dejaría dudas el cálculo del efecto causal (Imbens & Rubin, 2015, p.41).

Bajo estos supuestos, sí le podemos dar una interpretación causal sin sesgos a los resultados potenciales observados de los grupos de tratamiento y control.

iii. Estudios Observacionales con Mecanismos de asignación irregular. En estos casos, se asume que la asignación al tratamiento no tiene factores de confusión, pero permite que la recepción del tratamiento sí los tenga, porque la probabilidad de recibir el tratamiento (activo o control) depende de los resultados potenciales. Esto se suele presentar en investigaciones sociales donde algún participante decide no seguir un protocolo de recepción del tratamiento. Por esta razón, la comparación entre grupos puede estar sesgada y ser errónea. Para abordar esta situación, se suele requerir condiciones adicionales como la *Restricción de exclusión* (p.528), para

descartar aquellos participantes que no cumplen con el tratamiento, e incluir solo los que reciben y cumplen el tratamiento; asegurando en lo posible de que las diferencias observadas en los resultados se consideren sin factores de confusión. También se utiliza el método de *Análisis de Variable Instrumental* (p.526) en la que se parte del criterio de restricción de exclusión y se selecciona un variable que afecte la probabilidad de recibir el tratamiento que no influya con el resultado generado por este; con el propósito de calcular la Tasa Promedio local (G. W. Imbens, 2014; Wright, 1934; Wrigth, 1929), clasificando los diferentes perfiles de recepción.

2.1.9.Muestras, poblaciones y superpoblaciones

Cabe resaltar que partimos de las observaciones de una unidad de análisis de interés y en la cual se genera una población de unidades observadas. De la unidad de análisis se extraen conclusiones condicionadas a esta población específica sin pretender extraer conclusiones de otras poblaciones. Sin embargo, se pueden estudiar estas unidades observadas como (muestras finitas) aleatoriamente extraídas de una población infinita, y en estos casos se llamará superpoblación a la población infinita (Imbens and Rubin, 2015, p.20).

En otras palabras, la *muestra* es un subconjunto de unidades extraídas de una población infinita con el fin de hacer inferencias sobre la población completa (finita o infinita), pues examinar toda la población sería impráctico y costoso. La *población* es el conjunto completo de unidades sobre las cuales se desea realizar inferencias y pueden ser finitas o infinitas. *Superpoblación* es una concepción de población infinita, de la cual se considera que la población finita es una muestra; es decir, la población finita de interés es a la vez una muestra de una población mucho más grande e incluso infinita (Hartley & Sielken, 1975).

2.1.10. Variabilidad aleatoria

Existen principalmente dos fuentes de variabilidad aleatoria para medir un resultado potencial. La primera es la *variabilidad de la muestra* y la segunda es el *contrafactual no determinístico.* Ambas pueden presentarse simultáneamente. La primera proviene del proceso de muestreo, ya que la muestra seleccionada puede tener proporciones inexactas de las proporciones poblacionales, introduciendo un error en la estimación del resultado potencial (Hernán and Robins, 2020, p.9).

La segunda fuente de error proviene cuando el resultado contrafactual no es determinístico o fijo. Es decir, el resultado contrafactual es producto de un proceso estocástico en el cual la probabilidad de un resultado potencial individual no es 0 ni 1 (como en el determinístico), sino que fluctúa entre estos valores porque depende de una distribución estadística específica del individuo. De esta manera, se introduce otra fuente de error en la estimación del resultado potencial (Hernán and Robins, 2020, p.9). El impacto de este error afecta especialmente en la estimación de los intervalos de confianza.

2.1.11. Causalidad versus asociación

Es muy conocido el aforismo "*Correlación no implica causalidad*" planteado por Reichenbach (1956), porque la causalidad es mucho más que la mera asociación entre variables, y nos recuerda que no se puede inferir relaciones causales solamente a partir de asociaciones estadísticas. Existen múltiples conceptos de asociación estadística (regresión, estimación, pruebas de hipótesis, probabilidad, índice de riesgo, marginación, puntuación de propensión, etc.) que parten de una distribución conjunta de variables observadas afectadas por la variabilidad aleatoria, e intentan inferir asociaciones, probabilidades *a priori o posteriori*, o la actualización de probabilidades ante nuevas evidencias (Pearl, 2009a, 2010). Sus inferencias pretenden responder preguntas acerca de mediciones, pero limitadas a los datos actuales. Para realizar inferencias asociacionales extraen de la población dos subconjuntos disjuntos, a partir del valor del resultado del tratamiento *a posteriori* en una población (Hernán & Robins, 2020, p.11). La asociación puede presentar una correlación estadística de dos aspectos lógicamente inconexos, generando así *correlaciones espurias* (Pearl, 2009a). En estudios

observacionales, la correlación estadística es una condición necesaria pero no suficiente. Así, las probabilidades no distinguen la dependencia estadística de la dependencia causal.

El concepto causal necesita relaciones que van más allá de una simple función de distribución (Pearl, 2009a) porque requiere de supuestos, de intervenciones presentes en la misma población con al menos dos valores posibles. Para hacer inferencias causales se precisa determinar las condiciones bajo las cuales los datos del mundo real se pueden usar para la inferencia causal (Hernán & Robins, 2020, p.12). Por esta razón, el sesgo de confusión no puede detectarse ni corregirse únicamente con métodos estadísticos; porque requiere plantear supuestos causales, no verificables inicialmente en el problema, antes de realizar un ajuste (Pearl, 2009a).

Además, es insuficiente la notación de probabilidades para las relaciones causales. Requiere de una notación matemática causal para expresarlas, con su respectiva sintaxis y semántica (Pearl, 1995, 2000, 2009a).

2.1.12. Estimando Causal y Estimando Estadístico

Little & Lewis (2021) resumen estos conceptos en el contexto de ensayos aleatorizados, mientras que Hernán & Robins (2020), G. W. Imbens & Rubin (2015) y Morgan & Winship (2015) los definen para el contexto de estudios observacionales.

En un estudio causal, el principal objetivo cuando se aplica un tratamiento a una unidad de análisis es extraer conclusiones sobre su efecto. El verdadero efecto se denomina *estimando causal*, el cual se estima a partir de los datos de sus unidades observadas. El *estimando causal es la conceptualización de la cantidad objetivo (target quantity) que el estudio aspira a medir* (Little & Lewis, 2021). Para ello, requiere determinar cómo se puede identificar el efecto a partir de supuestos y metodologías que permitan eliminar o controlar los factores de confusión (sesgo de confusión), los cuales se establecen en el diseño observacional (Morgan & Winship, 2015, p.79). Además, implica un marco teórico sólido para su correcta interpretación con la posibilidad de desarrollar nuevas teorías (p.326,341). La calidad de las estimaciones dependerá de supuestos claramente especificados, que se puedan defender de forma teórica o empírica (p.388). Por tanto, el estimando causal aporta el marco teórico para la selección del estimador causal, y se expresa en términos de los resultados potenciales, covariables, y la asignación del tratamiento (G. W. Imbens & Rubin, 2015, p.19,486).

Un *estimador causal* es una fórmula o algoritmo utilizado para estimar la cantidad objetivo a partir de los datos de las unidades observadas, por ejemplo, una diferencia de medias muestrales (Little & Lewis, 2021). El paso de estimando causal a estimador causal involucra trasladar la conceptualización del estimando causal a la selección de métodos matemáticos y/o estadísticos apropiados, los cuales son considerados en el diseño del estudio. La estadística inferencial aporta estimadores y medidas de precisión de acuerdo a los datos. La *validez interna de un estimador* es la capacidad que tiene para estimar *el efecto causal resumido del tratamiento*, sin violar los supuestos estadísticos de la estrategia de análisis utilizada, porque de lo contrario perdería su capacidad de reflejar el verdadero estimando (Little & Lewis, 2021), por ser vulnerable al sesgo (estadístico). Además, la solidez de los hallazgos depende también del grado en que el análisis estadístico esté basado en los supuestos de inferencia causal, por ejemplo, la ausencia de confusión (Little & Lewis, 2021).

La *estimación causal* es el valor numérico obtenido cuando se aplica el estimador causal a los datos de las unidades observadas y dependerá de la calidad del estimador utilizado (Little & Lewis, 2021).

Un *estimando estadístico* intenta describir y resumir características de un conjunto de datos sin establecer relaciones causales sino asociacionales. Utiliza medidas descriptivas, estimadores, test estadísticos, modelos regresivos, análisis de correlación, entre otras. Los supuestos estadísticos son menos exigentes porque se enfocan en describir la relación entre variables, sin tener en cuenta el marco teórico de estas relaciones ni el control del sesgo de confusión.

