# Capítulo 4

# Aprendizaje Causal y su Aplicación a Textos

## Resumen

El creciente volumen de información textual disponible abre nuevas posibilidades para el análisis de diferentes episodios del mundo real (análisis del precedente, desarrollo y secuelas de una crisis o guerra, entre otros). La extracción de variables relevantes a estos episodios, y su posterior vinculación con relaciones causa-efecto, son de gran interés para permitir a los expertos que están analizando el dominio, entender o explicar los diferentes eventos que acontecieron durante ese periodo, o incluso asistirlos en la predicción de posibles desenlaces en episodios en curso. El presente capítulo constituye la etapa final del trabajo de extracción de relaciones causales a partir de medios de noticias digitales. Dicho trabajo se divide en dos grandes etapas: (i) la extracción de variables relevantes al episodio a partir de los textos de artículos periodísticos y (ii) el aprendizaje de la estructura causal a partir de las variables extraídas en (i). En capítulos previos se examina la tarea (i) desde dos posibles ópticas: extracción de términos relevantes y extracción de eventos relevantes. Para estas tareas se utiliza una técnica de pesaje de términos ($\text{FDD}_\beta$) y un modelo predictivo para detectar eventos en curso, respectivamente. El presente capítulo ofrece un extenso análisis comparativo de diferentes técnicas de aprendizaje de estructuras causales para su posible aplicación a la tarea (ii). Las contribuciones principales de este capítulo son: (a) la formulación completa del marco de trabajo (*framework*) para obtener estructuras causales para expertos a partir de artículos periodísticos (pasos (i) y (ii));

(b) la presentación de un extenso análisis comparativo de técnicas de aprendizaje de estructura causal en series de tiempo que compara 9 técnicas del estado del arte en 64 conjuntos de datos sintéticos y un conjunto de datos reales sobre demanda eléctrica en el Gran Buenos Aires (GBA) (c) la presentación de un caso de estudio de la aplicación del *framework* completo a texto. El análisis comparativo presentado (b) es el primero que compara nueve técnicas de distintas áreas de las ciencias (computación, econometría, sistemas complejos), permitiendo sacar conclusiones originales sobre los métodos y su aplicación a diferentes conjuntos de datos. Por otra parte, el *framework* presentado (a) y evaluado (c) es el primero que combina tantas herramientas diferentes para resolver distintos aspectos de la tarea global y que muestra resultados prometedores para continuar en esa dirección para la elaboración de estructuras causales a partir de textos que serán de gran interés para expertos que quieren entender un dominio o escenario.

## 4.1. Introducción

La inferencia de la existencia y cuantificación de efectos causales detrás de fenómenos observados es uno de los focos principales de muchos de los esfuerzos científicos [RBB+19]. Por ejemplo, durante muchos años se estudió el efecto causal entre fumar cigarrillos y el desarrollo de cáncer de pulmón en el individuo fumador [DH50]. De manera similar, a mediados del 1700 James Lind delineó el primer experimento aleatorizado (*randomized experiment*) para develar causa-efecto entre el consumo de cítricos y la recuperación del escorbuto [Lin57]. A mediados del 1800, John Snow descubrió que el agua contaminada con materia fecal causaba el cólera [Sno56].

Durante muchos años las herramientas para sacar conclusiones a partir de datos observados eran principalmente estadísticas, con poco desarrollo de perspectivas o formalismos causales. De acuerdo a Pearl el debate sobre si fumar causa cáncer de pulmón (que se extendió desde 1950 hasta 1964) podría haber sido más corto si los científicos hubieran tenido disponible una teoría de causalidad más formal [PM18]. Las herramientas estadísticas típicamente usadas no permiten sacar conclusiones causales (correlación no implica causalidad). Esto transforma la inferencia y cuantificación de efectos causales en un problema mucho más difícil sin una teoría formal de casualidad. Por ejemplo, la correlación entre fumar y el desarrollo de cáncer se podía medir de los datos sin mucha ambigüedad y sin embargo esto no implicaba causalidad. Uno de los argumentos más importantes en

contra de la teoría de que fumar causa cáncer era la posible existencia de factores no medibles que causen deseo por fumar y cáncer de pulmón, sugiriendo que el fumar no causaba cáncer, sino que estaban correlacionados por una causa en común.

La importancia de poder responder este tipo de preguntas, que se encuentran en el centro de los esfuerzos científicos dio lugar a un creciente interés por definir herramientas y formalismos para poder determinar y medir efectos causales. Múltiples marcos de trabajo para inferencia y razonamiento causal han surgido desde entonces [Pea09, PJS17, SGSH00]. Estos esfuerzos se pueden dividir en dos grandes categorías: (i) inferencia causal y (ii) razonamiento causal. El primero pretende partir de datos e inferir el modelo causal que dio origen a los datos, mientras que el segundo parte de un modelo causal ya definido y lo utiliza para contestar preguntas de razonamiento causal (o incluso preguntas de índole estadístico).

Para entender la diferencias entre un modelo causal y un modelo puramente estadístico hay que entender cómo se relacionan entre sí y qué preguntas permite contestar uno y cuáles el otro. Un resumen de estas relaciones se puede ver en la Figura 4.1 adaptada de [PJS17]. Un aspecto importante de la relación entre estos dos, es que el modelo causal subsume al estadístico, pero no al revés. Esto es, el modelo causal provee un entendimiento mayor del proceso generativo de los datos de lo que un modelo puramente estadístico puede proveer. Con el modelo causal se pueden contestar las mismas preguntas que con el modelo estadístico y potencialmente algunas más. En específico, mientras que ambos modelos pueden contestar preguntas observacionales, solo el modelo causal permite potencialmente contestar preguntas sobre intervenciones y preguntas contrafactuales.

Las preguntas intervencionales son cruciales, por ejemplo, para la creación de políticas y toma de decisiones sobre tratamientos médicos, y son inherentemente distintas a las preguntas observacionales. Por ejemplo, consideremos que se tiene una base de datos de pacientes que recibieron dos posibles tratamientos, A o B (siendo el A mejor que el B). La pregunta de cuál es la probabilidad de recuperarse dado que un paciente recibió el tratamiento A, no es lo mismo que preguntar cuáles son las probabilidades de recuperarse dado que efectivamente se intervino en el sistema y se le dio el tratamiento A a un paciente. El tratamiento A en los datos recolectados puede tener una mala proporción de recuperados versus no recuperados simplemente porque los médicos que atendieron a esos pacientes asignaron el tratamiento A (que es mejor que el B) a todos los pacientes difíciles (porque eran los que más necesitaban el tratamiento A). Por otra parte, intervenir

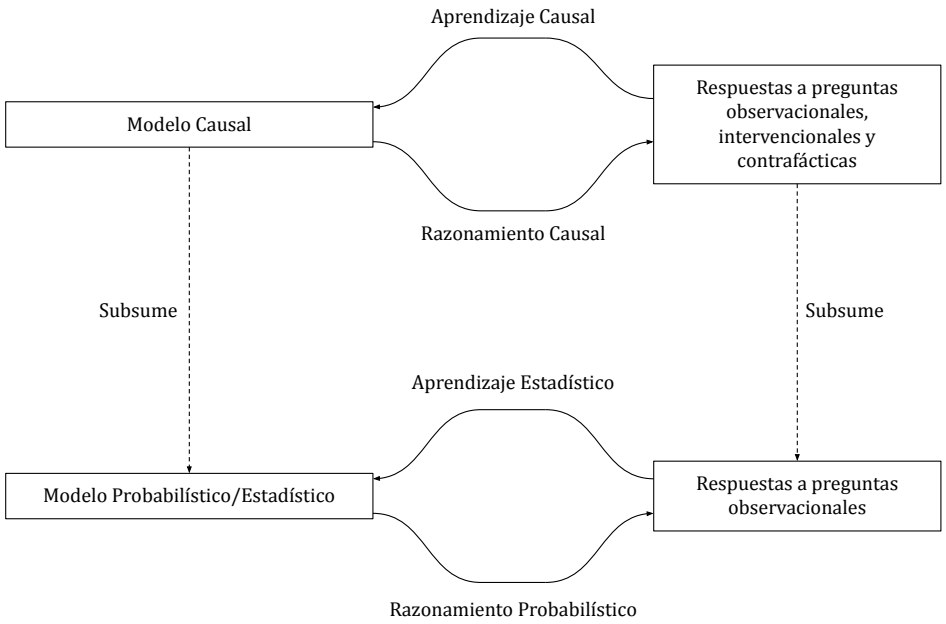



Figura 4.1: Diferencias y relaciones entre un modelo causal y un modelo estadístico/probabilístico, Figura adaptada de [PJS17]. En la parte superior se puede ver cómo información observacional, intervencional y contrafáctica puede ser usada para aprender un modelo casual, o un modelo causal puede usarse para consultar esta información. A esto se lo llama aprendizaje causal y razonamiento causal, respectivamente. Análogamente en la parte inferior se puede observar que de datos observados se puede construir un modelo estadístico/probabilístico, o si ya se cuenta con dicho modelo se lo puede usar para responder preguntas observacionales. A esto lo llamamos aprendizaje estadístico y razonamiento probabilístico, respectivamente.

en el sistema y asignar el tratamiento A a un paciente no hace que sea automáticamente un caso difícil, ya que se lo asignó independientemente de todas las demás variables del sistema (o en otras palabras, de manera aleatoria). Entonces en términos formales $P(R|T = A) < P(R|do(T = A))$, donde $P(R|T = A)$ es la probabilidad de recuperarse dado que se observa que el paciente recibió el tratamiento A, y $P(R|do(T = A))$ es la probabilidad de recuperarse de un paciente dado que se interviene en el sistema y se lo asigna al tratamiento A.

De manera similar, entender los mecanismos causales nos permite entender cuándo las preguntas observacionales y las intervencionales tienen el mismo o distinto resultado. Por ejemplo dado el gráfico de dispersión de la Figura 4.2, se puede sospechar de una relación lineal entre X e Y y realizar una regresión lineal. Dicha regresión nos daría una forma de estimar el valor de Y dado el valor observado de X. Sin embargo, la regresión lineal no nos permite estimar el efecto de una intervención en X sin conocer el modelo

causal. Si X causa Y en el sentido que el cómputo del valor de Y está afectado por el valor de X (por ejemplo cómo está modelado en la ecuación 4.1), entonces el efecto de una intervención se puede medir con el uso de la regresión lineal. Sin embargo si el verdadero modelo generativo es el descrito por la ecuación 4.2, en ese caso al intervenir en X el valor de Y no cambia (en este modelo Y causa a X, y no viceversa). Por ende, la regresión de Y en función de X no serviría para estimar el efecto de una intervención.

En resumen, los datos observacionales presentados en la Figura 4.2 permiten determinar la distribución probabilística observada de ambas variables (la cual está descrita en la ecuación 4.3). Sin embargo, esos datos y esa distribución pueden ser generados por distintos modelos generativos de datos (por ejemplo modelos 4.1 y 4.2), y es importante recuperar esta información para poder recuperar el efecto de una intervención. No siempre es posible determinar el verdadero modelo generativo de los datos a partir de un conjunto de datos observacionales. Ese es uno de los grandes desafíos del área de inferencia causal y muchas veces requiere de supuestos adicionales. Por otro lado, las preguntas contrafácticas son aquellas que consultan sobre mundos posibles que no ocurrieron. Dado el estado actual del mundo, ¿Tomás no se habría recuperado si no le hubieran dado el tratamiento A? Al igual que para las preguntas intervencionales, con datos puramente observacionales no siempre es posible recuperar el modelo causal necesario para responder este tipo de preguntas.

$$\begin{cases} X := N_X & N_X \sim \mathcal{N}(0;1) \\ Y := 2X + N_Y & N_Y \sim \mathcal{N}(0;1) \end{cases} \tag{4.1}$$

$$\begin{cases} Y := N'_Y & N'_Y \sim \mathcal{N}(0;\sqrt{5}) \\ X := 0,4Y + N'_X & N'_X \sim \mathcal{N}(0;\sqrt{0,2}) \end{cases} \tag{4.2}$$

$$P(X,Y) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix} \; ; \begin{pmatrix}1 & 2\\2 & 5\end{pmatrix}\right) \tag{4.3}$$

Una vez obtenido el modelo causal (ya sea de manera automática o construido por un experto), este tiene la capacidad de explicar las relaciones entre las variables y poder potencialmente responder preguntas observacionales, intervencionales y contrafácticas. Esto hace que se genere un gran interés por parte de la comunidad para diseñar técnicas para aprender estos modelos de manera automática a partir de los datos. Adicionalmente, estos modelos causales permitirían la construcción de modelos de aprendizaje automático

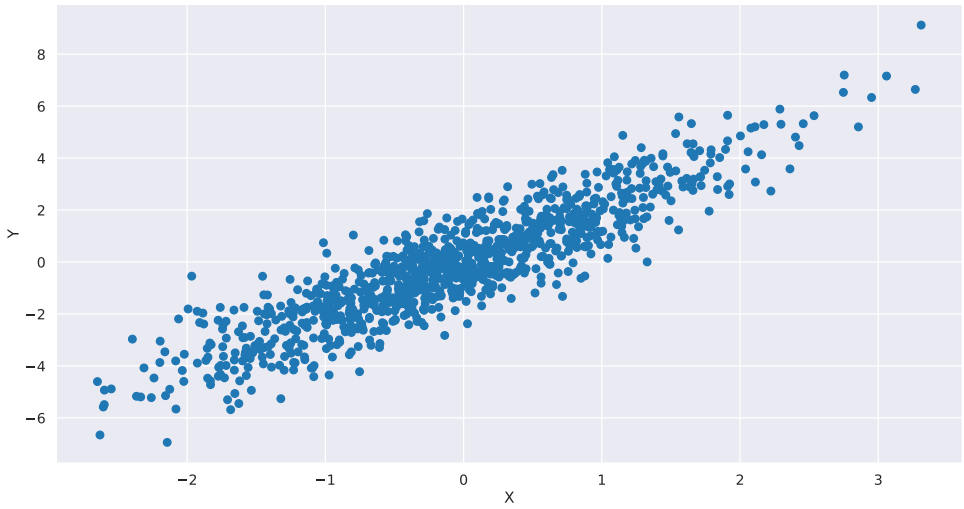



Figura 4.2: Gráfico de dispersión que muestra un conjunto de datos observacionales con distribución probabilística descripta por la ecuación 4.3. Los datos pueden haber sido generados con los procesos generativos 4.1 o 4.2. Sin información adicional no es posible determinar cuál de los dos es el proceso generativo real de los datos y por ende no es posible determinar si $X \to Y$ o si $Y \to Y$. Más aún, si no se cuenta con el supuesto de que todas las causas en común son conocidas (*causal sufficiency*) entonces puede haber una tercera variable que cause a ambas variables (*confounder*) y que no exista causalidad directa de $X$ a $Y$ ni de $Y$ a $X$. Los supuestos e información adicional sobre el proceso generativo sobre los datos a veces son fundamentales para poder determinar si existe causalidad y la dirección de la misma.