2.2. Métodos de evaluación de efectos causales

En estudios observacionales existen dos fuentes que alteran el resultado: el tratamiento y el confusor. Cuando se realizan los cálculos de los efectos causales, es posible que el estimador causal incluya *efectos espurios* a causa de la existencia de factores de confusión o confusores no controlados, que afectan la asignación del tratamiento y producen grupos con diferencias sistemáticas. Los confusores son el principal defecto de los estudios observacionales porque obtienen resultados sesgados.

Los confusores se definen como variables pretratamiento, observadas o no observadas, que afectan tanto a la asignación del tratamiento como a la variable resultado (causa común), produciendo sesgos de selección, inadecuada estimación del contrafactual y, por tanto, efectos espurios.

El sesgo de selección es la alteración que producen las variables confusoras en la asignación del tratamiento generando una inadecuada representación de la distribución del grupo observado y el grupo de interés, haciéndolos no comparables para la medición del efecto causal.

Por tanto, uno de los *problemas esenciales a la hora de realizar la inferencia causal* es realizar un adecuado control de las variables confusoras, también conocido como *ajuste de los factores de confusión*.

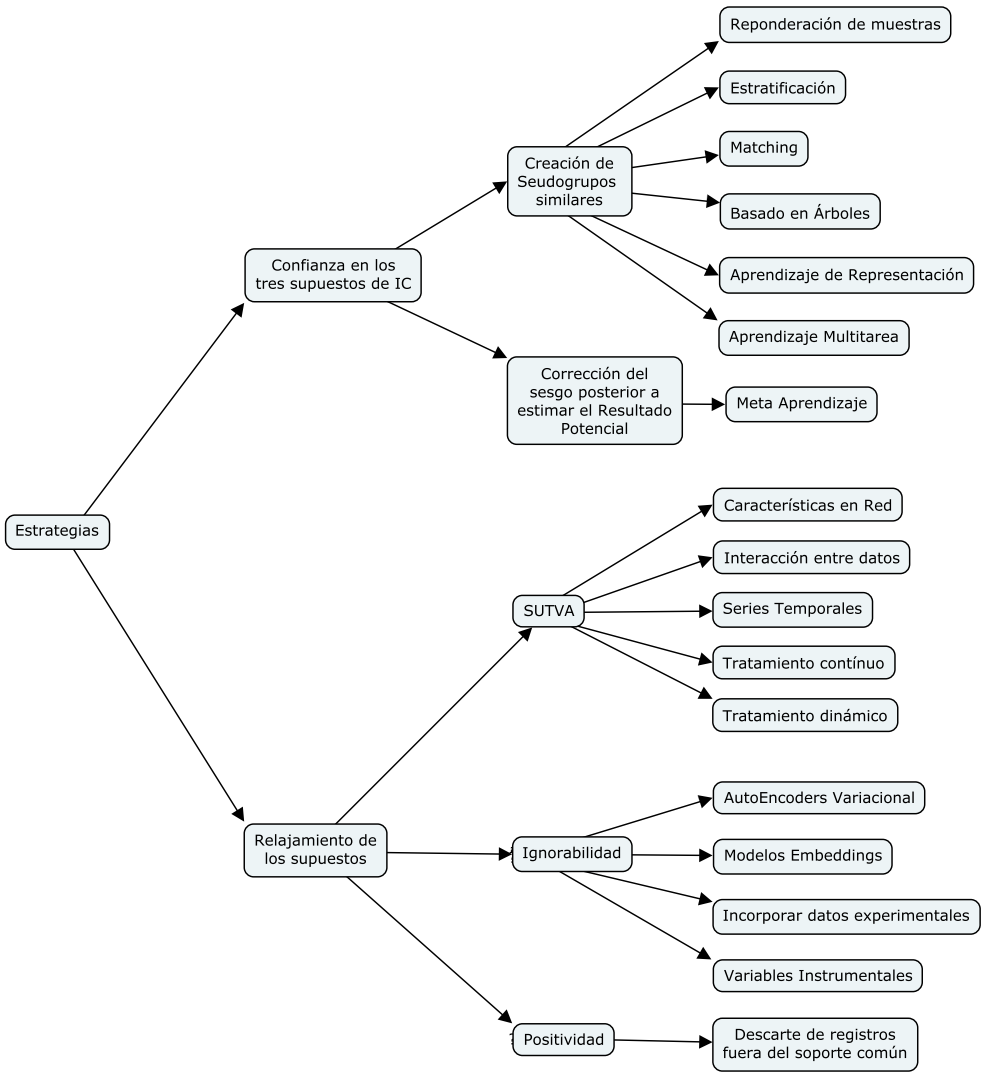

Figura 1 – Estrategias de Ajuste de Confusores

Una vez controlados los confusores observados, los grupos de tratamiento y control ya son comparables. Las distribuciones de ambos grupos son aproximadamente iguales, las unidades previas al tratamiento son intercambiables entre los dos grupos, y por tanto, se los puede tratar como si fuera un diseño experimental aleatorio. Así, cualquier diferencia observada entre los dos grupos posterior a la intervención, podrá ser atribuido al tratamiento. De esta manera, la estimación del efecto ya podrá ser interpretado como causal. Esta comparabilidad se mantiene siempre y cuando no existan confusores no observados que incidan considerablemente en el resultado.

La Figura 1 muestra una taxonomía de las diferentes estrategias de ajuste de confusores, agrupados en métodos que confían en los tres supuestos de Inferencia Causal (Sección 2.1.7), y los métodos que relajan estos supuestos (Sección 2.2.10). Estos métodos se abordarán brevemente en el resto de esta sección, introduciendo previamente los marcadores de propensión.

2.2.1. Marcador de equilibrio y Marcador de propensión (Propensity score)

Existe dos conceptos fundamentales que se utilizan en buena parte de los métodos de reducción del sesgo: el marcador de equilibrio y el marcador de propensión.

El *marcador de equilibrio* es un concepto utilizado para agrupar unidades en los grupos de control y tratamiento con el propósito de realizar comparaciones más significativas (Rosenbaum & Rubin, 1983). El marcador de equilibrio $b(x)$ es una función de las covariables observadas X tal que, la distribución condicional de X dado los tratados en ambos grupos son similares: $X \perp T \mid b(x)$ (Imbens & Rubin, 2015).

El *marcador de propensión*, en adelante *propensity score* (PS), es un tipo especial de marcador de equilibrio muy utilizado. Fue definido por (Rosenbaum & Rubin, 1983) como la probabilidad condicional de asignación de una unidad observada a la exposición de un tratamiento particular, dado su vector de covariables observadas.

$$e(x) = P(T=1 \mid x) \tag{5}$$

Según Morgan & Winship (2015), las principales ventajas del PS en estudios observacional son:

i. Se interpreta como la probabilidad de recibir el tratamiento dadas sus covariables, ya sea a nivel poblacional o para un determinado estrato (p.151-152).
ii. Resume todas las covariables confusoras de una unidad observada en una sola medida probabilística (p.155).
iii. Permite una comparación más cercana a la realidad y facilita la estimación precisa del efecto causal, dados los confusores que influyen en la asignación del tratamiento (p.152).
iv. Simplifica el análisis cuando se tienen un gran número de covariables.
v. Permite la creación de subgrupos y la estratificación, en función de probabilidades similares y, de esta manera, comparar los grupos de tratamiento y control (p.151).

Para muestras grandes o pequeñas, se puede obtener una estimación no sesgada del estimador causal mediante el ajuste por el PS (en lugar de usar el vector de covariables), siendo esto suficiente para remover el sesgo dadas las covariables observadas (Rosenbaum & Rubin, 1983), incluso en un mismo estrato (Rubin & Thomas, 1996). Sin embargo, para confusores no observados, el PS no tiene la capacidad de remover este tipo de sesgo.

El PS se utiliza para reducir el sesgo de selección, equilibrar las covariables en los dos grupos, o ajustar los confusores observados mediante diferentes técnicas: matching (Abadie & Imbens, 2006; Rosenbaum, 1989; Rosenbaum & Rubin, 1985a), subclasificación (Hansen, 2004; Rosenbaum & Rubin, 1984), ponderación (P. R.

Rosenbaum, 1987), regresión, o sus combinaciones (Abadie & Imbens, 2006; Ho et al., 2007 ; D'Agostino, 1998).

El PS, en tratamientos binarios, se estima frecuentemente con una regresión logística, configurando al tratamiento como variable dependiente y a los factores de confusión como variables independientes (Menard, 2002). También, se puede calcular con algoritmos de aprendizaje automático como Random Forest (Breiman, 2001), Support Vector Machines (Vapnik, 2000), Árboles de Decisión CART (Breiman et al., 1998), Metaclasificadores (Khan et al., 2020) y Redes Neuronales (Westreich et al., 2010).

Según Rubin (2004), los PS son herramientas críticas que contribuyen al diseño apropiado de estudios observacionales para hacerlos de manera análoga a los experimentos aleatorios, equilibrándolas tanto como sea posible, sin forzar el equilibrio de las covariables y sin acceder a las variables de resultado.

En los estudios observacionales, el método de PS para controlar los confusores ha sido un estándar (Jović et al., 2015). Antes de implementarlos, es importante considerar cuáles son las covariables que se incorporan en la estimación del PS (Rubin, 2004).

2.2.2.Selección de variables para los modelos de PS

Normalmente, el investigador no tiene todo el conocimiento del fenómeno con el cual está trabajando y se encuentra con el interrogante de seleccionar, en medio de muchas variables, con cuáles trabajar. Este proceso se llama comúnmente "*selección de variables*" o "*feature selection*" (Jović et al., 2015).

En el contexto de los PS, el investigador no conoce con profundidad cómo se realiza la asignación del tratamiento y se encuentra con el mismo interrogante sobre qué variables de pretratamiento escoger. Así, la reducción del sesgo de confusión también dependerá de las variables que se seleccionan en el modelo. Por tanto, una selección cuidadosa de las covariables que participaran en los modelos de PS es esencial. En la Sección 3.3.3.1. resumimos un conjunto de criterios que han sido recomendados por los principales investigadores en este campo.

2.2.3.Métodos de Reponderación (ReWeighting)

El primer método utilizado para controlar los confusores observados y estimar el efecto causal es la reponderación, conocido como *Inverse Propensity Weighting* (IPW) o *Inverse Probability of Treatment Weighting* (IPTW) (Rosenbaum, 1987; Rosenbaum & Rubin, 1983). Consiste en incorporar una ponderación *r* a cada muestra, aplicando esta reponderación a las unidades analizadas:

$$r = \frac{T}{e(x)} + \frac{1-T}{1-e(x)}, \tag{6}$$

Donde T es la variable de asignación de tratamiento y *e(x)* es el PS definido en la ecuación (5). Después de aplicada la reponderación se estima ATE.

Existen varios métodos IPW, siendo los más comunes:

i. La *versión normalizada*, utilizada cuando el PS se obtiene mediante estimaciones (G. W. Imbens, 2004).

ii. El *estimador doblemente robusto* (*IPW aumentado*), cuando existe un error dramático a causa de una especificación incorrecta de los PS (J. M. Robins et al., 1994).

iii. La *combinación de regresión de resultados con ponderación por PS* (*Doubly robust estimation*) para garantizar que las estimaciones sean robustas si existe una especificación incorrecta de los PS (Bang & Robins, 2005; Robins et al., 2007),

iv. La *propensión de equilibrio de covariables* (*Covariate balancing propensity score*) que utiliza los PS como probabilidades y como puntuación de equilibrio para aumentar la solidez, si el modelo de PS tiene una especificación errónea (Imai & Ratkovic, 2014).

v. La extensión del modelo anterior llamada Puntuación de propensión generalizada de equilibrio de covariables *(Covariate balancing propensity score for a continuous treatment)* para minimizar la correlación entre la asignación del tratamiento y las covariables, y reducir el efecto negativo de los PS especificados erróneamente (Fong et al., 2018).

vi. La desventaja cuando valores pequeños de PS se presentan en los denominadores del estimador IPW, se aprecia alrededor de las colas del modelo de regresión logística generando inestabilidad. Para mejorar la precisión, en estos casos se utiliza la *estrategia de regularización por recorte de pesos* (*Weight trimming & propensity score weighting*) después de la regresión logística por debajo de un umbral definido (B. K. Lee et al., 2011). Sin embargo, esta estrategia ha demostrado ser muy sensible a qué tan cerca de cero estén los pesos de probabilidades y qué tan grande sea el umbral de recorte, ya que genera distribuciones asintóticas no gaussianas diferentes al estimador IPW (F. Li et al., 2018).

vii. La robustez se consigue con técnicas de re-muestreo y corrección de sesgo empleadas en la estimación *IPW de robustez bidireccional* (Ma & Wang, 2020), donde los pesos de superposición *h(x)* están entre [0 ; 0.5] para hacerlo menos sensible a los valores extremos de PS, y donde *h(x) es proporcional a 1 – e(x)*, mostrando así, una mínima varianza asintótica entre todas las ponderaciones equilibradas (F. Li et al., 2018).

2.2.4. Métodos de Estratificación

También conocido como método de subclasificación o bloqueo (Imbens & Rubin, 2015). Es utilizado para la reducción de sesgo a causa de la diferencia entre las distribuciones de los grupos de tratamiento y control. El método divide cada grupo en subgrupos utilizando diferentes estrategias para separarlos según sus covariables, ya sea por igualdad de frecuencia (Rosenbaum & Rubin, 1983) o por IPW (Hullsiek & Louis, 2002). Luego, encuentra subgrupos homogéneos de cada grupo tal que puedan considerarse como muestras de un ensayo controlado aleatorizado. Esta homogeneidad de los grupos permite calcular el ATE del estrato y combinarlos de alguna manera para obtener el ATE de la muestra.

$$ATE_{estrato} = \sum_{j=1}^{J} q(j)[\bar{Y}_t(j) - \bar{Y}_c(j)], \tag{7}$$

Donde $\bar{Y}_t(j)$ y $\bar{Y}_c(j)$ son los resultados promedios de los subgrupos tratados y de control, respectivamente, para el estrato *q(j) = N(j)/N*, donde *q(j)* es la proporción del estrato *j*. El sesgo del método es el promedio ponderado del sesgo de cada estrato.

2.2.5. Métodos de Emparejamiento

El método de matching reduce el sesgo mediante el emparejamiento de unidades de cada grupo (tratamiento y control) para estimar el contrafactual. Para esto, utiliza varias estrategias de vecindad más cercana en función de sus covariables. Así, emplea diferentes *métricas de distancia*: Euclídea, Mahalanobis, etc.; o mediante *funciones de transformación*. Existen varias funciones de transformación como las que utilizan PS en cuyo caso

la distancia se calcula como el valor absoluto de la diferencia de los PS o de los *logit* de los PS (Stuart, 2010). Adicional a esta métrica de distancia basada en PS se pueden incluir otras métricas existentes para comparar covariables claves (Rubin & Thomas, 2000; Stuart, 2010).

Existen otras funciones de transformación que utilizan las covariables o los resultados, tal que el espacio transformado preserve más información. Una primera aproximación es el *prognostic score,* estimado bajo el resultado del grupo de control (Hansen, 2008) pero ignora el resultado del grupo de los tratados. La inclusión del criterio de independencia de Hilbert Schmidt supera esta desventaja ya que es estimado usando los resultados de ambos grupos (Chang & Dy, 2017). Otras transformaciones son la *representación equilibrada y no lineal* (Li & Fu, 2017), la *coincidencia aleatoria del vecino más cercano* (S. Li et al., 2016), entre otras.

Una vez elegida la métrica de similitud, el siguiente paso es encontrar el vecino más cercano mediante los algoritmos de matching. Caliendo & Kopeinig (2008) clasifica estos algoritmos en cuatro grupos:

i. Coincidencia del vecino más cercano – NNM (*Nearest Neighbor Matching*),
ii. Con calibrador,
iii. Estratificación y
iv. Kernel.

Los tres primeros grupos pueden utilizar pocas observaciones para calcular el contrafactual. El grupo (*i.*), elige una unidad del grupo de control y lo empareja con una unidad del grupo de tratamiento más cercana, en función de una puntuación de similitud; como por ejemplo el PS, con dos modalidades: con reemplazo y sin reemplazo.

El número de unidades de tratamiento N_T con el número de unidades de control N_C emparejadas reciben la nomenclatura N_C: N_T; y por tanto, 1:1 significará que una unidad tratada se empareja con una unidad de control. Además*, es importante compensar el número de vecinos porque si es grande (pequeño), el sesgo del estimador del efecto del tratamiento puede ser alto (bajo) y la varianza baja (alta)* (Dehejia & Wahba, 1999; H. L. Smith, 1997).

NNM obtiene coincidencias malas si el vecino más cercano está lejos, para lo cual se emplean tolerancias (caliper) para la distancia de los PS que garantice la condición de un soporte común (Lunt, 2014; Q.-Y. Zhao et al., 2021), grupo (*ii*.).

Los *métodos de estratificación* también conocido como coincidencia, bloqueo y subclasificación (Rosenbaum & Rubin, 1985a), grupo (*iii*.), particiona el soporte común del PS en un conjunto de intervalos, y luego toma la diferencia de medias en los resultados entre las observaciones tratadas y de control, para calcular el efecto dentro de cada intervalo.

Los métodos *Kernel*, grupo (*iv*.), en sus variantes de *Kernel Matching*, KM (Heckman et al., 1998) y *Local Linear Matching,* LLM (Heckman et al., 1997), son métodos no paramétricos que utilizan promedios ponderados de las observaciones en el grupo de control para crear el resultado contrafactual. Producen menor varianza, debido a que utilizan más información para crear los resultados contrafactuales, pero KM requiere una elección apropiada del parámetro *ancho de banda* (J. A. Smith & Todd, 2005).