(*machine learning*) más robustos ante cambios en la distribución de los datos que pueden aparecer en problemas del mundo real [SLB+21]. Esto es, en los modelos de *machine learning* tradicional se asume que los datos están independientemente e idénticamente distribuidos (i.i.d.), esto es, en general no están pensados para predecir bajo cambios en la distribución (intervenciones). Por esto es que aprender modelos de *machine learning* con perspectiva causal permitiría modelar explícitamente las intervenciones posibles y hacer sistemas de predicción más robustos ante cambios en la distribución que ocurren normalmente en problemas del mundo real. Más aún, la construcción de modelos predictivos basados en estructuras causales permitiría aprender mecanismos causales independientes transferibles entre modelos, permitiendo así la transferencia de aprendizaje (*transfer learning*) entre diferentes tareas abordadas con técnicas de *machine learning* [SLB+21].

Los datos a partir de los cuales se pretende recuperar el modelo causal pueden tener distintas características, y dependiendo de esas características involucrar diferentes dificultades para la recuperación de dicho modelo causal. Por ejemplo, los datos pueden ser de corte transversal, o de series de tiempo. Los datos también pueden ser puramente obser-

vacionales, pueden contener intervenciones conocidas obtenidas a través de experimentos controlados [KSv+14, SPP+05] o pueden ser conjuntos de datos sintéticos con intervenciones [CY13]. Por otro lado, en algunos trabajos [EM07] se extraen estructuras causales a partir de una combinación de datos observacionales e intervencionales con intervenciones desconocidas [SPP+05].

Este capítulo revisa nueve propuestas de aprendizajes de estructura causal a partir de datos observacionales de series de tiempo y compara sus desempeños para sesenta y seis conjuntos de datos con diferentes características. A partir del análisis realizado se toman las cuatro técnicas con mejor desempeño para ser incluidas como parte del *framework* de recuperación de estructuras causales a partir de textos de artículos periodísticos. Para maximizar la precisión de los vínculos obtenidos por la herramienta, se construye el algoritmo de descubrimiento causal utilizando las cuatro mejores técnicas con votación unánime (*ensemble* de técnicas). Un caso de estudio es llevado a cabo para mostrar el potencial del *framework* completo para cumplir el objetivo de obtener una representación causal de un episodio del mundo real que les permita a expertos tener un mejor entendimiento del mismo.

En la Sección 4.2 se revisan conceptos base, presentando las diferentes granularidades posibles para un modelo causal y se introduce el modelo causal que se busca capturar en este capítulo (modelo causal gráfico), así como también las diferentes técnicas del estado del arte relevadas y comparadas para realizar la tarea del aprendizaje del modelo causal gráfico. En la Sección 4.3, se revisan todos los conjuntos de datos de las cuatro fuentes de datos usadas para el análisis comparativo de las técnicas relevadas: (i) 56 conjuntos de datos sintéticos (categorizados en 5 escenarios) generados con la herramienta de simulación *TETRAD* [SSG+98], (ii) 8 experimentos sintéticos del conjunto *nonlinear-VAR* obtenidos de la plataforma de evaluación comparativa (*benchmarking*) de técnicas de causalidad *CauseMe* [RBB+19] (*CauseMe*), (iii) un conjunto de datos reales de demanda de energía eléctrica en el aglomerado urbano del Gran Buenos Aires (GBA) proporcionada por la empresa proveedora de energía eléctrica *CAMMESA*[1] y (iv) un conjunto de datos de series de tiempo de menciones de términos y eventos extraídos del conjunto de datos de noticias del *New York Times* [San08]. Una descripción detallada del proceso para generar las series de tiempo a partir de los textos es dada en esa sección.

[1] https://cammesaweb.cammesa.com/

En la Sección 4.4 se presentan y discuten los resultados de aplicar las técnicas relevadas sobre los datos de origen sintético (las primeras dos fuentes) (i, ii). Luego, en las Secciones 4.5 y 4.6, se revisan los resultados de aplicar las técnicas sobre los datos de origen real: *CAMMESA* y *The New York Times*, respectivamente. El *framework* completo de extracción de causalidad es delineado en este capítulo a partir de herramientas presentadas en capítulos anteriores. La generación completa de los datos a partir de los textos es explicada en la Sección 4.3 y las estructuras causales resultantes de aplicar las técnicas de causalidad estudiadas sobre estos datos son reportadas en la Sección 4.6. Finalmente, en la Sección 4.7 se discuten las conclusiones generales del capítulo y se presentan los posibles trabajos futuros que se desprenden del presente trabajo.

## 4.2. Conceptos Base y Trabajos Relacionados

En la Tabla 4.1 adaptada de [SLB+21] se pueden observar tres tipos de modelos causales y sus características en comparación con un modelo puramente estadístico. Este último modelo es el que se encuentra en la última fila de la tabla y es el más sencillo de los presentes en la tabla. La complejidad de los modelos va creciendo a medida que se sube en las filas llegando al modelo más completo, el modelo mecánico/físico que consiste en un conjunto de ecuaciones diferenciales acopladas que modelan los mecanismos físicos responsables de la evolución de las variables en el tiempo [SLB+21].

Para muchos autores ([PJS17, SLB+21]) el vínculo entre los modelos causales y los modelos estadísticos está dado por el principio de sentido común de Reichenbach: si dos variables X e Y están correlacionadas, entonces o (1) $X \to Y$, o (2) $Y \to X$, o (3) están vinculados por una causa común, o (4) una combinación de las tres opciones. Por lo tanto, según dicho principio, la tarea de detección de causalidad se la puede pensar como la tarea de, dado un vínculo de correlación, determinar de cuál de las cuatro opciones se trata.

Como se puede observar, el modelo estadístico es el único que, sin supuestos adicionales, puede ser aprendido a partir de datos observacionales (datos sin intervenciones que cambien la distribución). Esto es, usando datos i.i.d. El signo de pregunta en los otros modelos indica que la respuesta no es afirmativa o negativa, sino que depende del contexto particular del problema. Con supuestos adicionales puede ser posible, pero sin éstos puede que no lo sea.

| Modelo | Predecir en i.i.d | Predecir ante intervenciones | Responder contrafactuales | Aprender de datos i.i.d. |
|---|---|---|---|---|
| Mecánico/Físico | yes | yes | yes | ? |
| Estructural Causal | yes | yes | yes | ? |
| Gráfico Causal | yes | yes | no | ? |
| Estadístico | yes | no | no | yes |

Tabla 4.1: Tabla adaptada de [SLB+21], que resume los distintos tipos de modelos discutidos al principio de la Sección 4.2 y sus características. Los modelos son reportados de más complejo (en la primera fila) a menos complejo (en la última fila). Se puede ver cómo el modelo más sencillo (estadístico) puede ser aprendido de los datos i.i.id. pero solo permite contestar preguntas observacionales (en i.i.d.). Por otro lado, a medida que se gana en complejidad, por ejemplo, con los modelos Gráficos o los Estructurales, se pueden contestar preguntas adicionales: intervencionales y contrafactuales, respectivamente. El modelo más complejo es el mecánico/físico que consiste en un conjunto de ecuaciones diferenciales acopladas que modelan los mecanismos físicos responsables de la evolución de las variables en el tiempo. El aprendizaje causal se ubica entre medio de los dos extremos (modelo estadístico y modelo mecánico/físico) tratando de recuperar o bien un modelo gráfico causal o uno estructural causal.

De la Tabla 4.1 se puede observar que todos los modelos permiten predecir en i.i.d., esto es, estimar el resultado de preguntas observacionales. Sin embargo, para respuestas del tipo intervencional es necesario tener al menos el modelo gráfico causal, y para responder preguntas contrafactuales es necesario tener al menos el modelo estructural causal. Teniendo un modelo causal de filas superiores se puede simplificar y obtener un modelo de filas inferiores. Esto es, por ejemplo, a partir del modelo estructural causal se puede extraer el modelo gráfico causal o el modelo estadístico, pero no se puede ir en la dirección opuesta (de abajo hacia arriba). Esto es, no se puede obtener el modelo estructural gráfico a partir de un modelo de filas inferiores (modelo gráfico causal o modelo estadístico).

El modelo mecánico/físico es el estándar de oro (*ground truth*) de la causalidad, conteniendo toda la información posible sobre el fenómeno físico/mecánico. Por otro lado, el modelo estadístico es el modelo más sencillo que no tiene en cuenta las relaciones causales entre las variables, solo permiten predecir cuando las condiciones experimentales no cambian y la distribución se mantiene (sin intervenciones). El modelado causal se encuentra entre estos dos extremos [SLB+21].

**Un modelo estructural causal** (SCM por sus siglas en inglés) está compuesto por (i) un conjunto de variables exógenas, (ii) un conjunto de variables endógenas, (iii) un

conjunto de funciones que relacionan estas variables y (iv) una distribución de probabilidad que se desprende del conjunto de funciones [Pea09]. En la práctica se lo representa como un conjunto de asignaciones que sirven para determinar el valor de cada variable en función de otras (pudiendo ser estas exógenas o endógenas). Un ejemplo de modelo SCM son los SCMs de la ecuación 4.1 y la ecuación 4.1. Las variables exógenas ($N_X$,$N_Y$,$N'_X$,$N'_Y$) las variables endógenas ($X$,$Y$) y la distribución de probabilidad que describen se pueden deducir de este conjunto de asignaciones. Las relaciones causales se pueden leer directamente de un SCM. Cuando una variable es función de otra, entonces estamos en la situación en la que la variable dependiente causa la variable independiente, esto es, en el ejemplo del SCM definido por la ecuación 4.1 se puede ver que Y es causado por X, mientras que en el SCM definido por la ecuación 4.1 Y causa X.

**Un modelo gráfico causal** sobre un conjunto de variables $\boldsymbol{X} = (X_1, \ldots, X_d)$ consiste de un grafo $\mathcal{G}$ y una colección de funciones $f(X_j, Pa(X_j)_{\mathcal{G}})$ cuya integral da 1. Donde $Pa(X_j)_{\mathcal{G}}$ es el conjunto de variables padres de $X_j$ (variables que tienen arcos salientes cuyo destino es $X_j$). Las funciones del modelo causal inducen una distribución probabilística $P_{\boldsymbol{X}}$ sobre $\boldsymbol{X}$:

$$P_{\boldsymbol{X}} = p(X_1, \ldots, X_d) = \prod_{j \neq k} f(X_j, Pa(X_j)_{\mathcal{G}}) \tag{4.4}$$

Los nodos del grafo $\mathcal{G}$ representan las variables de interés y sus arcos representan una relación de causalidad, donde el arco va desde la causa al efecto. Es importante resaltar que en el grafo los arcos representan relaciones de causalidad directa, esto es si X causa a Y, e Y causa a Z, es cierto que un cambio en X produce un efecto en Z, pero este efecto se puede explicar a través de Y, por ende, no es una relación directa y no debería haber una flecha de X a Z. Como se mencionó previamente a partir de un modelo estructural causal se puede obtener un modelo causal gráfico, ya que el primero contiene estrictamente más información que el segundo [PJS17].

En este capítulo se analizan diferentes técnicas del estado del arte para aprendizaje de modelos gráficos causales a partir de datos observacionales (i.i.d.) de series de tiempo. En la Figura 4.3b se puede observar una representación de la tarea a realizar en este capítulo. En contraste con la tarea de aprendizaje de estructura causal a partir de datos de corte transversal (Figura 4.3a), en el aprendizaje de estructuras a partir de series de tiempo se tiene un gráfico que se "desenrolla" en el tiempo con flechas que se pueden clasificar en contemporáneas y no contemporáneas. Las **flechas contemporáneas** son aquellas que parten en un determinado instante de tiempo y afectan (se dirigen a) variables en ese

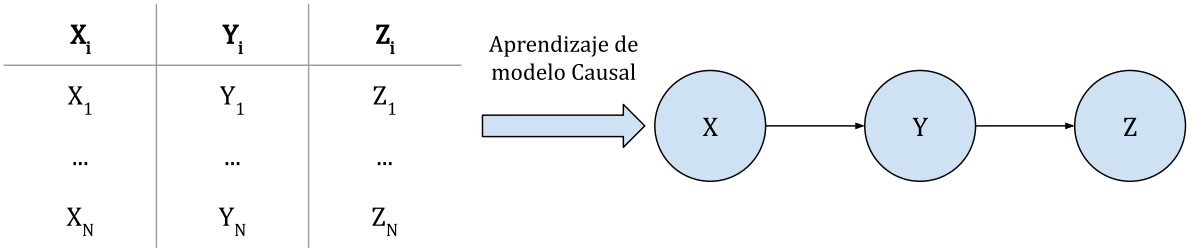
mismo instante de tiempo. Por ejemplo, en la Figura 4.3b, la única relación contemporánea que existe es de $Z$ a $Y$ y se representa con las flechas de la forma $Z_{t-i} \to Y_{t-j}$ con $i = j$. Las **flechas no contemporáneas** tienen la misma forma que las anteriores pero con $i < j$. En el caso de la Figura 4.3b existen dos relaciones no contemporáneas, de $X$ a $X$ (relación autorregresiva (AR)) con un periodo de una unidad de tiempo ($AR$ de orden 1 ($AR(1)$)), y la relación de $Y$ a $X$ con un período de 2. Diferentes herramientas de aprendizaje causal pueden enfocarse en encontrar un tipo o ambos tipos de relaciones. Esto es, existen herramientas de aprendizaje causal que solo apuntan a encontrar las relaciones no contemporáneas y otras herramientas que apuntan a descubrir las dos.