Por su importancia para este trabajo, los métodos de matching lo abordaremos con mayor profundidad en la Sección 3.3.

2.2.6. Métodos Basados en Árboles (Tree-based)

Los árboles de decisión son métodos robustos de aprendizaje supervisado no paramétrico utilizados para clasificación o regresión, dependiendo si la variable target es discreta o continua, respectivamente. Estos

métodos están categorizados como análisis de "*Classification And Regressión Tree*" - *CART* (Breiman et al., 1998). El método consiste en particionar los datos y aplicar un modelo simple de predicción (lineal o polinómica) para cada partición, las cuales se represen gráficamente como un árbol de decisión (Loh, 2011).

Para la estimación del efecto causal basados en CART, utiliza la misma muestra para construir la partición y después estimar el efecto del tratamiento promedio realizando pruebas de hipótesis sobre la diferencia en los efectos (base para dividir en subpoblaciones utilizando magnitudes distintas). Luego, se utiliza otra muestra para calcular los efectos del tratamiento promedio condicionales en cada subpoblación, en lugar de directamente predecir el resultado. Finalmente, construye un solo árbol (cultiva y poda) hasta que el nivel de tolerancia de partición sea alcanzado (Athey & Imbens, 2015).

En los *Bayesian Additive Regression Tree* (BART) no paramétrico, cada árbol tiene un aprendizaje débil y limitado por una regularización previa mediante árboles binarios como base. Cada árbol aporta una pequeña parte del ajuste, tiene su estructura y sus reglas de decisión para acceder a nodos inferiores, y cada nodo contiene la respuesta media de las observaciones que corresponden a este nodo. Sólo requiere ingresar el resultado, el tratamiento asignado y las covariables de confusión, pudiendo manejar una gran cantidad de predictores, generar intervalos de incertidumbres coherentes, modelar relaciones no lineales, modelar iteraciones entre variables, y manejar variables de tratamiento continuo y datos faltantes. Además, puede identificar con facilidad el efecto de tratamientos heterogéneos, y estimar con más precisión el efecto del tratamiento promedio que el método de matching con PS, porque incorpora la técnica Bayesiana que facilita la estimación de la incertidumbre (Hill, 2011).

Random Forest (RF) es otro clasificador que combina árboles predictores (Breiman, 2001). La versión para estimar efectos heterogéneos se basa en el algoritmo de *Random Forest Causal.* Este método permite comparar los bosques causales con métodos de emparejamiento de vecinos más cercanos, siendo más eficaces que los métodos tradicionales porque soporta un gran número de covariables y la cercanía se define con respecto a un árbol de decisión (Wager & Athey, 2018). También puede extenderse a tratamientos uni o multidimensionales, donde cada dimensión puede ser discreta o continua (P. Wang et al., 2015).

2.2.7.Métodos de Aprendizaje de Representación Equilibrada

Uno de los grandes desafíos en inferencia causal es el *problema de adaptación de dominio* que ocurre cuando la distribución de los datos de entrenamiento y de prueba violan el supuesto de la teoría del aprendizaje estadístico que asume que los datos provienen de una misma distribución, dando como resultado distribuciones de prueba solo relacionadas a los datos de entrenamiento, pero no idénticas. Esto trae consigo que las distribuciones de interés no son independientes de las covariables, y como consecuencia, la distribución del contrafactual sea con frecuencia diferente de la distribución factual (Yao et al., 2021).

Para abordar esta situación, algunos autores proponen un método de representaciones de covariables que permita la adaptación de dominio con el fin de: minimizar la diferencia entre los dominios fuente y destino, maximizar el margen del conjunto de entrenamiento (Ben-David et al., 2006), y ajustar una distancia de discrepancia entre las distribuciones mediante una función de pérdida arbitraria (Mansour et al., 2009).

Un método para equilibrar las distribuciones y preservar la similitud local es propuesto con el nombre *Similarity preserved Individual Treatment Effec* (SITE) que tiene sus bases en el *aprendizaje de representación profunda* (Yao et al., 2018; Yao, Li, Li, Huai, et al., 2019). El método *Conditional Treatment-Adversarial learning based Matching* (CTAM) filtra la información de variables cuasi instrumentales al aprender las representaciones, realiza el emparejamiento entre las representaciones aprendidas, y estima el efecto causal del tratamiento (Yao,

Li, Li, Xue, et al., 2019). Sin embargo, los método de matching son más interpretables que los métodos de representación.

2.2.8. Métodos de Aprendizaje Multitarea

El objetivo de estos métodos es controlar los factores de confusión y estimar el efecto causal de forma precisa, realizando ambos pasos de forma simultánea. El modelo de *redes neuronales multitareas profundas* (Alaa et al., 2017) permite inferir efectos causales en datos observacionales utilizando técnicas de *aprendizaje automático* y *técnicas de regularización* basados en puntajes de propensión. El modelo utiliza una red compartida entre resultados potenciales factuales y contrafactuales, y un conjunto de capas idiosincráticas para cada resultado potencial. Para reducir el sesgo de selección, se adelgaza la red mediante una probabilidad de *dropout* (abandono) en función del puntaje de propensión (*Propensity-Dropout*). El entrenamiento se realiza de forma alternada entre los grupos tratados y control. Luego, se actualizan los pesos de las capas compartidas e idiosincráticas para cada grupo. El uso del *Propensity-Dropout* muestra beneficios marginales significativos en la precisión del efecto.

Otra aproximación del aprendizaje multitarea es el método de inferencia bayesiana de efectos de tratamiento individualizado, usando el proceso gaussiano multitarea (Alaa & Van Der Schaar, 2017). Una extensión del modelo multitarea para múltiples tratamientos es el método de aprendizaje que utiliza redes neuronales para la representación contrafactual, con el fin de estimar curvas individuales de dosis-respuesta para cualquier número de tratamiento con parámetros de dosificación continua (Schwab et al., 2020)

2.2.9. Métodos Metaprendizaje

Los métodos de metaprendizaje estiman el efecto promedio condicional del tratamiento (CATE) a partir de algoritmos de aprendizaje automático supervisado, como *Random Forest*, BART, o redes neuronales (Künzel et al., 2019). Estos métodos dividen el problema en dos partes abordadas por separado que son: controlar los factores de confusión y posteriormente estimar de forma precisa el efecto causal.

El estudio de metalerners tales como *T-learner*, *S-learner* y *X-learner* es abordado por Künzel et al. (2019). El más simple de los algoritmos es *T-learner,* y se utiliza cuando los datos permiten una separación clara y precisa entre los grupos tratados y de control. Primero, utiliza un método de regresión para estimar el resultado para los tratados a partir de sus características a nivel individual, y luego hace lo mismo para el grupo de control. Finalmente, estima el efecto causal.

El método *S-learner* evalúa cómo los tratamientos inciden sobre un resultado específico a partir de un indicador de tratamiento, considerado como una variable más. Es relevante en contextos donde el efecto causal promedio condicional es cero en gran parte del espacio de las características. Puede utilizar varios algoritmos como *Random Forest* para hacer las estimaciones y LASSO (Tibshiranit, 1996) para la selección de variables. Su desventaja es el sesgo en situaciones donde el tratamiento es diferente de cero (Künzel et al., 2019).

El método *X-learner* primero evalúa el efecto del tratamiento y control para cada unidad de observación por separado, es decir, se busca su respectivo contrafactual o factual, similar al primer paso del *T-learner*. Luego estima el CATE usando un método de regresión, pero considerando el conocimiento estructural de CATE: linealidad, escasez y suavidad; mediante un estimador adaptativo que aprenda estas características, el cual es importante cuando se acerque a cero o tenga un comportamiento aproximadamente lineal (Künzel et al., 2019).

El *método cuasi-oráculo de estimación del efecto de tratamiento heterogéneo* (Nie & Wager, 2021) tiene dos etapas. En la primera, estima efectos marginales y propensiones de tratamiento. En la segunda, optimiza una función objetivo que se adapta a los datos para aislar el componente causal de la señal y estimar el efecto causal,

basándose en la diferencia de los resultados de la primera etapa. En ambas etapas permite usar cualquier método de minimización de pérdida tales como regresión penalizada, *boosting* o redes neuronales profundas; y ajustadas con técnicas de validación cruzada, lo que le da mayor capacidad de adaptación al contexto de la investigación, flexibilidad y robustez. Además, los métodos *R-learner* y *U-learner*, se implementan en R con la libería "*glmnet"* y estiman el efecto del tratamiento a partir de los PS.

U-learner está sujeto a alta varianza e inestabilidad. Para su control, se vuelve crítico establecer un caliper de 0.05 y minimizar las tasas en la validación cruzada mediante un parámetro de regularización máximo que logre una desviación estándar alejada del mínimo (lambda.1se) (Nie & Wager, 2021).

El método *R-learner* permite determinar cómo varía el efecto del tratamiento condicionado en diferentes poblaciones o individuos a partir de sus covariables con el fin de informar decisiones de tratamiento más precisas y personalizadas (Nie & Wager, 2021).