En el presente capítulo se pretende obtener solo los arcos no contemporáneos del grafo causal $\mathcal{G}$ a partir de datos de series de tiempo. En esta capítulo no se calculan las funciones $f(X_j, Pa(X_j)_{\mathcal{G}})$. Aunque teniendo el grafo $\mathcal{G}$ se las puede computar como la distribución de probabilidad condicional de cada variable dado sus padres, el cálculo de dichas funciones no es el objetivo de este trabajo y por ende no se computan ni reportan. Los grafos obtenidos, por una cuestión de interpretación, no son reportados "desenrollados" sino que se muestra un grafo resumido donde las variables no están rezagadas y cada arco de $X \to Y$ se tiene que interpretar como que valores pasados de $X$ causan valores futuros de $Y$.

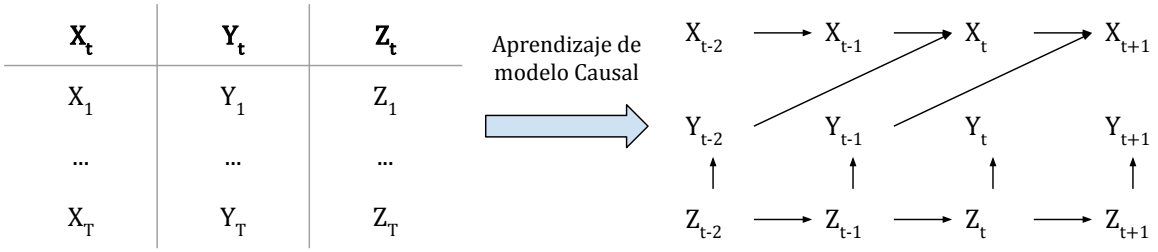
### 4.2.1. Modelos Comparados

Como se mencionó previamente se estudian nueve diferentes técnicas en el marco de aprendizaje de estructuras causales a partir de series de tiempo en un contexto i.i.d. Estas técnicas son: (i) *BigVAR* [NMB17], (ii) *Direct-LiNGAM* [SIS+11], (iii) *ICA-LiNGAM* [SHHK06], (iv) *Lasso-Granger* [Gra69, Tib96], (v) *PC* [SG91], (vi) *PCMCI* [RNK+19], (vii) *SIMoNe* [CSG+08], (viii) *Transfer Entropy* [Sch00] y (ix) *VAR* [Sim80].

Para poner las técnicas en contexto se las divide en tres categorías principales: (1) basadas en independencias, (2) basadas en modelos estructurales causales restringidos y (3) basadas en modelos autorregresivos. En lo que resta de esta sección se contextualizan las técnicas relevadas, explorando las bases teóricas y supuestos sobre los que están apoyados y se revisan otras técnicas de las mismas categorías que por motivos que se detallan no formaron parte del estudio.




a




b

Figura 4.3: Representación gráfica del proceso de aprendizaje causal para el caso de datos de corte transversal (a) y de datos de series de tiempo (b). En (a) se observan "mediciones" de los valores de $X$, $Y$ y $Z$ para $N$ individuos o entidades. Utilizando esos datos se construye un modelo causal entre las variables. La dimensión del tiempo no forma parte de los datos ni del grafo resultante. En (b) se observan $T$ mediciones de las variables $X_t$, $Y_t$, $Z_t$ a lo largo del tiempo para el mismo individuo o entidad. A partir de estos datos, usando aprendizaje causal, se obtiene un grafo causal "desenrollado" en el tiempo que muestra las relaciones entre los valores de las variables en distintos momentos del tiempo. Se puede observar una relación contemporánea ($Z_{t-i} \rightarrow Y_{t-j}$ para $i = j$). También se pueden ver dos relaciones autorregresivas de orden uno ($Z_{t-i} \rightarrow Z_{t-j}$ para $i = j + 1$ y $X_{t-i} \rightarrow X_{t-j}$ para $i = j + 1$) y una relación no contemporánea ($Y_{t-i} \rightarrow X_{t-j}$ para $i = j + 2$).

Las **técnicas basadas en independencias** (1) se apoyan en dos supuestos fundamentales: *propiedad de Markov para grafos dirigidos* y *faithfulness* [KF09]. Asumiendo estos dos supuestos como válidos se tiene una correspondencia uno a uno entre las independencias condicionales del grafo (d-separaciones [KF09]) y las independencias condicionales que se pueden estimar en los datos observacionales a partir de test de estadísticos. Al existir esta correspondencia se puede estimar la existencia de arcos en el grafo causal que originó los datos solo con realizar test estadísticos (test de independencia condicional) sobre el conjunto de datos observacional y sin tener información previa de la estructura

causal. Por ejemplo, si solo se tienen dos variables X e Y, llevando a cabo un test de independencia entre las variables se puede saber si estamos en la situación donde X e Y son independientes. Si ese es el caso ya sabemos que el grafo causal está compuesto por dos nodos ({X,Y}) y el conjunto de arcos vacío ({}). Si X e Y son dependientes reducimos el conjunto de arcos posible a una de dos opciones: {X $\rightarrow$ Y} o {Y $\rightarrow$ X}. Aunque no siempre es posible determinar el grafo original (como el caso anterior donde nos pueden quedar dos o más opciones posibles), se ha demostrado que es posible identificar la clase equivalente de *Markov* [KF09]. Esto es, la clase de todos los grafos que comparten el mismo esqueleto (arcos sin orientar) y *colliders* (conjuntos de tres nodos con la estructura: $X \rightarrow Y \leftarrow Z$). La estructura *colliders* puede ser recuperada de los datos por las independencias condicionales que exhiben, esto es, los *colliders* presentan un conjunto de independencias que cuando son analizadas con test estadísticos dan como resultado una única posible estructura y no múltiples alternativas.

Al utilizar técnicas basadas en independencias se obtiene la clase equivalente de *Markov*, la cual puede tener flechas sin orientar (se obtiene correctamente el esqueleto del grafo, pero no necesariamente se obtiene la dirección de todas las flechas). Como en este trabajo se analizan causalidades no contemporáneas se puede usar la dirección del tiempo para orientar los arcos que queden sin orientar. Esto es, como la causa tiene que suceder antes que el efecto, se sabe que las flechas no pueden ir hacia atrás en el tiempo. Usando ese criterio se obtiene un grafo totalmente dirigido a partir de las técnicas basadas en independencia usadas. Para el presente trabajo se utilizan dos técnicas basadas en independencias: *PC* [SG91] (v) y *PCMCI* [RNK$^+$19] (vi). Se utiliza para tal efecto el paquete *TIGRAMITE*[2] [RNK$^+$19] que tiene ambas técnicas implementadas. Para analizar las independencias condicionales presentes en los datos observados se utiliza el test de correlaciones parciales presente en el mismo paquete de software (*ParCor*). Dicho test estima las correlaciones parciales mediante una regresión lineal computada con mínimos cuadrados ordinarios y un test de correlación lineal de Pearson distinto de cero en los residuos. Otros test de independencias condicionales no lineales no son incluidos por su tiempo de cómputo prohibitivo.

El algoritmo *PC* [SG91] comienza con el grafo no dirigido completo, y luego reduce el conjunto de arcos primero probando test de independencia condicional de orden cero (conjunto condicional vacío), luego de nuevo con test de independencia condicional de

[2]https://github.com/jakobrunge/tigramite

orden uno (conjunto condicional de una sola variable), y así siguiendo. Luego de este procedimiento ya se tiene el esqueleto del grafo (estructura final con todos los arcos no dirigidos). En una segunda etapa, el algoritmo *PC* orienta algunos de estos arcos utilizando las propiedades de las estructuras *colliders*. En el presente capítulo se utiliza una adaptación del algoritmo *PC* presentada en [RNK+19] diseñada específicamente para series de tiempo, que computa un grafo totalmente dirigido pero que no tiene en cuenta relaciones contemporáneas.

La técnica *PCMCI* [RNK+19] está también basada en el enfoque de test de independencias condicionales. De acuerdo a los autores, es una adaptación para series de tiempo altamente interdependientes. El algoritmo consiste de dos pasos: (i) la aplicación del algoritmo *PC* para identificar todos los posibles conjuntos de padres relevantes para cada variable $X_t^j$, notado como $\hat{Pa}(X_t^j)$ (el conjunto de nodos directamente conectados de acuerdo al algoritmo *PC*); y luego (ii) la aplicación del test de independencia condicional momentáneo (MCI por sus siglas en inglés) que definido de la siguiente forma nos permite probar si $X_{t-\tau}^i \to X_t^j$:

$$MCI : X_{t-\tau}^i \perp\!\!\!\perp X_t^j | \hat{Pa}(X_t^j) \backslash \{X_{t-\tau}^i\}, \hat{Pa}(X_{t-\tau}^i) \tag{4.5}$$

De este modo, usando MCI se condiciona en los padres de $X_t^j$ y en los padres rezagados (*time-shifted*) de $X_{t-\tau}^i$. Los dos pasos del algoritmo sirven para lo siguiente: la aplicación de *PC* permite descubrir el conjunto de *Markov* de cada nodo, descartando variables irrelevantes sobre las que no es necesario condicionar, obteniendo de resultado un conjunto de padres candidatos para ser usados en el segundo paso del algoritmo *PCMCI*, el test MCI. Este segundo paso sirve para abordar el problema de controlar los falsos positivos para series altamente interdependientes. Por ejemplo, para probar $X_{t-2}^1 \to X_t^3$ usar el conjunto de padres de $X_t^3$ ($\hat{Pa}(X_t^3)$) es suficiente para detectar causas en común o relaciones indirectas. Condicionar también sobre el conjunto de padres de $X_{t-2}^1$ ($\hat{Pa}(X_{t-2}^1)$) permite controlar por la autocorrelación, lo que posibilita controlar la tasa de falsos positivos al nivel esperado de acuerdo a los resultados presentados por los autores.

Las **técnicas basadas en modelos estructurales causales restringidos** (2) utilizan supuestos adicionales sobre la forma funcional de las relaciones causales para ganar identificabilidad. Por ejemplo, como se mencionó en la Sección 4.1, dos modelos generativos distintos (4.1 o 4.2) pueden haber generado los datos de la figura 4.2, y solo con los datos es imposible determinar cuál de los dos modelos es el que produjo dichos datos. Se

puede plantear (1) un modelo donde $X$ sea función de $Y$ y (2) otro donde $Y$ sea función de $X$ y ambos van a explicar correctamente los datos. Sin embargo, si la función real que explica los datos no es invertible (restricción adicional), generando y comparando ambos modelos ((1) y (2)) es posible detectar cuál es el modelo real que generó los datos observados.

Por ejemplo, para el caso de modelos lineales no gaussianos (LiNGAM por sus siglas en inglés), es posible analizar la asimetría entre la causa y el efecto para poder determinar cuál es la causa y cuál el efecto. En la Figura 4.4, se puede observar un conjunto de datos generados a partir del modelo causal real $X \to Y$, donde $Y = f(X)$ con $f$ siendo una función lineal con ruido uniforme. Los mismos datos son mostrados de dos formas, como Y en función de X (izquierda) y como X en función de Y (derecha). El modelo correcto es el de la izquierda (modelo generativo real de los datos), y se puede observar que la regresión lineal computada tiene residuos homogéneos y correctamente distribuidos. Por otra parte, en el modelo de la derecha, donde se trata de computar $X = f(Y)$, se puede observar que por las características del modelo y por no tratarse del modelo real, se tiene una regresión lineal cuyos residuos están relacionados con la variable $Y$. Esto sugiere que el modelo causal correcto de estos datos es $X \to Y$ y no $Y \to X$. Usando esta estrategia se pueden usar solo datos observacionales para determinar causas y efectos, siempre y cuando se tengan restricciones adicionales sobre el modelo (en este caso que sea lineal no gaussiano). Existen otras combinaciones de restricciones que nos permiten romper la simetría entre causa y efecto para detectar causalidad usando solo datos observacionales. Un resumen de estas configuraciones para ruido gaussiano se puede observar en la Tabla 4.2 adaptada de [PJS17].

En el presente capítulo se analizan y reportan resultados para dos técnicas basadas en modelos estructurales restringidos:*ICA-LiNGAM* [SHHK06] y *Direct-LiNGAM* [SIS$^+$11]. En 2006 la técnica *ICA-LiNGAM* de aprendizajes de estructuras causales fue propuesta utilizando la asimetría entre causa y efecto que presentan los modelos lineales no gaussianos (ver Figura 4.4). Sin embargo, de acuerdo a los autores, *ICA-LiNGAM* presentaba varios problemas: (i) el algoritmo podría no converger a la solución correcta en un número finito de pasos si el estado inicial no era el adecuado o si el tamaño del paso no era seleccionado adecuadamente para las versiones del método que buscaban la solución a través del método del gradiente. (ii) Algunos pasos del algoritmo no eran invariantes ante problemas de escala, por ende, podrían tener problemas de desempeño dependiendo de

| | | Restricciones | Identificabilidad |
|---|---|---|---|
| Modelo Estructural Causal | $X_i = f_i(Pa(X_i), N_i)$ | ninguna | No |
| Modelo de Ruido Aditivo | $X_i = f_i(Pa(X_i)) + N_i$ | no lineal | Si |
| Modelo Causal Aditivo | $X_i = \sum_{k \in Pa(X_i)} f_{ik}(X_k) + N_i$ | no lineal | Si |
| Modelo Lineal Gaussiano | $X_i = \sum_{k \in Pa(X_i)} \beta_{ik} X_k + N_i$ | lineal | No |
| Modelo Lineal Gaussiano (con varianzas de error iguales) | $X_i = \sum_{k \in Pa(X_i)} \beta_{ik} X_k + N_i$ | lineal | Si |

Tabla 4.2: Combinaciones de restricciones para la forma funcional y el impacto que tienen sobre la identificabilidad del vínculo causal usando solo datos observacionales. Esta tabla es adaptada de [PJS17] y son todas para configuraciones con ruido gaussiano.

la escala o la desviación estándar de las variables, especialmente para cuando hay una gran variedad de escalas. Por estos motivos en 2011 Shimizu junto con otros autores de los cuales varios trabajaron en *ICA-LiNGAM*, propusieron un nuevo algoritmo llamado *Direct-LiNGAM*. De acuerdo a los autores, este nuevo algoritmo propuesto garantiza que va a converger a la solución en una cantidad finita de pasos igual al número de variables si los datos siguen estrictamente el modelo [SIS+11]. En el presente capítulo se analiza el desempeño de ambas técnicas en varios conjuntos de datos.

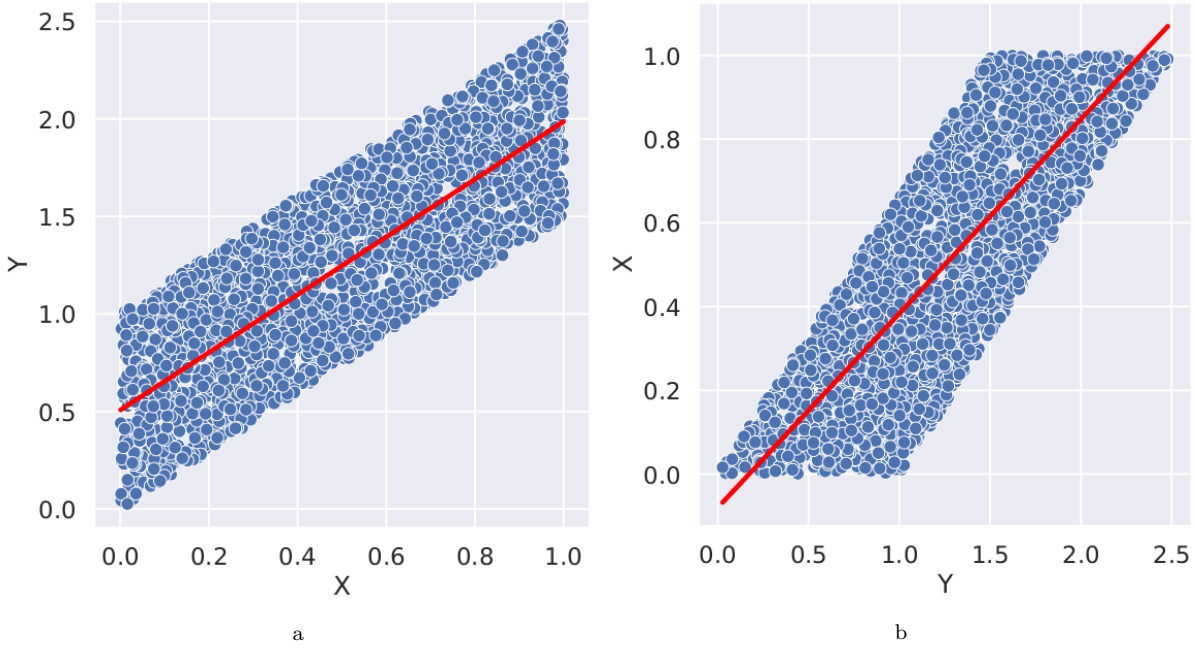



Figura 4.4: Dos gráficos de dispersión que muestran el mismo conjunto de datos invirtiendo los ejes: primero X en el eje horizontal e Y en el vertical (a), y luego al revés (b). Los datos fueron generados a partir de un modelo causal real $X \to Y$, donde $Y = f(X)$ con $f$ siendo una función lineal con ruido uniforme. En esta figura se puede apreciar la asimetría entre la causa y el efecto. Al tratar de modelar los datos con una regresión lineal $Y \sim f(X)$ (a) se pueden observar residuos uniformemente distribuidos. Por otra parte, al modelar los datos con la regresión lineal $X \sim f(Y)$ (b) se obtienen residuos no uniformes (dependientes de $Y$). Las técnicas de aprendizaje de estructuras causales basados en modelos funcionales restringidos aprovechan este tipo de asimetría para detectar el modelo correcto, en este caso el (a).

Las tercera y última categoría de las técnicas analizadas en este capítulo son las **técnicas basadas en modelos autorregresivos** (3). Este categoría es exclusiva a series de tiempo, y se basa en tratar de ver cómo valores del pasado de una variable $X$ dan información única (no presente en otras variables) para predecir o explicar valores del futuro de otra variable $Y$; si esto sucede se hipotetiza que $X \to Y$. Esta idea está resumida en los dos principios en los que se basa la técnica propuesta por Clive Granger en 1969

denominada Causalidad de Granger [Gra69]: (1) la causa tiene que preceder al efecto, (2) la causa produce cambios únicos en el efecto, por lo que el pasado de la causa aporta información única a la tarea de predecir o explicar el efecto. En este capítulo se analizan cinco técnicas que, apoyándose en esos dos principios, son usadas para inferencia de estructuras causales en series de tiempo: (1) *Lasso-Granger* [Gra69, Tib96], (2) *Transfer Entropy* [Sch00], (3) *VAR* [Sim80], (4) *BigVAR* [NMB17] y (5) *SIMoNe* [CSG+08].

La propuesta *Lasso-Granger* está basada en la aplicación de la técnica de regularización lasso (*least absolute shrinkage and selection operator*) como técnica de selección de variables, seguida de la aplicación de la técnica de causalidad de Granger entre pares de variables. Para cada variable $X_i$ se realiza un primer paso de selección de variables, para esto se utiliza lasso para modelar una regresión lineal penalizada con $X_i$ como variable dependiente, y todas las variables del sistema rezagadas como variables independientes. Todas las variables acompañadas por coeficientes significativos (distintos de cero) son utilizados como variables padres candidatas de $X_i$. Finalmente se realiza el test de causalidad de Granger de a pares entre $X_i$ y todas las variables candidatas $X_j$, para determinar si $X_j \rightarrow X_i$. Se computa la causalidad de Granger usando 4 rezagos y se establece un link causal entre las variables cuando el p-valor asociado al estadístico F está por debajo de $0,05$.

La técnica *Transfer Entropy* [Sch00] está basada en los mismos dos principios que el test de Granger pero en lugar de utilizar regresión lineal utiliza el concepto de transferencia de información, el cual está basado en nociones de teoría de la información. Esto es, para probar si $X_j$ causa a $X_i$ ($X_j \rightarrow X_i$), la técnica analiza la transferencia de información del pasado de $X_j$ hacia valores futuros de $X_i$. Si el pasado de la primera variable aporta información única (no presente en el pasado de $X_i$) para predecir valores futuros de $X_i$ se dice que $X_j$ causa a $X_i$. La técnica *Transfer Entropy* se puede entender como una extensión no paramétrica del test de causalidad de Granger [ONSH20]. En [BBS09] se mostró que estas dos técnicas son equivalentes para procesos Gaussianos. Aunque la técnica de causalidad de Granger fue definida concretamente usando un test F en los coeficientes de una regresión lineal, utilizando los dos principios de la técnica y otro tipo de modelos de predicción se pueden definir otras técnicas de detección de causalidad. Este es el caso para la técnica *Transfer Entropy* que reemplaza la regresión lineal por el concepto de transferencia de información. De manera similar, en [Hma20] se define una técnica de extracción de causalidad basada en Granger pero que reemplaza la regresión

lineal por una red neuronal.

La técnica Vector Autorregresivo (*VAR*) [Sim80] se puede utilizar para capturar la relación entre múltiples variables a lo largo del tiempo. Es la extensión del modelo autorregresivo univariado de orden $p$ (*AR(p)*) donde una variable es modelada usando $p$ rezagos de sí misma. Dicho modelo se define como sigue:

$$X_t = c + \beta_1 X_{t-1} + ... + \beta_p X_{t-p} + \varepsilon_t \tag{4.6}$$

En este modelo $\beta_1, \ldots, \beta_p$ son las constantes que acompañan a las variables rezagadas, $c$ es una constante y $\varepsilon_t$ es ruido blanco. La extensión de este modelo para múltiples variables es el modelo *VAR(p)* donde se describe la evolución de $k$ variables, llamadas endógenas, a lo largo del tiempo. Un modelo *VAR(p)* con $k$ variables se lo puede describir con una ecuación similar a la Ecuación 4.6 pero donde cada $X_i$ es un vector de $k$ dimensiones. También se lo puede describir en términos de variables de una sola dimensión usando múltiples ecuaciones. Por ejemplo, un modelo *VAR(p)* con $k = 2$ se lo puede describir como:

$$\begin{cases} X_{1,t} = c_1 + \beta_1 X_{1,t-1} + \cdots + \beta_p X_{1,t-p} + \alpha_1 X_{1,t-1} + \cdots + \alpha_p X_{1,t-p} + \varepsilon_{1,t} \\ X_{2,t} = c_2 + \gamma_1 X_{1,t-1} + \cdots + \gamma_p X_{1,t-p} + \delta_1 X_{1,t-1} + \cdots + \delta_p X_{1,t-p} + \varepsilon_{2,t} \end{cases} \tag{4.7}$$

Como se puede observar, un modelo *VAR(p)* representa cada variable con una regresión lineal de todas las variables endógenas del sistema rezagadas (de 1 hasta $p$), de forma que cada variable se modela usando el pasado de todas las demás variables del sistema. En este modelo $\alpha_i, \beta_i, \gamma_i, \delta_i, c_1$ y $c_2$ son constantes y $\varepsilon_{1,t}$ y $\varepsilon_{2,t}$ son variables de ruido blanco. Los modelos *VAR(p)* son ampliamente usados en diferentes áreas como economía y ciencias naturales. Aunque no son siempre utilizados con el objetivo de estudiar causalidad. Guiados por los principios (1) y (2) en los que se basa la causalidad de Granger podemos pensar al modelo *VAR* como una extensión multivariada del test de causalidad de Granger. En este capítulo se estudia la técnica *VAR* para la extracción de causalidad al estilo Granger, donde de todos los coeficientes del modelo $(\alpha_i, \beta_i, \gamma_i, \delta_i)$ consideramos que los que son significativos (p-valor $< 0{,}05$) indican la existencia de un vínculo causal de la variable independiente a la dependiente. Por ejemplo, si en la Ecuación 4.7 el coeficiente $\gamma_1$ da significativo entonces se considera que $X_{1,t-1}$ causa $X_{2,t}$.

Las técnicas *BigVAR* [NMB17] y *SIMoNe* [CSG+08] están basadas en un modelo *VAR(p)* pero tienen la característica adicional que tienen factores de penalización o regularización para obtener modelos más estables e interpretables (y con más coeficientes

con valor cero). De esta manera, estas técnicas generan modelos con menos *overfitting* y menos sensibles al ruido inherente de los datos. En este contexto de extracción de causalidad es posible obtener modelos con mayor precisión en la recuperación de los arcos (potencialmente afectando a la cobertura), ya que efectos pequeños serán penalizados y llevados a cero (potencialmente descartando efectos espurios pequeños detectados por las técnicas). Al igual que para la técnica *VAR*, se consideran como efectos causales a todos aquellos coeficientes significativos encontrados en las ecuaciones del modelo. A diferencia de la técnica *VAR* sin penalizar (en el cual se determinaba la significancia usando el p-valor de un test estadístico), para *BigVAR* y *SIMoNe* se consideró coeficiente significativo a todo aquel que tenga magnitud distinta de cero (que no haya sido penalizado de tal forma de alcanzar el valor cero).

Es importante notar que las técnicas basadas en modelos regresivos permiten detectar la dirección de la causalidad apoyándose en los principios de la causalidad de Granger, por ende, solo es posible detectar causalidades rezagadas (no contemporáneas). Por otra parte, las técnicas basadas en modelos estructurales causales restringidos permiten detectar la direccionalidad incluso para efectos contemporáneos o para conjuntos de datos de corte transversal porque asumen restricciones adicionales a la forma funcional que permiten romper la simetría entre la causa y el efecto (y así obtener la dirección de la causa al efecto sin necesidad del tiempo). Por otro lado, las técnicas basadas en independencias, solo en algunos casos permiten distinguir efectos causales contemporáneos. Para muchos arcos la dirección no es conocida dando lugar a un grafo parcialmente dirigido. Para las técnicas basadas en independencias es posible determinar alguna de las direcciones faltantes del grafo utilizando los mismos principios aplicados para la causalidad de Granger. Esto es, si existe un vínculo causal no dirigido entre dos variables en distintos instantes de tiempo (relación no contemporánea), usando la intuición de Granger, de que la causa tiene que preceder al efecto, se puede orientar la flecha en la dirección del paso del tiempo. En este trabajo se utiliza una versión de *PC* [SG91] adaptada para series de tiempo que computa un grafo totalmente dirigido utilizando el criterio antes mencionado (solo considerando relaciones no contemporáneas).