2.2.10. Métodos con los supuestos relajados

En la Figura 1 se incluyen las diferentes estrategias para el ajuste de confusores cuando es necesario relajar los supuestos de Inferencia Causal.

2.2.10.1. Relajamiento del supuesto SUTVA

Para este caso, las estrategias que se utilizan son las siguientes:

i. Características en red:

Cuando existen confusores no observados o dependencia de datos se requiere relajar este supuesto. Si las características están interconectadas entre sí mediante una estructura de red, existe una afectación entre los datos, como son los casos de las redes sociales, la inmunidad colectiva y el procesamiento de señales (Yao et al., 2021). En los casos de datos en estructura de red, puede utilizarse las redes convolucionales gráficas (R. Guo et al., 2020). Este método, también relaja el supuesto de ignorabilidad y utiliza una *red desconfusora* para aprender la representación de los confusores ocultos. Esto permite identificar y mejorar la capacidad para controlar los sesgos de confusión utilizando la información de la red, lo que consigue una estimación más precisa de los efectos causales individuales (R. Guo et al., 2020).

ii. Interacción entre datos

Cuando existe interacción entre datos, porque algunas unidades de observación afectan los resultados de otras unidades, algunos autores lo abordan desde la identificación y estimación de parámetros causales bajo interferencia (Sherman & Shpitser, 2018).

iii. Series Temporales

Cuando se requiere modelar datos de series temporales que, por sus características intrínsecas no satisface el supuesto (*i.i.d.)*, se suele utilizar el método de regresión, pero es difícil encontrar el modelo correcto por lo que se reemplazan los modelos de regresión por un clasificador (Chikahara & Fujino, 2018). Sin embargo, para la presencia de confusores se ha utilizado un *desconcertador de series temporales* mediante una red neuronal recurrente con salida multitarea y *drop-out* variacional para inferir variables latentes y poder estimar sin sesgos los efectos del tratamiento (Bica et al., 2020).

iv. Tratamiento continuo

Cuando se tiene un tratamiento continuo, se viola el supuesto SUTVA porque no existe una sola versión del tratamiento. En estos casos, se discretiza la variable causal (dosis-tratamiento) para que vuelva a satisfacer el supuesto (Schwab et al., 2020).

v. Tratamiento dinámico

Cuando existe un régimen de tratamiento dinámico con reglas de decisión por cada etapa de intervención (Chakraborty & Moodie, 2013), se estiman estos tratamientos con reglas de decisión dinámicas como el *Q-learning* o *A-learning* mediante ajuste recursivo hacia atrás, como en la programación dinámica (Bather & John, 2000). *Q-learning* estima el régimen de tratamiento dinámico óptimo en cada punto de decisión, dada la información de las unidades. *A-learning* se utiliza para la regresión con el contraste entre tratamientos, y la probabilidad de asignación de tratamiento en cada punto de decisión, dada la información de las unidades.

2.2.10.2. Relajamiento del supuesto de Ignorabilidad

Debido a la dificultad o la insostenibilidad práctica de establecer el supuesto de ignorabilidad indicado en la Sección 2.1.7.2, se han abordado diversas estrategias para relajar este supuesto, pero destacando la importancia de modelar adecuadamente los confusores. Una de ellas utiliza *Variational AutoEncoder* para inferir las relaciones no lineales entre los confusores observados y latentes (Louizos et al., 2017). Para esto, utiliza un modelo basado en redes neuronales profundas que aproxime la distribución conjunta de las variables latentes y observadas, y aproxime las interacciones entre el tratamiento y su efecto.

Otra aproximación reduce el problema a una predicción semi-supervisada de tratamientos y resultados utilizando *Embedding Model* como alternativa a los modelos generativos, con el fin de asignar *Embedding Vectors* a cada unidad y así desacoplar las propiedades de la unidad y la estructura de la red. Además, utiliza una red como proxy para los confusores no observados en la inferencia causal (Veitch et al., 2019).

Existen otros mecanismos como mezclar datos experimentales con observacionales, o utilizar *Variables Instrumentale*s. En el caso de utilizar variables instrumentales, ésta sólo debe afectar a la asignación del tratamiento sin que afecte a la variable resultado (Thomas et al., 2020). De tal manera que, los cambios en la variable instrumental conduzcan a una asignación del tratamiento diferente e independiente de las variables latentes, aíslen el efecto del confusor y midan adecuadamente el efecto causal.

2.2.10.3. Relajamiento del supuesto de Positividad

Este supuesto es más complejo relajarlo cuando se presentan conjunto de datos de alta dimensionalidad en las cuales existen zonas donde no se cumple este supuesto, por una escasa superposición de los grupos, recurriéndose a la poda, o al descartarte de registros sin superposición (Yao et al., 2021).

2.2.11. Error de medición y medición del sesgo

Cuando se realiza un estudio causal es posible que se presente un sesgo de medición cuando hay una asociación entre el tratamiento y el resultado, como consecuencia del proceso de medición de los datos del estudio. Por ello, se vuelve necesario tenerlo en cuenta cuando se interpretan los resultados.

Según Hernán & Robins (2020, p.113), el *error de medición* son las discrepancias entre los valores medidos y sus valores verdaderos, sean estos del tratamiento o de resultado, que aparece durante el proceso de recolección de datos por diversos factores: instrumentales, humanos, etc.

La *medición del sesgo* determina cómo el error de medición afecta sistemáticamente la estimación del efecto, o cómo altera la percepción de la variable tratamiento y el resultado (Hernán & Robins, 2020). El sesgo de medición puede debilitar o reforzar un efecto e incluso puede invertir la dirección percibida de esta asociación. Siempre existe el riesgo de sesgo de medición en cualquier tipo de estudio. Evidentemente, se puede reducir con un cuidadoso proceso de medición, mediante métodos y herramientas más precisas y menos susceptibles a errores. Los diagramas causales permiten representar las fuentes de sesgo o error (nodos), ayudan a visualizar las relaciones confusoras y sus posibles direcciones, pero no miden el sesgo.

Existen varios criterios sobre las fuentes de errores de medición, dada su relación entre el tratamiento y el resultado que se describen a continuación (Hernán & Robins, 2020, p.114-115):

Los errores de medición pueden ser *dependientes* o *independientes,* de acuerdo a la forma cómo se obtuvieron los datos; *dependientes,* cuando el error de medición del tratamiento incide en el error de medición del resultado; e *independientes* cuando el error de medición del tratamiento no incide en el error de medición de resultado, por ejemplo, cuando se cometen errores aleatorios al ingresar datos de tratamiento o de resultados.

Existen cuatro tipos de errores de medición:

i. *Independiente no diferencial*: El error de tratamiento y el error de resultado son independientes entre sí.
ii. *Dependiente no diferencial*: El error de tratamiento y el error de resultado son dependientes, pero no afecta la relación entre las variables tratamiento y resultado.
iii. *Independiente diferencial*: Cuando el valor verdadero de la variable tratamiento incide directamente en el error de medición de la variable resultado.
iv. *Dependiente diferencial*: Cuando la variable tratamiento incide directamente en el error de medición de la variable resultado, o el resultado incide directamente en el error de medición de la variable tratamiento.

Dependiendo del tipo de error se escogen los métodos apropiados para corregirlo. Sin embargo, la mejor forma de evitarlos es mejorando los procedimientos de medición.

2.3. Aplicaciones previas de inferencia causal en la industria

En la Ingeniería de manufactura o de procesos, los estudios causales analizan cómo los cambios en las salidas se deben a intervenciones controladas sobre el sistema o su ambiente (Hund & Schroeder, 2020).

Con el avance tecnológico, las industrias se han vuelto más integradas y complejas, por lo que es vital evaluar la seguridad y confiabilidad de sus procesos, asociadas con la necesidad de detección y diagnóstico de fallos (*Fault Detection and Diagnostic*, FDD).

En este contexto, los fallos ocurren generalmente por el desgaste o envejecimiento de equipos o instrumentos, materia prima con desviaciones en la calidad, cambios en la lógica de control, o errores del operador. Estos fallos pueden propagarse lenta o rápidamente, y pueden generar desperdicio, producir accidentes en la planta, y poner en riesgo al personal y a las máquinas (Zope et al., 2023).

Para detectarlos, requieren procedimientos de FDD para identificar en tiempo real la secuencia del subconjunto de variables impactadas al inicio del fallo (propagación) y determinar la causa raíz (*Root Cause Identification*, RCI).

La RCI es fundamental para detectar el origen del fallo y tomar las acciones correctivas, preventivas, o mantenimientos prescriptivos que brinden fiabilidad a los procesos. Además, la RCI permite reducir los tiempos de paradas no planificados, el consumo de energía y los costos de mantenimientos. Sin embargo, las interacciones entre variables de entrada y de salida, en la gran mayoría de procesos indsutriales, presentan interacciones no lineales que son difíciles de monitorizar, controlar y optimizar (Zope et al., 2023).

En el contexto del Monitoreo Estadístico de Procesos, que involucran disciplinas como la Ingeniería de Sistemas de Procesos y el Control Estadístico de Procesos, la investigación causal se centra, según Chiang et al. (2004), en cuatro aspectos principales:

i. *Detección de perturbaciones o fallos* en presencia de comportamientos anómalos en el proceso. Esta tarea se dificulta cuando existe inundaciones de alarmas con alto riesgos de seguridad. Las inundaciones de alarmas son producidas por anomalías que se propagan a través del proceso, por medio de conexiones de materia, energía o información (Rodrigo et al., 2016).

ii. *Identificación y aislamiento de variables influyentes en los fallos*, para centrarse en la variable monitoreada más relevante en la generación del fallo.

iii. *Diagnóstico de causa raíz (RCI)* para detectar la causa raíz en fases tempranas, evitar que se produzca un comportamiento anómalo en el proceso, y reducir los incidentes en la planta industrial. La causa raíz activa la generación de una secuencia de alarmas.

iv. *Recuperación del proceso,* posterior al fallo y a la RCI.

Estos cuatro aspectos, son el itinerario que siguen la mayoría de los métodos de RCI. La investigación de la RCI se ha abordado desde dos vertientes (Yang & Xiao, 2012):

i. *Métodos basados en mecanismos,* que dependen de la complejidad del proceso industrial y del acceso a información detallada del sistema (conocimiento disponible). Se adaptan a procesos específicos que impiden la reutilización o transferencia de modelos a otros entornos.

ii. *Métodos orientados a datos* (*data-driven*), que utilizan series temporales y se pueden agrupar por el tipo de método, tales como la Causalidad de Granger (GC), la Transferencia de Entropía (TE), las Redes Bayesianas (BN) y los métodos combinados.

2.3.1. Métodos basados en mecanismos (Mechanism driven)

En el ámbito de la gestión de alarmas y detección de fallos, los principales métodos de causalidad basados en mecanismos son:

i. *Los Sistemas dinámicos*: Desde la teoría de control y los sistemas dinámicos, la causalidad ha sido abordada con la intención de predecir y controlar el comportamiento de procesos industriales. Para esto, modelan sistemas de ecuaciones diferenciales del sistema según los tipos de variables (continuas o discretas), las dependencias o no del tiempo, si las relaciones son lineales o no lineales, si llevan o no registros de memoria, o si son procesos determinísticos o estocásticos (Mellodge, 2016).

ii. *El Análisis de Árbol de Fallos (Faul Tree Analysis, FTA):* La raíz del árbol representa el evento no deseado y las ramas son las causas principales. Cada nodo no raíz, relaciona eventos intermedios con sus causas. Sin embargo, se vuelve complejo modelar sistemas grandes y altamente integrados porque requiere un conocimiento detallado del sistema, lo que limita su efectividad y actualización (Fussell, 1975).

iii. *La Matriz de Adyacencias:* Muestra las interconexiones de las diferentes unidades del proceso a partir de la información visual del diagrama de flujo del proceso (Jiang et al., 2009). Para encontrar la causa raíz, el método detecta y diagnostica las oscilaciones que pueden propagarse hacia otras etapas del proceso. El método no utiliza datos históricos, pero es una herramienta poderosa para el diagnóstico de oscilaciones en la planta.

iv. *Los Modelos de Flujos Multinivel* (*Multilevel Flow Models*, MFM): Para encontrar la causa raíz, Dahlstrand (2002) representó gráficamente los flujos de materiales, energía e información, y analizó los cambios que se producen en diferentes niveles del proceso, lo que facilita la identificación de interacciones entre componentes. Sin embargo, no cuantifica la relación causal y requiere de un profundo conocimiento del sistema industrial, dificultando su implementación en sistemas grandes.

v. *Análisis de Similaridad*: Rodrigo et al. (2016) introdujeron el análisis de similaridad de secuencias de alarmas usando datos de varias fuentes, para descartar alarmas irrelevantes (parlanchinas); y así concentrarse en alarmas que se activan en el periodo de inundación de alarmas y en intervalos de tiempos definidos (por ejemplo, 10 minutos). Luego, el método agrupa las alarmas por patrones de secuencias construyendo una matriz de similaridad; identifica las señales y activos asociados; finalmente, encuentra la alarma causal.

Sin embargo, estos métodos carecen de una cuantificación de la incidencia causal, dependen del conocimiento del proceso, e introducen sesgos porque requieren del conocimiento de posibles fallos descartando aquellos desconocidos y que también pueden participar en la generación del efecto. Además, no se apoyan en los datos generados por el proceso de fabricación y, por lo tanto, son difíciles de aplicar en sistemas reales. Más aún, se vuelven inviables en procesos de producción altamente complejos (J. G. Wang et al., 2024).

2.3.2. Métodos orientados a datos (Data-Driven)

Para la implementación de la Industria 4.0, la automatización y la transformación digital se ha realizado progresivamente con la adopción de nuevas tecnologías como el internet de las cosas (IoT), la computación en la nube y el aprendizaje automático. Este fenómeno produce un incremento vertiginoso de datos recopilados de los procesos de producción y han puesto en auge el desarrollo de modelos orientados a datos. Estos modelos se destinan hacia el soporte a la toma de decisiones para reducir costos, aumentar la productividad y la rentabilidad del negocio, y mejorar la calidad de la producción.

Sin embargo, los modelos derivados del aprendizaje automático supervisados son complejos, difíciles de explicar e interpretar, se comportan como una caja negra, están basados sobre asociaciones o correlaciones, no evalúan el contrafactual y no permiten el razonamiento causal (Scholkopf et al., 2021), limitando su uso.

En este contexto, las dos principales estrategias causales que han ganado gran popularidad son aquellos que utilizan los métodos de causalidad de Granger o la Transferencia de Entropía. Estos métodos estiman relaciones causales entre variables, identifican caminos de propagación y detectan la causa raíz de las perturbaciones o fallos de un proceso, con un mínimo conocimiento previo (J. G. Wang et al., 2024; Zope et al., 2023).

La ventaja de estos métodos es que trabajan con datos actuales o históricos de la actividad de los equipos e instrumentos; sin embargo, los mapas causales generados por estos métodos presentan ciclos y retrasos temporales (Zope et al., 2023) que vuelve desafiante identificar estructuras causales.

2.3.2.1. Métodos basados en la Causalidad de Granger

La causalidad de Granger (*Granger Causality*, GC), o *G causalidad,* es un método estadístico econométrico bivariados de influencia causal para determinar si una serie temporal proporciona información útil para predecir otra serie temporal; basado en la predicción de modelos autorregresivos vectoriales (Granger, 1969). Su uso se ha extendido a una amplia gama de disciplinas porque el método y su notación es sencilla de aplicar, que incluso se la ha utiliado en sistemas complejos (Barnett et al., 2009).

Estos modelos se extendieron a múltiples variables y su ventaja es que tienen un bajo coste computacional, son más fáciles de implementar, automatizar e interpretar; y funcionan bien con pocos datos (H. S. Chen et al., 2018). Es más fácil de interpretar porque la GC se basa en una regresión lineal simple, y la definición de los modelos solo requiere especificar dos hiper parámetros: el orden del modelo que minimice el indicador AIC (*Akai Information Criteria*) y el valor de *alfa* para el *F-Test* (Lindner et al., 2019). El análisis de componentes principales (*Principal Component Analysis*, PCA) se utiliza tradicionalmente como método de selección de características (Seyed Alinezhad et al., 2022).

La GC se fundamenta en varios supuestos: la causa se produce antes que su efecto, los valores de las series temporales son continuos y gaussianos, el efecto causal es lineal, las series temporales son estacionarias (Barnett et al., 2009), la frecuencia de muestreo es discreta y no existen variables de confusión no medidas (Shojaie & Fox, 2022). Sin embargo, algunos de estos supuestos rara vez se cumplen (Shojaie & Fox, 2022).

Landman et al. (2014) identifican la ruta (directa o indirecta) de propagación de oscilaciones de un tipo particular de fallo, mediante un algoritmo de búsqueda de trayectorias. A partir de un diagrama de tuberías e instrumentación (*Piping and Instrumentation Diagrams*, P&ID) en Autocad, genera la topología del modelo base representado en un diagrama causal de señales que fluyen por los componentes (equipos, controladores, sensores y válvulas). Luego, se exporta a formato XML (*eXtensible Markup Language*) y en MatLab se convierte este diagrama causal en una matriz de conectividad de relaciones entre componentes. A continuación, aplica el método de GC y los métodos de dominio de frecuencia, con el fin de obtener la matriz de causalidad inicial y el nivel de interacción entre los componentes, respectivamente. Entonces, depura la matriz de causalidad eliminando las entradas que no representan interacciones causales directas, quedándose con la información estructural de la ruta de propagación. Finalmente, el algoritmo de búsqueda utiliza la matriz de conectividad para determinar si existe un camino físico de propagación del fallo y si lo encuentra, verifica si el camino es directo o indirecto. El método facilita lo identificación exitosa de la trayectoria de propagación de fallos, pero mantiene enlaces redundantes que posteriormente se eliminan con el conocimiento del experto.