La extracción de causalidad también se ha abordado desde la perspectiva de sistemas complejos [SMY+12, MAC14, HA10]. De esta literatura surgen una serie de herramientas de detección de causalidad que se apoyan en el modelo de espacio de estados, siendo el mayor representante de esta categoría la técnica *Convergent Cross Mapping*

(CCM) [SMY+12]. Diversos autores mencionan que las técnicas tradicionales de causalidad asumen separabilidad, lo cual resulta inadecuado para algunos contextos [HLSP17]. En particular los autores, en [SMY+12], aseguran que la técnica de causalidad de Granger no es adecuada para sistemas complejos, ya que asume la propiedad de separabilidad, requerimiento que no siempre se cumple en este dominio. Esto es, la relación causa efecto no siempre está correctamente diferenciada, muchas veces la relación entre variables o sistemas presentan simultaneidad (el comportamiento del predador afecta a la presa y viceversa). Las técnicas de extracción de causalidad de esta categoría fueron desarrolladas para el dominio de sistemas complejos. Dicho dominio tiene algunas particularidades, como por ejemplo los sistemas suelen estar altamente interconectados, vinculados de forma determinística y con comportamientos caóticos (pequeñas variaciones en las condiciones iniciales provocan cambios arbitrariamente grandes en la evolución del sistema). Ya que todos los conjuntos de datos trabajados en este capítulo no se corresponden con este dominio, esta categoría de técnicas no es considerada.

Se puede considerar la existencia de una cuarta categoría de métodos de extracción de causalidad que tampoco es analizada en este capítulo porque han sido propuestos para escenarios donde no solo se tienen datos observacionales sino que además se cuenta con datos intervencionales, esto es, se tienen observaciones del sistema bajo diferentes intervenciones, las cuales pueden ser conocidas o desconocidas [PBM16, HDPM18, EM07, CY13]. Estas técnicas no son consideradas ya que no se cuenta con este tipo de datos para el presente trabajo.

Si bien en el presente trabajo se presentan técnicas de extracción de causalidad a partir de series de tiempo, vale la pena mencionar que existe una gran cantidad de herramientas y literatura dedicada a extraer causalidad directamente de textos [ZLZ+16, FHK+20, PK07, LLZR21, KBR91, KCN00, RDM12, ZWM+17, GM02, Gar97, DSDN18]. En estos escenarios los vínculos causales están explícitos en el texto y se busca la creación de herramientas que los detecten y extraigan. Por ejemplo, en la oración: “La Crisis Financiera Global de 2008 se desató de manera directa debido al colapso de la burbuja inmobiliaria en los Estados Unidos en el año 2006”, se podría detectar y extraer la relación causa-efecto entre la burbuja inmobiliaria del 2006 y la crisis financiera del 2008. Una desventaja de este enfoque es que este tipo de herramientas requieren que el reportero que escribe la noticia o el texto conozca la relación causal y la deje explícita en el texto (directamente con la palabra “causa” o con palabras que indiquen la causa “el desarrollo de X desembocó

en la ocurrencia de Y"). Adicionalmente el reportero puede estar dando su opinión sobre las posibles causas, pero no necesariamente se trata de una representación correcta de la situación. Más aún, puede estar hablando de un posible efecto causal refiriéndose a una causa que aún no sucedió: "Si las empresas especulan una subida del precio de un factor importante en su proceso productivo y deciden subir paulatinamente sus precios causarán inflación".

En [ZWM+17] los autores usan conectores causales para detectar pares $(x, y)$ causales en el texto (como por ejemplo "x because y" y "x leads to y"). Luego para obtener patrones generales de alto nivel generalizan los sustantivos a sus hiperónimos utilizando *WordNet* [Mil95] y los verbos a sus clases con *VerbNet* [KS05] (por ejemplo, "kill" pertenece a la clase "murder-42.1"). De esta manera logran obtener patrones generales de causalidad que no necesariamente estaban explícitos en los textos pero que se pueden deducir. Por ejemplo, se plantean como objetivo transformar un par causal como "*a massive 8.9-magnitude earthquake hit northeast Japan on Friday* → *a large amount of houses collapsed*" en un par causal más general no presente en el texto: "*earthquake hit* → *house collapse*" a través de generalizaciones con *WordNet* y *VerbNet*. Con estos pares de causalidad construyen una red causal y una representación distribuida de la misma (embedding) que argumentan sirve para tareas posteriores. En su caso la usan para predicción de movimientos de los precios de la bolsa. Esta estrategia sirve para agregar flexibilidad y capacidad de generalización a la extracción de causalidad de texto, pero aún se requieren menciones de causalidad explícitas en textos y se necesitan herramientas sofisticadas para lograr detectar y extraer esas menciones.

En [RDM12] construyen redes causales a partir de textos con el objetivo de predecir. El algoritmo que presentan, *Pundit*, extrae pares causales de textos de noticias no estructurados buscando por patrones causales gramaticales ("*because*", "*due to*", "*lead to*", y otros). Luego generalizan estos pares usando conocimiento del mundo que proviene de diferentes ontologías. El grafo está compuesto por conceptos de *Wikipedia*, *ConceptNet* [LS04], *WordNet* [Mil95], *Yago* [SKW07] y *OpenCyc*. Por otra parte, para las relaciones entre los conceptos (por ejemplo "CapitalOf) usan el proyecto *LinkedData* [BHBL11]. El objetivo es poder predecir a través de la generalización de eventos y sus relaciones. Por ejemplo, detectar "Earthquake hits [Country Name]" causa "Red Cross help sent to [Capital of Country]" (a través de múltiples acontecimientos de este tipo de evento), lo que permitiría predecir, ante la ocurrencia de un terremoto en un país, el envío de ayuda por parte

de la Cruz Roja a la Capital del país.

Un trabajo relacionado al realizado en este capítulo es el de [BCFS19]. En dicho trabajo construyen un grafo causal midiendo cómo la ocurrencia de una palabra puede influenciar la ocurrencia de otras en el futuro. Utilizan el concepto de Causalidad de Granger para medir la capacidad de una palabra de influenciar la ocurrencia de otra. En su trabajo realizan un trabajo de filtrado de palabras ad-hoc, filtrando palabras demasiado frecuentes (si aparece en más del 50 % de los días) o demasiado infrecuentes (si aparecen en menos de 100 artículos). También utilizan entidades como parte de su vocabulario, las cuales son recuperadas con un reconocedor de entidades (NER). Por último, utilizan un algoritmo para reconocer *triggers* de eventos [Ahn06] para ser incorporados al vocabulario. A este vocabulario le agregan bigramas frecuentes. Una de las principales limitaciones de este trabajo es que trabajar a nivel de términos y bigramas aislados puede conducir a un vocabulario difícil de interpretar en una red causal. En este trabajo los autores no reportan ninguna red causal como resultado, sino que muestran la utilidad de la herramienta para hacer predicciones del precio de la bolsa de valores. Una representación semánticamente significativa de cada variable es necesaria para poder mostrar un grafo causal con nodos y relaciones causales interpretables. Otra limitación es el uso de una única herramienta para la parte de extracción de causalidad (Causalidad de Granger). La tarea de descubrimiento causal es una tarea con una gran complejidad y no existe una única técnica que se adecue a todos los dominios. Por tal motivo es importante el análisis de múltiples herramientas y la adopción de la o las mejores opciones para el dominio de trabajo.

En el presente trabajo se utiliza la técnica de pesaje de términos relevantes a un dominio presentada en el Capítulo 2 ($\text{FDD}_\beta$) para obtener términos relevantes (unigramas, bigramas y trigramas). Luego se detectan *event-triggers* de eventos en curso utilizando el modelo para tal fin presentado en el Capítulo 3. Estos *event-triggers* son luego representados utilizando una representación vectorial que tiene en cuenta todo el contexto, y posteriormente mostrados utilizando una representación que permite visualizar el *event-trigger* y todo el contexto, dando lugar a descripciones de eventos más interpretables. Por último, los términos relevantes y los eventos completos son unidos en un solo conjunto de datos para la construcción de la estructura causal. Esta construcción de la estructura causal es llevada a cabo a través de un *ensemble* de cuatro técnicas de descubrimiento causal (*PC*, *PCMCI*, *Direct-LiNGAM* y *VAR*).

## 4.3. Conjuntos de Datos

Para los experimentos realizados en el presente capítulo se utilizaron datos de cuatro fuentes distintas: (1) *TETRAD* [SSG+98], (2) *CauseMe* [RBB+19], (3) *CAMMESA*[3] y (4) *The New York Times* [San08]. La primera fuente se trata de una herramienta de simulación con la cual se generaron 56 conjuntos de datos sintéticos con diferentes características. La segunda fuente es una plataforma de *benchmarking* de técnicas de extracción de causalidad con varios conjuntos de datos disponibles. De esta plataforma se utilizaron los 8 conjuntos de datos correspondientes a los experimentos *nonlinear-VAR*. Estas dos primeras fuentes (1) y (2), son conjuntos de datos sintéticos. Por otro lado, las otras dos fuentes se corresponden con datos observados del mundo real. La fuente (3) es la Compañía Administradora del Mercado Mayorista Eléctrico Sociedad Anónima (*CAMMESA*) que puso a disposición un conjunto de datos con mediciones de demanda de energía eléctrica en el área metropolitana de la ciudad de Buenos Aires (Gran Buenos Aires (GBA)) junto con mediciones de variables climáticas para el mismo área geográfica. La fuente (4) son los textos completos del corpus del *The New York Times*. A partir de estos se construye un conjunto de datos de series de tiempo de menciones de términos y eventos en curso detectados en dichos textos. En la presente sección se describe en detalle cómo están constituidos cada uno de los conjuntos de datos y sus diferentes características. Estos conjuntos de datos son presentados en ese orden ((1), (2), (3) y (4)) para respetar el orden en el que se reportan los resultados en las Secciones 4.4, 4.5 y 4.6

### 4.3.1. Fuente #1: *TETRAD*

*TETRAD* [SSG+98] es una aplicación de escritorio escrita en Java que permite crear, estimar o buscar modelos causales. A partir de estos modelos se pueden realizar tests y predicciones. Adicionalmente permite crear modelos causales aleatorios y a partir de estos generar conjuntos de datos. Esta última funcionalidad es la que se utilizó para el presente trabajo. Dado que esta herramienta permitía de manera flexible modificar varios de los parámetros de creación de los modelos y datos simulados, se aprovechó la herramienta para generar diversas configuraciones para probar diferentes aspectos de las técnicas de aprendizaje de estructura causal estudiadas. Las diversas configuraciones probadas se pueden categorizar en cuatro escenarios distintos: (i) variar el número de nodos (**N**) del

[3] https://cammesaweb.cammesa.com/

modelo causal, (ii) variar el tamaño de la serie de tiempo (**T**), (iii) variar la cantidad de variables latentes (**H**), y (iv) variar la cantidad de rezagos presentes en el modelo causal real (**L**). Para todos los escenarios se fijaron todos los parámetros excepto el que estaba siendo analizado el cual se lo variaba dentro de un rango de valores elegidos. Se reportan los parámetros fijos y los rangos utilizados en la Tabla 4.3.

Se utilizó la configuración predeterminada de *TETRAD* para todos los parámetros excepto para los que varían (N, T, H, L) y para el valor máximo del rango de coeficientes (de 0,7 se lo cambió a 0,5). Los valores mínimos y máximos de los coeficientes son los valores a partir de los cuales se muestrean los coeficientes asociados a los vínculos causales. En el caso aquí presentado se muestrean de manera uniforme aquellos correspondientes al intervalo (-0,5; -0,2) ∪ (0,2; 0,5). Utilizando un parámetro de 0,7 se obtenían series no estacionarias con facilidad (las series explotaban hacia infinito) y por esta razón se utilizó un valor máximo menor (0,5). Una descripción detallada de cada parámetro se puede encontrar en los menús flotantes de cada parámetro en la aplicación *TETRAD*[4]. Una descripción detallada de cada uno de los cuatro escenarios es dada a continuación.

Para el **escenario #1** se utilizaron nueve valores distintos para la cantidad de nodos ($N \in \{6, 9, 12, 15, 18, 21, 24, 27, 30\}$). Para poder medir el desempeño de las diferentes técnicas ante un número creciente de variables se mantuvieron el resto de los parámetros en una configuración sencilla: cantidad de variables ocultas igual a cero ($H = 0$), cantidad de rezagos incluidos en el modelo igual a uno ($L = 1$) y longitud de la serie se la fijó a 1.069 ($T = 1.069$). Esta longitud para la serie fue elegida para concordar con el conjunto de datos obtenido del *New York Times* que consiste en de 1069 semanas (desde enero 1987 hasta junio 2007). Aunque finalmente no se usó esta frecuencia para este conjunto de datos (se usó frecuencia mensual para los datos extraídos del *New York Time*), se mantuvo la configuración $T = 1.069$ ya que igual constituye un tamaño adecuado para los presentes experimentos.

Para el **escenario #2** se utilizaron siete diferentes valores para la longitud de la serie, $T \in \{100, 500, 1.000, 2.000, 3.000, 4.000, 5.000\}$. Se utilizó $N = 30$, y al igual que para el escenario #1 se utilizó $H = 0$, $L = 1$ y $T = 1.069$.

Para el **escenario #3**, además de las variables observadas se agregaron variables no observadas que podían introducir complejidad a la tarea de descubrimiento causal. Además de tener 20 variables observadas ($N = 20$) se crearon siete conjuntos de datos

[4]https://www.ccd.pitt.edu/tools/

con diferente cantidad de variables no observadas (ocultas), $H \in 0, 2, 4, 6, 8, 10, 12$. Al igual que para los escenarios anteriores se usaron $L = 1$ y $T = 1.069$.

Para el **escenario #4** se varió la cantidad de variables rezagadas incluidas en el modelo causal real. El objetivo del presente capítulo es detectar correctamente las relaciones causales no contemporáneas de exactamente una unidad de tiempo en el pasado (relaciones del tipo $X_{j,t-1} \rightarrow X_{i,t}$). Sin embargo la herramienta crea también vínculos contemporáneos($X_{j,t} \rightarrow X_{i,t}$), y en este escenario, vínculos causales con mayor distancia ($X_{j,t-\tau} \rightarrow X_{i,t}$ con $\tau \in \{2, 3, 4, 5\}$). Lo que se pretende analizar a partir de este escenario es la capacidad de las técnicas de encontrar las relaciones causales directas correctas con distancia uno, a pesar de tener correlaciones adicionales originadas por arcos adicionales que no son los buscados ($X_{j,t-\tau} \rightarrow X_{i,t}$ con $\tau \in \{0, 2, 3, 4, 5\}$). La cantidad de nodos se la fijó en diez ($N = 10$).