En procesos por lotes, Fei et al. (2019) proponen un algoritmo de detección de fallas y análisis de causa raíz que combina: *i.* el análisis de componentes de entropía de kernel (*Kernel Entropy Component Analysis, KECA*) para reducir la dimensionalidad y extraer características no lineales a través de un mapeo no lineal con entropía cuadrática de Renyi, *ii.* el cálculo de índices de disimilitud (DISSIM) entre datos de prueba y datos de referencia en condiciones de operación normal (no defectuosas); *iii*. si los datos de prueba del lote no son de una operación normal, se realiza el análisis de GC comparativo para identificar la causa raíz. Para esto, el conjunto de datos de prueba se descompone en segmentos de datos a través de una ventana temporal móvil. Se realiza el análisis de GC en estos segmentos y se obtienen valores de causalidad para cada par de variables. Al comparar la causalidad entre lotes de operaciones normales y defectuosas, identifica la (variable) causa raíz como aquella que tiene el mayor número de causalidades de operaciones defectuosas (no normales). Los resultados muestran que el análisis comparativo de GC funciona eficientemente en el análisis de causa raíz.

J. G. Wang et al. (2024) introducen una técnica de detección de redundancia de variables para mejorar la precisión de la identificación de la causa raíz y de las rutas de propagación. Para esto, se pueden seleccionar las variables candidatas usando PCA (*Principal Component Analysis*), Lasso (*Least Absolute Shrinkage and Selection Operator*), NNG (*NonNegative Garrote*) o Bayesian NNG. Luego, emplea una técnica de detección de redundancia para identificar la proximidad de cada variable a la perturbación de la causa raíz, mediante el índice SOR (*Sum of Redundancy*), que mide cuantitativamente la proximidad de una variable a la perturbación. Posteriormente, a partir de los valores del índice SOR agrupa las variables en capas; y para el análisis causal, utiliza el análisis jerárquico mediante pruebas de GC dentro de cada capa de variables y entre capas. Finalmente, construye el mapa causal. El método es efectivo en detectar la causa raíz, pero en fallos de hardware puede mostrar relaciones no lineales o no estacionarias ya que, en tales casos, las pruebas GC no son apropiadas.

Tabla 1 – Métodos de causalidad de Granger para la detección de fallos y causa raíz

Método	Descripción	Referencias
Spectral Granger Causality	Aplica la Transformadas de Fourier a las variables oscilantes. Reduce la ruta de propagación.	(Yuan & Qin, 2014)
No lineal y no estacionario	Regresión de Proceso Gaussiano	(H. S. Chen et al., 2018)
Disimilaridad no lineal	KECA para reducir la dimensionalidad, DISSIM para comparar datos de prueba con datos de operación normal, y GC en datos de operación defectuosa.	(Fei et al., 2019)
Grouping Multivariate Granger	Variables oscilantes: Multivariate non-linear chirp mode decomposition.	(Q. Chen et al., 2021)
Análisis jerárquico de GC	Utiliza el índice de suma de redundancia SOR para identificar la proximidad de cada variable a la perturbación, agrupa las variables en capas y realiza el análisis jerárquico con pruebas de GC.	(J. G. Wang et al., 2024)

Lindner et al. (2019) realizaron un análisis comparativo entre GC y TE. Encontraron que el modelo lineal obtenido de GC, puede no representar adecuadamente las relaciones entre las variables del proceso; la no estacionariedad puede detectar conexiones causales espurias, y puede omitir relaciones causales reales. Además, el método GC presenta mayor número de conexiones espurias que TE, lo que afecta la precisión del análisis y dificulta la identificación de trayectorias de propagación de fallos. Sugiere que, para descartar conexiones causales espurias, es preciso validarlas con el conocimiento previo del proceso. Una forma de evitar estas limitaciones es combinar la GC con otros métodos.

2.3.2.2. Métodos basados en la Transferencia de Entropía

La TE mide la cantidad de información transferida, de forma dirigida y asimétrica, entre dos series temporales interdependientes y proporciona información de pronóstico sobre la variable objetivo (Schreiber, 2000). Esta medida representa la influencia causal, pero no distingue si dicha influencia se produce a través de una vía directa o indirecta. La TE se utiliza para el análisis de causalidad de procesos industriales complejos porque permite determinar la dirección y la magnitud de la transferencia de información.

Para reemplazar las probabilidades en el cálculo de TE, se requiere estimar la Función de Densidad de Probabilidad (*Probability Density Function, PDF*) conjunta de dos variables mediante Estimadores Kernel de series discretizadas, pues ofrece robustez y precisión (Bauer et al., 2007). Además, compara la transferencia de entropía (en una u otra dirección), establece una medida causal, y fija un umbral para establecer que los valores por encima de este umbral puedan considerarse significativos. El nivel de significación de esta medida (seis sigmas, se considera robusta) se calcula utilizando simulaciones Monte Carlo de datos subrogados (Bauer et al., 2007). Las medidas de causalidad de TE se trasladan a una matriz de causalidad, se reagrupan las variables en el orden de ocurrencia del evento y se procede a construir el mapa causal (Bauer et al., 2007). Con el fin de reducir la carga computacional, los datos subrogados pueden generarse mediante datos aleatorizados o mediante la *Transformada de Fourier interactiva de amplitud-ajustada* (Duan et al., 2013).

Duan et al. (2013) proponen el método de Transferencia Directa de Entropía (*Direct Transfer Entropy, DTE*) mediante a eliminación de los efectos de las variables intermedias, lo cual abre las puertas para construir una topología precisa del modelo. Posteriormente, mejora la precisión del método DTE introduciendo el método D0TE en el que no utiliza PDF ni probabilidades, se aplica a datos continuos o discretos, pero tiene un alto consumo computacional (Duan Ping et al., 2015).

C. Guo et al. (2015) proponen aplicar la TE sobre las tendencias de series temporales estacionarias con el fin de reducir la influencia del ruido. El método Transferencia de Entropía de Tendencia (*Trend Transfer Entropy, TTE*) utiliza una serie simbólica de tendencias, para lo cual discretiza la serie continua en segmentos isométricos, según una escala de compresión (longitud del segmento). Esto reduce la carga computacional, porque se reduce la cantidad de datos a procesar, y disminuye la presencia de ruido. Sin embargo, su precisión depende de la longitud de la serie.

Lindner et al. (2019) aportan un flujo de trabajo para detectar fallos utilizando el método de TE. En este flujo de trabajo, primero identifica la presencia de oscilaciones mediante el análisis espectral usando la Transformada Rápida de Fourier, encuentra las frecuencias pico de oscilación y determina las variables candidatas que comparten esta oscilación. Luego, extrae los tiempos de retardo entre las variables candidatas para obtener los parámetros del modelo de TE. Se analiza la TE entre cada par de variables y se grafican los enlaces resultantes.

Wen et al. (2022) proponen el método de transferencia de entropía condicional simbólica (SCTE) basado en gráficos de control. Primero, selecciona un conjunto de variables relevantes candidatas usando una herramienta de aislamiento de fallos para reducir la carga computacional del método de transferencia de entropía y hacer más fiable la estimación de la función de densidad. Discretiza y captura la dinámica de la serie mediante la simbolización basada en los límites sigma de los gráficos de promedios móvil de ponderación exponencial (EWMA). Luego realiza el cálculo de la transferencia de entropía condicional en los datos del proceso simbolizados para identificar relaciones causales. La información de causalidad identificada se visualiza en un mapa causal de forma simplificada y fáciles de interpretar. El rendimiento del método se cuantifica con los índices de sensibilidad, especificidad y densidad, y frecuentemente proporciona un buen equilibrio entre los tres índices. Es muy útil para analizar la propagación de perturbaciones porque son más precisos, detectan mejor las

relaciones causales y evitan falsas alarmas. Además, identifica adecuadamente las variables más próximas a la causa raíz de la perturbación del proceso.

Zope et al. (2023) propone un algoritmo de identificación de trayectorias de recorridos de fallas y causa raíz (FTRCI) a partir de mapas causales. Primero, calcula la transferencia de entropía temporal multivariada con lo cual puede capturar relaciones no lineales e incorporar desfases temporales entre las variables. Luego, aprende las relaciones causales entre estados de las variables de falla sin conocimiento previo del proceso y las representa en un mapa causal. Finalmente, el algoritmo FTRCI rastrea las trayectorias de recorrido de fallas dentro del mapa causal e identifica las variables de causa raíz. El algoritmo FTRCI identifica con precisión las trayectorias de recorrido de fallas y aísla de manera efectiva las variables de causa raíz. Funciona bien para procesos dinámicos no lineales.

Los métodos TE tienen la ventaja de detectar relaciones causales lineales o no lineales entre variables de procesos industriales complejos, son más precisos y generalizables, ofrecen una mejor interpretación visual, y presentan menos enlaces falsos positivos (Lindner et al., 2019). La desventaja es el alto coste computacional debido a la prueba de significancia. Otras desventajas de TE y GC es la incapacidad de determinar la presencia de variables ocultas (no medidas). Las principales aplicaciones de TE en el contexto de FDD y RCI se resumen en la Tabla 2.

Tabla 2 – Métodos Data-Driven usando TE y en combinación con GC y BN

	Método	**Descripción**	**Referencias**
TE	*Direct Transfer Entropy (DTE)*	Detecta causalidades espurias y trayectorias directas o indirectas.	(Duan et al., 2013)
TE	*Direct Transfer zero Entropy (DT0E)*	No asume PDF. Mejora la detección de causalidad directas. Mejora la precisión. Se aplican a datos discretos o continuos.	(Duan Ping et al., 2015)
TE	*Trend Transfer Entropy (TTE)*	Aplica la TE sobre segmentos isométricos de la serie para eliminar el ruido y reducir la carga computacional.	(C. Guo et al., 2015)
GC, TE	*Procesos dinámicos no estacionarias*	Detecta la causa raíz en procesos cuasi estacionarios o no estacionarios. Detecta fallos con DPCA y selecciona variables candidatas con RBC. Aplica GC a procesos cuasi estacionarios y TE para no estacionarios basados en DTW.	(G. Li et al., 2016)
TE, BN	*Active Dynamic Transfer Entropy Bayesian Network*	Aprendizaje de BN multibloque basado en la transferencia de entropía dinámica promedio (ADTE) y el algoritmo de aprendizaje codicioso.	(Luo et al., 2020)
TE	*Symbolic Conditional Transfer Entropy (SCTE)*	Aísla las variables relevantes candidatas de fallas, discretiza la serie y la simboliza basada en los límites sigma de los gráficos de promedio móvil de ponderación exponencial. Calcula la TE en las series simbolizadas e identifica las relaciones causales.	(Wen et al., 2022)
TE, BN	*MultiBlock Transfer Entropy (as scoring) and Bayesian Network*	Aprendizaje de BN multibloque utilizando puntuación basada en transferencia de entropía directa y búsqueda codiciosa. Utiliza principios de fusión para unir redes de los segmentos y descubrir ciclos, mejorando la precisión de la red causal.	(Kumari et al., 2022)
TE	*Fault Traversal and Root Cause Identification (FTRCI)*	Calcula la TE temporal multivariada, aprende relaciones causales sin conocimiento previo del proceso. Las representa en un mapa causal y ejecuta FTRCI para identificar las trayectorias de falla y aíslar las variables de causa raíz.	(Zope et al., 2023)

2.3.2.3. Métodos basados en Redes Bayesianas

Yu & Rashid (2013) proponen un método para la monitorización de procesos en red, la identificación de trayectorias de propagación de fallos y el diagnóstico de causa raíz. Combinan el conocimiento previo del proceso y los datos históricos para el diseño de una red temporal bayesiana dinámica (*Dynamic Bayesian Network, DBN*). La estimación de los parámetros del modelo que captura las interacciones entre las variables y las relaciones causales, se realizan mediante probabilidades de transición entre variables e índices de probabilidad bayesiana. El método ha mostrado ser efectivo en la detección temprana de fallos y el diagnóstico de la causa raíz, pero requiere del conocimiento previo del proceso, la complejidad computacional para calcular las probabilidades condicionales y maximizar funciones de verosimilitud.

Amin et al. (2021) presentan una metodología que integra el análisis de componentes principales y la red bayesiana. Aplica el método *Correlation Dimension* para reducir la dimensionalidad de los datos, selecciona los componentes principales y elimina la necesidad de insertar la opinión del usuario durante la construcción del modelo PCA. Luego, utiliza la divergencia de *Kullback-Leibler* para el aprendizaje de la topología de BN y la *Teoría de Cópulas* para la estimación de sus parámetros. El método captura la dependencia no lineal de datos de proceso de alta dimensionalidad. Este método es orientado a datos y no requiere el juicio de un experto, proporciona menos alarmas falsas, facilita la detección temprana de fallos, el diagnóstico de causa raíz es preciso, y no requiere la discretización de los datos para construir las tablas de probabilidad condicional. El diagnóstico preciso ahorra costos de mantenimiento y mejora la integridad de los activos. Sin embargo, el método no ofrece un rendimiento óptimo cuando los datos presentan una extrema no linealidad y no gaussianidad.

2.3.2.4. Métodos basados en el Modelo Causal Estructural

En el contexto del análisis de fiabilidad para cuantificar la incertidumbre del funcionamiento adecuado de un sistema a lo largo del tiempo, Hund and Schroeder (2020) utilizaron el juicio de expertos para identificar la estructura causal de tres variables, especificaron un modelo estadístico lineal para los datos gaussianos con ruido aleatorio $N(0,\sigma)$ y evaluaron las violaciones de los supuestos con estudios de sensibilidad y especificación de parámetros. Sin embargo, al ser un modelo sencillo dificulta su aplicabilidad en entornos complejos.

2.3.2.5. Métodos combinados

Causalidad de Granger y Transferencia de Entropía

G. Li et al. (2016) combinan la TE y GC para identificar la causa raíz cuando se detecta un fallo en procesos dinámicos no estacionarios o cuasi estacionarios, respectivamente. Para ello utiliza un modelo DPCA (*Dynamic PCA*) para la detección de fallos mediante correlaciones entre variables, y autocorrelaciones entre variables retrasadas (*lags*), identificando los componentes principales y calculando un índice de detección de fallos. Luego, se identifican las variables candidatas que contribuyen al fallo basado en el análisis de trazabilidad con el método Multidireccional RBC (*Reconstruction Based Contribution*). Así, emplea el análisis de causalidad de Granger para localizar la causa raíz de fallos que conducen a procesos cuasi-estacionarios, Para fallos que conducen a procesos defectuosos no estacionarios, propone un índice de causalidad (*Dynamic Causality Index, DCI*), basado en el agrupamiento de series temporales (*k-means*) según la distancia de alineación temporal dinámica (*Dynamic Time Warping, DTW*), y de este modo comparar series temporales e investigar su relación causal.

Transferencia de Entropía y Redes Bayesianas

Luo et al. (2020) proponen la construcción de una BN basada en la transferencia de entropía dinámica activa (*Active Dynamic Transfer Entropy, ADTE*) y el algoritmo de aprendizaje codicioso. ADTE permite identificar trayectorias de propagación de alarmas relacionadas causalmente. El algoritmo identifica las relaciones causales, pero requiere del conocimiento de un experto y de un importante costo computacional.

Kumari et al. (2022) proponen un método de aprendizaje de BN multibloque utilizando la Transferencia de Entropía Directa, DTE (Duan et al., 2013). A partir del conocimiento del proceso (diagrama de flujo del proceso), se segmenta en múltiples bloques pequeños tomando en cuenta las variables compartidas entre los bloques. Para cada bloque, aprende una BN acíclica utilizando la puntuación basada en DTE (para medir la fortaleza de la relación causal) y la búsqueda codiciosa. Luego fusiona las BN de cada segmento, en base a principios de fusión, considerando la variable compartida entre los bloques. Esto permite descubrir estructuras secuenciales, colisionadoras y cíclicas; y de esta manera construir un diagrama causal que incluya ciclos. El método demostró un alto rendimiento en descubrir bucles cíclicos, obteniendo una red causal precisa.

Causalidad usando Autoencoder variacional ortogonal autoatento

Bi & Zhao (2021) proponen un modelo de autoencoder variacional ortogonal autoatento (*Orthogonal Self-Attentive Variational Autoencoder, OSAVA*) que consta de dos componentes: atención ortogonal y autoencoder autoatento variacional (*Variational Self-Attentive Autoencoder, VSAE*). El primero extrae las correlaciones entre diferentes variables y la dependencia temporal entre diferentes pasos de tiempo, de tal forma que extrae las relaciones causales entre múltiples variables del proceso. El segundo se entrena para detectar fallos a través de un método que emplea *mecanismos de autoatención*, agrega información a lo largo de todos los pasos de tiempo y reconstruye la salida de OA (*Orthogonal Attention*). Según los autores, el método detecta de manera efectiva los fallos con un bajo retardo en la detección y proporciona resultados interpretables con una tasa prometedora de detección de fallos.

Hasta donde sabemos, no existe un estudio comparativo de los métodos orientados a datos descritos en esta sección que permita establecer cuál es el mejor, cuándo es recomendable usarlo y bajo qué criterios o circunstancias.