En resumen, se varió T en el escenario 2, pero para todos los demás se lo fijó a $T = 1.069$ para que tuviera la misma dimensión que el conjunto de datos originado con *The New York Times* con frecuencia semanal. Por otro lado, los valores de H y L fueron variados en los escenarios 3 y 4 respectivamente, pero en todos los demás, por simplicidad, se los fijó en $H = 0$ y $L = 1$. Por último, se varió N en el escenario 1, pero para todos los demás se usó el N más grande posible. Es importante destacar que no se podían elegir valores de N arbitrariamente grandes en la herramienta, ya que con un valor grande de N se tiene una gran cantidad de arcos y se generan situaciones problemáticas en las que las series tienen comportamientos no estacionarios (explotan hacia infinito).

La estructura causal real de los conjuntos de datos generados a partir de TETRAD puede ser representada usando Grafos Acíclicos Dirigidos (DAG por sus siglas en inglés). Para cada uno de los cuatro escenarios se utilizaron las dos posibles configuraciones para construcción de DAG provistas por *TETRAD*: *Random Foward DAG* (*RFDAG*) y *Scale-free DAG* (*SFDAG*). La primera estrategia crea un DAG aleatoriamente agregando arcos hacia adelante (arcos que no apunten a antecesores de la variable), donde los arcos son insertados de a uno. Por otro lado, la estrategia *SFDAG* crea un DAG cuyas variables tienen un grado de conectividad que sigue una ley de potencias. Se utilizan y reportan resultados para ambas configuraciones de construcción de DAG.

Habiendo **nueve** configuraciones posibles para el escenario 1, **siete** para el escenario 2, **siete** para el escenario 3 y **cinco** para el escenario 4, se tiene un total de 28 configuraciones. Para cada una de estas configuraciones se crea un conjunto de datos usando la estrategia

| Descripción del Parámetro | Valor |
|---|---|
| Valores mínimo y máximo del rango de los coeficientes | (0,2; 0,5) |
| Valores mínimo y máximo del rango de la covarianza | (0; 0) |
| Grado máximo del grafo | 100 |
| Varianza del ruido de medición aditivo | 0 |
| Cantidad de rezagos incluidos en el modelo (L) | {1, 2, 3, 4, 5} |
| Cantidad de variables ocultas incluidas en el modelo (H) | {0, 2, 4, 6, 8, 10, 12} |
| Cantidad de variables observadas incluidas en el modelo (N) | {6, 9, 12, 15, 18, 21, 24, 27, 30} |
| Longitud de la serie de tiempo (T) | {100, 500, 1.000, 2.000, 3.000, 4.000, 5.000} |
| Para grafos *scale-free*, el parámetro alfa | 0,05 |
| Para grafos *scale-free*, el parámetro beta | 0,9 |
| Para grafos *scale-free*, el parámetro delta_in | 3 |
| Para grafos *scale-free*, el parámetro delta_out | 2 |
| Valores mínimo y máximo del rango de varianza | (1; 3) |
| Coeficientes negativos | Si |
| Covarianza negativa | Si |
| Estandarizar datos | No |

Tabla 4.3: Descripción de los parámetros usados en la herramienta de simulación de datos *TETRAD* para generar los 56 conjuntos de datos reportadas. Se reportan en esta tabla tanto los parámetros fijos como los variables (siendo estos últimos los que están entre llaves). Por ejemplo, se varió T en el escenario 2 utilizando los valores entre llaves reportados en esta tabla, pero para todos los demás se lo fijó a $T = 1.069$. Por otro lado, los valores de H y L fueron variados en los escenarios 3 y 4 respectivamente (usando los valores entre llaves), pero en todos los demás, por simplicidad, se los fijó en $H = 0$ y $L = 1$. Por último, se varió $N$ en el escenario 1 (usando los valores entre llaves), pero para todos los demás se usó el $N$ más grande posible. Siendo estos $N$: $N = 30$, $N = 20$ y $N = 10$ para los escenarios 2, 3 y 4, respectivamente.

*RFDAG* y otro usando *SFDAG*. Finalmente se tiene un total de 56 conjuntos de datos diferentes para la fuente #1 (*TETRAD*).

### 4.3.2. Fuente #2: *CauseMe*

Para tener variedad en datos de origen sintético, y para agregar conjuntos de datos existentes (no solo los datos creados especialmente para este trabajo) se agregaron datos provenientes de la plataforma para evaluación comparativa (*benchmarking*) *CauseMe*[5] [RBB+19]. A la fecha en la que esta tesis fue escrita, había en dicha plataforma 19 conjuntos de datos disponibles para su descarga. Cualquiera de estos puede ser descargado

[5] causeme.net

y probado con un algoritmo de recuperación de estructura causal. Es importante resaltar que los arcos correctos (*ground truth*) no están disponibles para descargar. En su lugar, los resultados deben ser subidos a la plataforma para que allí se calculen las métricas de desempeño sobre la técnica propuesta.

En la documentación de *TETRAD* se aclara que los coeficientes son muestreados de una distribución uniforme, pero —de acuerdo a lo que se pudo observar— no se presenta la forma funcional con la que se computan los valores simulados. Como la forma funcional no está descrita, se asume que debe ser sencilla y por ende se asume lineal. Para agregar variedad a los conjuntos de datos sintéticos se toma de *CausaMe* el cojunto *nonlinear-VAR*[6].

De acuerdo a la descripción del conjunto de datos usado, los datos presentan tres desafíos usualmente encontrados en estos procesos estocásticos: autocorrelación, relaciones rezagadas en el tiempo y no linealidad. Se los combina con desafíos para las herramientas estadísticas/computacionales: dimensionalidad alta y series de tiempo cortas. El máximo rezago en las relaciones causales es 5 (distancia máxima entre la causa y el efecto).

El conjunto de datos *nonlinear-VAR* consiste de ocho repeticiones con diferentes parámetros. Cuatro repeticiones utilizando longitud de serie 300 ($T = 300$), para 3, 5, 10 y 20 nodos ($N \in \{3, 5, 10, 20\}$). Y cuatro repeticiones utilizando longitud de serie 600 ($T = 600$), para 3, 5, 10 y 20 nodos ($N \in \{3, 5, 10, 20\}$).

### 4.3.3. Fuente #3: *CAMMESA*

La tercera fuente de datos es la Compañía Administradora del Mercado Mayorista Eléctrico Sociedad Anónima (*CAMMESA*), una compañía argentina encargada de operar el mercado eléctrico mayorista de Argentina. Dicha compañía suministró para este trabajo los datos de la demanda de energía eléctrica en la zona metropolitana de la ciudad de Buenos Aires (Gran Buenos Aires (GBA)) para el periodo enero-2012 hasta diciembre-2018. Dentro de *CAMMESA* se trata de resolver el problema de predecir la demanda de energía eléctrica que va a haber en cada zona en el corto, mediano y largo plazo, para solicitar el suministro correspondiente a las compañías generadoras de energía eléctrica. Para realizar estas predicciones *CAMMESA* recopila información del clima y otros indicadores relevantes (por ejemplo, Estimador mensual de actividad económica (EMAE)). A

[6]https://causeme.uv.es/model/nonlinear-VAR/

partir de los datos suministrados por *CAMMESA* (que son de frecuencia horaria) se crea **un** conjunto de datos de tipo serie de tiempo con frecuencia diaria durante el periodo enero-2012 hasta diciembre-2018, teniendo un total de 2.558 observaciones y nueve variables. Las nueve variables describen diferentes aspectos de la zona analizada (GBA): (1) la demanda de energía eléctrica (DemGBA), (2) La temperatura promedio (Temp), (3) la componente vx del viento (vx), (4) la componente vy del viento (vy), (5) Irradiancia Horizontal Global (GHI), (6) la humedad (Hum), (7) la presión (Pres), (8) la sensación térmica (Ster) y (9) una variable que representa si el día actual es laborable o no (Wrk). Todas las variables son de tipo real excepto la última que se trata de una variable binaria.

El conjunto de datos provista por *CAMMESA* con frecuencia diaria contiene $2.558 \times 9$ datos (TxN), pero no contiene información respecto a la estructura causal real del problema. Sin embargo, analizando la naturaleza de las variables y conversando con los expertos de la compañía llegamos a las siguientes conclusiones respecto a la estructura causal real (*ground truth*).

***Ground truth CAMMESA*, simplificación**. Debido a que no se contaba con expertos en climatología no se tiene información detallada del modelo causal real respecto a las variables climatológicas entre sí. Por simplicidad se toman como referentes de las variables climáticas la humedad y la temperatura (que se entiende son las más vinculadas con la variable de interés, la demanda). Las demás variables no son consideradas para la *ground truth* ya que sumarlas no aportaba al conocimiento de este mismo y cada nueva variable climatológica agrega muchos arcos posibles sobre los que se tiene poca información. Esto quiere decir que para los experimentos y para reportar las métricas de desempeño solo se usan las dos variables climatológicas consideradas, la demanda y si es día laborable ($\{\text{Wrk}, \text{Hum}, \text{Temp}, \text{DemGBA}\}$).

***Ground truth CAMMESA*, arcos inexistentes**. Para construir la *ground truth* se parte de algunas relaciones que por sentido común se asumen incorrectas, esto es, arcos que no deberían ser encontrados por las herramientas de aprendizaje causal. Por ejemplo, (1) no debería existir un vínculo causal desde ninguna variable climática hacia la variable Wrk, el clima no afecta a que un día sea laborable o deje de serlo. Tampoco debería afectar la demanda de energía eléctrica a la variable Wrk. Análogamente, (2) se entiende que la demanda de energía eléctrica no puede afectar a ninguna variable climática ni tampoco determinar que un día sea feriado. Por último, (3) que un día sea feriado no puede afectar a ninguna variable climática. En resumen, sabemos que hay siete arcos incorrectos que no

deberían ser encontrados por las técnicas de descubrimiento causal. Estos arcos incorrectos están dibujados en rojo en la Figura 4.5. En dicha Figura la *ground truth* está representada con el grafo "desenrollado" (abajo) y con el grafo resumido (atemporal) (arriba). En el grafo de abajo los arcos deben leerse como no contemporáneos, esto es, una flecha de la forma $x \to y$ representa $x_{t-\tau} \to y_t$ con $\tau \neq 0$.

***Ground truth CAMMESA*, arcos existentes posibles**. De acuerdo a las discusiones con expertos de la empresa se asume un vínculo causal entre algunas variables climatológicas y la demanda de energía eléctrica en el Gran Buenos Aires. En particular, se sospecha un vínculo con la temperatura y la humedad ($\text{Temp}_{t-1} \to \text{DemGBA}$ y $\text{Hum}_{t-1} \to \text{DemGBA}$). También, de la misma discusión, se asume que si el día es laborable o no tiene un impacto en la demanda ($\text{Wrk}_t \to \text{DemGBA}_t$). Más aún, según los expertos, si el día anterior fue feriado tiene un impacto en la demanda (no es lo mismo la demanda un lunes, que un miércoles, que un viernes, o que un martes luego de un feriado) ($\text{Wrk}_{t-1} \to \text{DemGBA}_t$). Adicionalmente, si bien se desconoce con exactitud la relación entre las dos variables climáticas consideradas (Hum y Temp), se asume que se influencian entre sí en el tiempo ($\text{Hum}_{t-1} \to \text{Temp}_t$ y $\text{Temp}_{t-1} \to \text{Hum}_t$).

Por último se considera que las variables climáticas y la demanda tienen una fuerte inercia en sus valores, mostrando un comportamiento autorregresivo ($Temp_{t-1} \to \text{Temp}_t$, $\text{Hum}_{t-1} \to \text{Hum}_t$ y $\text{DemGBA}_{t-1} \to \text{DemGBA}_t$). Adicionalmente, se asume que la variable feriado tiene una componente autorregresiva, aunque menor que para las tres variables anteriores ($\text{Wrk}_{t-1} \to \text{Wrk}_t$). Esto se debe a que los días hábiles se tienden a agrupan juntos y los días no laborables también. En una semana regular sin feriados, luego de un día hábil (lunes, martes, miércoles, jueves o viernes) hay una probabilidad de 4/5 de que el día siguiente sea día hábil nuevamente y solo 1/5 de que sea día no laborable (solo si el día actual es viernes).

Un resumen de los arcos correctos esperados se puede ver en la Figura 4.5. Los arcos correctos están representados en negro del lado derecho de la Figura. En la parte superior está representado el grafo atemporal, y en la parte de abajo está representado el grafo "desenrollado". Como se puede observar, los arcos autorregresivos no son representados en la representación atemporal ya que no es el objetivo de este trabajo capturar relaciones causales de variables hacia sí mismas ($x_{t-1} \to x_t$).

La aplicación de técnicas de descubrimiento causal a este conjunto de datos es presentada y analizada en la Sección4.5. Las técnicas son comparadas en base a la *ground truth*

aquí discutida. Diferentes transformaciones de los datos para la eliminación de ciclos son consideradas para sumar al análisis comparativo de las técnicas de descubrimiento causal.

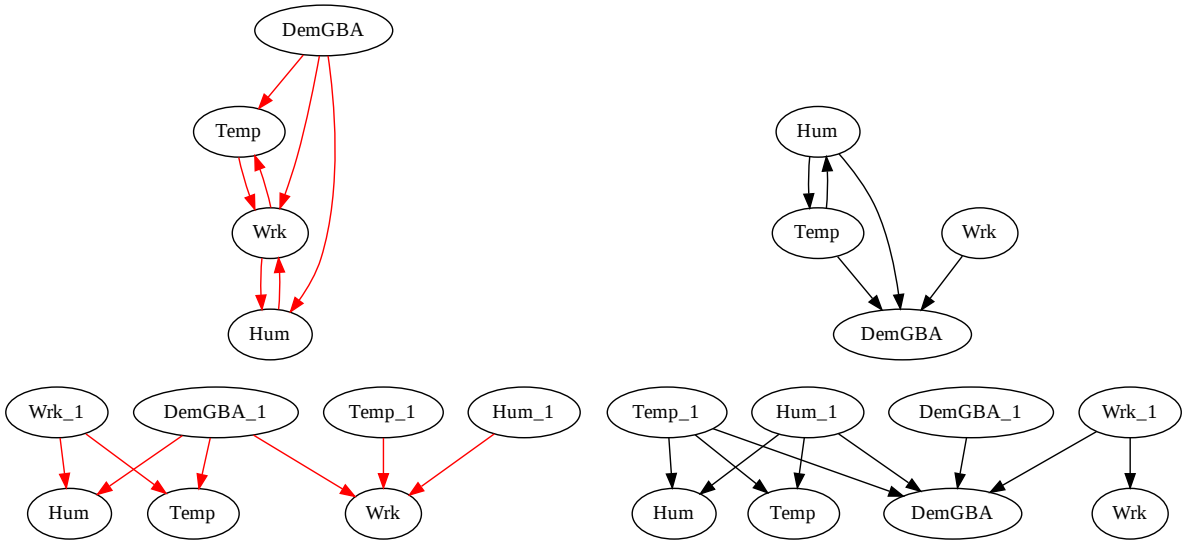



Figura 4.5: Descripción gráfica de la estructura causal real (*ground truth*) del conjunto de datos obtenido de la empresa *CAMMESA*. A la izquierda (a y c) están representados los arcos incorrectos en rojo. A la derecha (b y d) están representados en negro los arcos correctos. La *ground truth* se muestra en dos formatos distintos (pero codificando la misma información): abajo (c y d) se muestran los grafos "desenrollados" en el tiempo, mostrando cómo se afectan causalmente las variables de un instante al siguiente (como se trabaja con frecuencia diaria, de un instante al siguiente hay 24 horas de diferencia). Como las flechas contemporáneas no son buscadas, las mismas no se analizan ni reportan en este trabajo. En la parte superior de la figura (a y b) se muestra el grafo atemporal resumido, donde cada arco representa una relación de causalidad de un instante al otro. Esto es, la flecha Hum $\rightarrow$ Temp en (b) se debe leer como $\text{Hum}_{t-1} \rightarrow \text{Temp}_t$. Como se puede observar, los arcos autorregresivos no son representados en la representación atemporal ya que no es el objetivo de este trabajo capturar relaciones causales de variables hacia sí mismas ($x_{t-1} \rightarrow x_t$). Para obtener esta estructura de arcos incorrectos y arcos correctos se analizó la naturaleza de las variables y se conversó con los expertos de la compañía.

### 4.3.4. Fuente #4: *The New York Times*

La cuarta fuente de datos es el *The New York Times Annotated Corpus*[7] [San08]. El objetivo de la utilización de textos en el presente trabajo es generar un prototipo de una herramienta que permita a los expertos entender y sacar conclusiones sobre eventos y otras variables del mundo real que hayan sido reportados en artículos de noticias. La

[7] https://catalog.ldc.upenn.edu/LDC2008T19

herramienta tiene por objetivo detectar y mostrar variables relevantes a un dado dominio de manera semiautomática, y mostrar posibles vínculos causales entre esas variables para permitir a los expertos un mejor entendimiento del dominio analizado. En el presente trabajo se muestra un caso de uso de la herramienta aplicada no a un dominio específico sino a una entidad geopolítica (GPE por sus siglas en inglés) particular, en este caso Irak. Para elegir una GPE con una buena cantidad de menciones se utilizó un detector de entidades (*NER*) de la librería *spaCy* para detectar todas las menciones de dichas entidades en todo el corpus del *New York Times* en el periodo enero-1987 a junio-2007. Se detectaron las siguientes 10 GPEs como las más mencionadas en el corpus (se muestran ordenadas de las más mencionadas a las menos mencionadas): `New York`, `the United States`, `Manhattan`, `Washington`, `Iraq`, `New York City`, `America`, `China`, `Brooklyn` y `New Jersey`. Se prefirió elegir una GPE fuera del país de origen del corpus ya que sobre el mismo país probablemente haya mayor cantidad de noticias y más variadas que sobre países extranjeros. Por este motivo se eligió el primer GPE externo a Estados Unidos, Irak (con 180.206 menciones en un total de 170.497 oraciones distintas).

A partir del texto completo de las noticias de *The New York Times* en el período enero-1987 a junio-2007 se filtra por GPE, solo considerando como dominio de interés las oraciones con menciones explícitas del país Irak. A partir de estos textos se construye un conjunto de datos de tipo series de tiempo con variables de interés con el objetivo de aplicar sobre ella las técnicas de descubrimiento de estructura causal. Para la construcción de este conjunto de datos se consideraron dos tipos de variables de interés: términos mencionados en el texto (que pueden ser unigramas, bigramas o trigramas) y menciones de eventos en curso del mundo real. Para el primero se utiliza la técnica de pesaje de términos $\text{FDD}_\beta$ presentada en el Capítulo 2 y para el segundo tipo de variables se utiliza el modelo de detección de eventos en curso presentado en el Capítulo 3. Es importante mencionar que el filtro de GPE y todo el trabajo posterior es a nivel de oraciones y no de artículos completos para que sea compatible con el trabajo del Capítulo 3 donde se trabaja con la detección de eventos a nivel de oración y no de artículos.

Las variables de ambos tipos (términos y eventos en curso) son detectadas en el texto de manera independiente, y a cada mención de cada variable se le asocia el mes y año de publicación del artículo donde aparece reportada. De esta manera se generan dos conjuntos de datos preliminares, uno con menciones de términos (con $M$ términos) y otro con menciones de eventos (con $P$ eventos). Como se utiliza la misma frecuencia y

el mismo periodo de tiempo (frecuencia mensual en el periodo enero-1987 a junio-2007) ambos conjuntos tienen la misma longitud (246 meses) y por ende estos pueden ser unidos. Esta unión resulta en el conjunto de datos final de dimensiones $246 \times (P + M)$, donde $(P + M)$ es la cantidad de variables y 246 es la longitud de las series. A continuación, se describe el proceso a través del cual se generan los dos conjuntos de datos preliminares, el de menciones de términos $(246 \times M)$ y el de menciones de eventos en curso $(246 \times P)$.

**Conjunto de datos de menciones de términos.** Para poder construir este conjunto de datos de términos se debió detectar repeticiones de unigramas, bigramas y trigramas para construir un vocabulario y luego utilizar la técnica de pesaje de términos $\text{FDD}_\beta$ definida en el Capítulo 2 para obtener un conjunto reducido de términos altamente relevantes sobre el cual elegir los $M$ términos finales. Como la técnica de pesaje de términos a ser utilizada es supervisada, se necesita un conjunto de textos relevantes al contexto y otro conjunto de textos irrelevantes para poder estimar la relevancia de cada término para el contexto dado. En este caso el contexto es el país Irak, por ende, se usa como conjunto relevante las 170.497 oraciones que mencionan explícitamente al GPE, y se selecciona un conjunto de la misma cantidad de oraciones de forma aleatoria de las 63.734.239 oraciones restantes del corpus (las que no mencionan al GPE Irak). El conjunto final tiene 340.994 oraciones, donde la mitad mencionan al país de interés y la otra mitad no. Notar que el término "`Iraq`" va a estar presente en todas las oraciones relevantes y no estará en las irrelevantes, alcanzando un poder tanto descriptivo como discriminativo de $1,0$.

A partir de este conjunto de datos se construye un vocabulario de términos (unigramas, bigramas y trigramas) presentes en el corpus. Se descartan automáticamente términos que no aparezcan ni una vez en el conjunto de relevantes (ya que tienen poder descriptivo y discriminativo cero). Se descartan también de forma automática aquellos términos que aparezcan en menos del $0,1\,\%$ del corpus (341 menciones o menos), ya que esa cantidad de menciones no alcanzaría para construir una serie de tiempo con suficientes datos como para aplicar las técnicas de causalidad. Por último, se filtran también aquellos términos que representen cantidades o signos de puntuación, así como también *stopwords*. A la totalidad de los términos restantes del vocabulario se les aplica $\text{FDD}_\beta$ utilizando el mejor $\beta$ encontrado durante los experimentos del Capítulo 2 para la estimación de relevancia de términos asignados por usuarios, esto es $\beta = 0,477$. En la Tabla 4.4 se muestran tres listas con los 10 unigramas, 10 bigramas y 10 trigramas con mayor $\text{FDD}_\beta$ de todo el vocabulario utilizado. Como el objetivo de la aplicación de la técnica $\text{FDD}_\beta$ era descubrir

*nuevos* términos relevantes al dominio Irak, no se consideró "`Iraq`" como una palabra relevante a ser descubierta y por ende se la descartó de la lista de unigramas.

La detección de variables es una tarea que se la considera no totalmente automatizable, ya que para diferentes usuarios la definición de variable relevante puede variar. Por este motivo es que si bien las herramientas propuestas en este capítulo detectan variables de manera automática, se espera que el usuario del sistema sea el que finalmente decida, a partir de una lista reducida de variables posibles, si incorpora o no a cada una de estas. Para el caso de los términos, se espera que dada la lista de 30 palabras presentadas en la Tabla 4.4, el usuario decida qué variables le parecen relevantes para ser usadas en el próximo paso de detección de estructura causal. Vale aclarar que el usuario podría ajustar el valor de $\beta$ o la cantidad de términos a visualizar (en este caso 30) para tener más o diferentes opciones para elegir. A modo de ejemplo en este trabajo se eligen los 10 términos ($M = 10$) que se consideran más relevantes para el estudio: ('`weapons`', '`mass`', '`destruction`'), ('`Persian`', '`Gulf`', '`war`'), ('`United`', '`Nations`', '`Security`'), ('`Iraq`', '`invasion`', '`Kuwait`'), ('`chemical`', '`biological`', '`weapons`'), ('`military`', '`action`', '`Iraq`'), ('`United`', '`States`'), ('`war`', '`Iraq`'), ('`Saddam`', '`Hussein`') y ('`Bush`', '`administration`'). Varios trigramas son elegidos por tener una semántica fácil de interpretar y con mucha relevancia para el dominio. Por otra parte otros son ignorados por encontrarse representadas en otros. Por ejemplo (('`President`', '`Saddam`', '`Hussein`') subsume a los trigramas ('`Saddam`', '`Hussein`', '`Iraq`') y a ('`Saddam`', '`Hussein`'). Por ende se usa el término ('`Saddam`', '`Hussein`') que a la vez es el de mayor $\text{FDD}_\beta$ entre los tres. En la Tabla 4.4 además de mostrarse los diez unigramas, diez bigramas y diez trigramas con mayor $\text{FDD}_\beta$, también se indica con un asterisco los diez términos seleccionados previamente mencionados.

Finalmente, el conjunto de datos de menciones de términos se construye contando la frecuencia de aparición de cada uno de estos términos a lo largo de los 246 meses del periodo abarcado por el corpus de *The New York Times*. Las dimensiones finales del conjunto de datos de términos son de $246 \times 10$. En la Tabla 4.6 se reporta estadística descriptiva sobre el conjunto de datos de términos creado.

**Conjunto de datos de menciones de eventos.** Para poder construir un conjunto de datos de eventos en curso mencionados en el corpus del *New York Times* se realizaron los siguientes cuatro pasos. Primero, se utiliza el modelo presentado en el Capítulo 3 [MDT+21] para extraer los eventos en curso de todas las oraciones del corpus. Segundo,

| Unigrama | DISCR | DESCR | $FDD_\beta$ | Bigrama | DISCR | DESCR | $FDD_\beta$ | Trigrama | DISCR | DESCR | $FDD_\beta$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| war | 0,975 | 0,154 | 0,490 | ('United', 'States')* | 0,896 | 0,086 | 0,325 | ('President', 'Saddam', 'Hussein') | 0,998 | 0,014 | 0,073 |
| United | 0,920 | 0,144 | 0,460 | ('war', 'Iraq')* | 1,000 | 0,071 | 0,292 | ('weapons', 'mass', 'destruction')* | 0,996 | 0,012 | 0,061 |
| American | 0,902 | 0,131 | 0,432 | ('United', 'Nations') | 0,975 | 0,065 | 0,270 | ('Persian', 'Gulf', 'war')* | 0,995 | 0,011 | 0,056 |
| said | 0,623 | 0,181 | 0,428 | ('Saddam', 'Hussein')* | 0,994 | 0,041 | 0,189 | ('Saddam', 'Hussein', 'Iraq') | 1,000 | 0,010 | 0,052 |
| Mr. | 0,660 | 0,146 | 0,399 | ('Mr.', 'Bush') | 0,935 | 0,034 | 0,159 | ('United', 'Nations', 'Security'*) | 0,989 | 0,006 | 0,031 |
| Bush | 0,949 | 0,100 | 0,368 | ('President', 'Bush') | 0,965 | 0,033 | 0,154 | ('Nations', 'Security', 'Council') | 0,989 | 0,006 | 0,031 |
| States | 0,894 | 0,086 | 0,325 | ('Security', 'Council') | 0,989 | 0,028 | 0,134 | ('Iraq', 'invasion', 'Kuwait')* | 1,000 | 0,006 | 0,031 |
| military | 0,954 | 0,079 | 0,312 | ('Persian', 'Gulf') | 0,986 | 0,022 | 0,108 | ('Iraq', 'invaded', 'Kuwait') | 1,000 | 0,005 | 0,027 |
| Hussein | 0,991 | 0,069 | 0,285 | ('Bush', 'administration')* | 0,980 | 0,022 | 0,108 | ('chemical', 'biological', 'weapons')* | 1,000 | 0,004 | 0,022 |
| Iraqi | 0,985 | 0,067 | 0,279 | ('Mr.', 'Hussein') | 0,991 | 0,021 | 0,105 | ('military', 'action', 'Iraq')* | 1,000 | 0,004 | 0,022 |

Tabla 4.4: En esta tabla se presentan los 10 mejores unigramas, los 10 mejores bigramas y los 10 mejores trigramas de acuerdo a la técnica $FDD_\beta$ con $beta = 0,477$, siendo este valor de $\beta$ el que obtuvo el mejor desempeño como estimador de relevancia de términos para un dominio por parte de usuarios (análisis presentado en el Capítulo 2). Para cada término (unigrama, bigrama, trigrama) se reporta su poder descriptivo, su poder discriminativo y el puntaje de $FDD_\beta$ obtenido. Un usuario potencial podría elegir el $\beta$ y ajustar la cantidad de términos que se visualizan (en este caso 30) para luego manualmente elegir cuáles utilizar para el análisis causal. A modo de caso de uso se eligen los 10 que resultan más interesantes para el análisis posterior causal, estos 10 términos son marcados con un asterisco.

como cada evento es distinto (con diferentes contextos alrededor del *trigger*), para poder compararlos se construyó una representación vectorial para cada uno de ellos. Tercero, como cada representación vectorial es única, para poder agrupar menciones equivalentes (mismo evento con diferentes palabras) se aplicó una técnica de agrupamiento para encontrar $K$ grupos de menciones de eventos. Por último, el cuarto paso fue seleccionar de los $K$ grupos los $P$ más relevantes para ser utilizados como las variables del conjunto de datos. Cada grupo seleccionado es un evento distinto (variable) y cada instancia dentro del grupo constituye una mención de dicho evento.

Al igual que para el conjunto de datos de términos, se utiliza la totalidad del periodo cubierto por el *New York Times* (enero-1987 a junio-2007) usando frecuencia mensual (246 meses en total), resultando en un conjunto de datos de menciones de eventos con dimensión $246 \times P$. A continuación, se describen en detalle los cuatro pasos antes mencionados que permiten obtener las series de tiempo a partir de los textos completos del *New York Times*.

**Paso #1, detección de menciones de eventos en curso.** Al igual que para el conjunto de datos de menciones de términos se utiliza un dominio de interés para ilustrar un posible caso de uso de la herramienta. Nuevamente se utilizan todas las oraciones que mencionan al GPE "`Iraq`". Se utiliza el modelo de detección de eventos en curso presentado en el Capítulo 3 sobre las 170.497 oraciones que contienen al país de interés. Se detectan un total de 498.560 menciones de eventos en total (un promedio de 2,92 eventos por oración).

**Paso #2, construcción de una representación vectorial para cada mención.** A partir de cada una de las 498.560 menciones de eventos en curso, se define la tarea de agrupar menciones del mismo evento en un mismo grupo para poder construir la serie de tiempo de menciones de cada evento en el tiempo. Para poder construir los grupos se plantea el desafío de construir una representación vectorial para cada mención de tal manera que permita la comparación entre eventos. Se espera que eventos semánticamente similares estén cercanos en esta representación. Para poder comparar menciones de eventos se considera que no solo el *event-trigger* es importante sino todo su contexto. Por esta razón, en este trabajo se introduce la representación de la frase del evento (Event-Phrase Embedding Representation (EPER)), que se define como una suma de representaciones *GloVe* [PSM14] con un decaimiento cuadrático. Esto es, se considera como la parte más importante del evento al *event-trigger* y por tal motivo la representación *GloVe* de esta palabra es incluida en la representación sin penalización. Por otra parte, cada *token* a la

izquierda y a la derecha hasta terminar la oración, dan lugar a representaciones *GloVe*, las que son sumadas a la representación ajustadas por un coeficiente de penalización de tal manera que cuanto más lejos del *trigger* se encuentra el token representado, más fuerte es la penalización. Esta penalización es de orden cuadrático. La definición formal de la EPER se puede ver en la ecuación presentada a continuación:

$$\text{EPER}(e_k, P) = \sum_{w_i \in P} \frac{1}{(|k-i|+1)^2} \cdot \text{GloVe}(w_i) \tag{4.8}$$

En esta fórmula se tiene una frase $P$ compuesta de palabras $w_i$ definida como sigue: $P = w_1, w_2, \ldots, w_n$. Siendo el *event-trigger* $e_k$ la palabra $w_k$, para algún $k$, $1 \leq k \leq n$. De esta manera se agrega la representación de cada palabra de la oración, pero con un factor de penalización que crece cuadráticamente con la distancia al *event-trigger*. Así, se tiene una representación de 300 dimensiones de cada mención de evento que tiene en cuenta principalmente el *trigger* y en menor medida todas las palabras del contexto. Al finalizar la construcción de esta representación se tienen 498.560 vectores de 300 dimensiones, uno para cada mención de evento en curso detectado en las 170.497 oraciones que mencionan a Irak.

**Paso #3, aplicación de una técnica de agrupamiento para detectar menciones del mismo evento.** Finalmente, teniendo una representación vectorial para cada mención de evento se procede a realizar el agrupamiento planteado para poder unir los 498.560 vectores en $K$ grupos donde cada grupo representa el mismo evento semántico y cada instancia dentro del grupo representa una mención. El primer desafío que se plantea es la elección de la técnica de agrupamiento adecuada. Para el presente problema se emplea la clásica y ya bien establecida técnica de agrupamiento *KMeans* [Llo82].

El segundo desafío que se presenta es la elección del valor de $K$ (cantidad de clusters) adecuado para obtener la granularidad adecuada en cada grupo. Esto es, lo suficientemente grande como para tener múltiples menciones del mismo evento, sin comprometer la cohesividad del grupo y que se terminen agrupando eventos distintos. Para poder elegir el $K$ se procedió a utilizar la técnica gráfica *Elbow* [Tho53] para analizar el resultado de aplicar agrupamiento para diferentes valores de $K$. Se utiliza como métrica de cohesión de los grupos la distancia a los centroides al cuadrado (inercia). Se toma un conjunto de valores tentativos de $K$ ($K \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 50, 100, 500, 1.000, 5.000, 6.000, 7.000, 8.000, 9.000, 10.000, 50.000, 100.000\}$) y se grafica la inercia obtenida por cada

agrupamiento para esos valores de $K$. Debido a la gran cantidad de instancias, el costo computacional de la técnica *KMeans* y la cantidad de valores de $K$ a probar, para este análisis, no se pudo usar la técnica *KMeans* tradicional. En su lugar se utilizó la técnica *MiniBatch KMeans* [Scu10], la cual presenta modificaciones a la técnica original de tal modo que permite alcanzar soluciones similares a un costo computacional mucho menor. El resultado de este procedimiento se puede ver en la Figura 4.6.

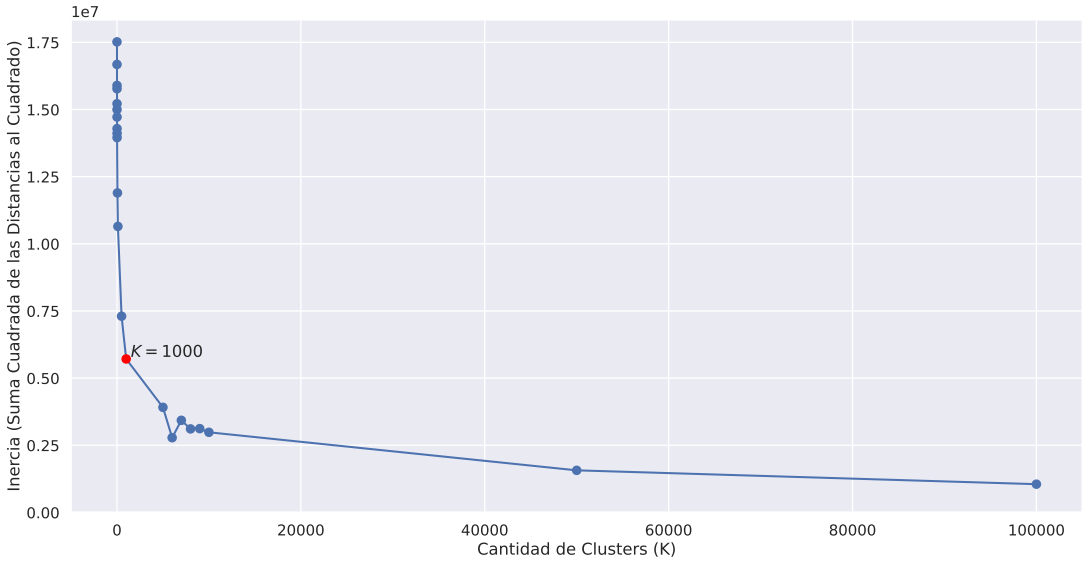



Figura 4.6: Visualización de la inercia de aplicar la técnica de agrupamiento *MiniBatch KMeans* para diferentes valores de $K$. La inercia se mide como la suma cuadrada de las distancias al centroide. A mayor $K$, mayor cantidad de grupos, y más cerca está el centroide de cada grupo a las instancias dentro del mismo, por ende, alcanzando una inercia menor. En el extremo se tendría un grupo por instancia, con un centroide que coincide con dicha instancia, en cuyo caso se tendría inercia cero. Se observa una caída pronunciada del valor de inercia hasta $K = 1.000$ (valor dibujado en rojo). Luego, cada aumento en $K$ no tiene un beneficio grande en disminución de inercia. Por este motivo se selecciona $K = 1.000$ como el valor óptimo de grupos a usar.

Como se esperaba para pocos grupos ($K$ pequeños) la distancia a los centroides es mayor. En el extremo, para $K = 1$ con un solo centroide en el centro de las 498.560 instancias, se tiene la máxima suma de distancias cuadradas, siendo esta igual a 17.516.440,97. En el otro extremo, con $K$ igual a la cantidad de instancias, se tiene que cada instancia es su propio grupo, y por ende cada centroide coincide con esa instancia. Esto da lugar a una inercia de cero. Se puede observar que hasta $K = 1.000$ se tenía una pendiente pronunciada, donde cada crecimiento en el valor de $K$ generaba una gran diferencia en la inercia resultante. A partir de ese valor se puede observar como la pendiente se achata.

Por este motivo, se concluye que el punto de inflexión más grande está en $K = 1.000$ y se continúa con ese valor de $K$ para los próximos pasos. Ya con el valor de $K$ definido, para construir el conjunto de datos final se utiliza el algoritmo *KMeans* tradicional (ya no su optimización *MiniBatch Kmeans*) con el valor de $K$ elegido ($K = 1.000$). La agrupación resultante tiene un valor de inercia de 5.341.146,00. Un histograma con la cantidad de menciones en cada grupo se reporta en la Figura 4.7.

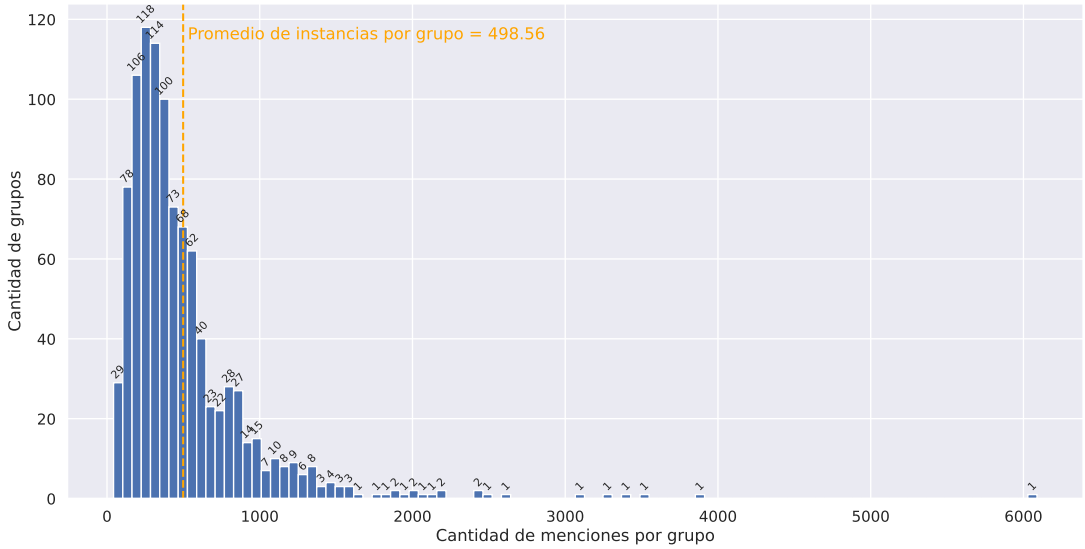



Figura 4.7: Histograma que muestra la dispersión de los tamaños de los grupos obtenidos al aplicar *KMeans* con $K = 1.000$. Como se puede observar, el promedio de instancias por grupo es de casi 500 instancias. Sin embargo, se pueden observar algunos grupos muy poblados: un grupo con más de 6.000 instancias, seis grupos con más de 3.000 instancias en cada uno. Por otra parte, hay 23 grupos con 103 elementos o menos y 107 con 163 elementos o menos.

**Paso #4, selección de los $P$ grupos más relevantes y construcción del conjunto de datos de eventos.** A partir del agrupamiento realizado por la técnica *KMeans* con $K = 1.000$ se analizan los 1.000 grupos resultantes en términos de la cohesión y cantidad de menciones por grupo ($p_i$). El objetivo es seleccionar grupos con alta cantidad de menciones para poder aplicar las técnicas de descubrimiento causal, y con alto nivel de cohesión para preferir grupos con una semántica específica y bien definida (de preferencia que se refieran a un único evento del mundo real bien definido). La cohesión, en este trabajo, se computa utilizando la distancia cuadrada promedio al centroide para cada grupo ($d_i$). Usando este valor, la cohesión del grupo ($c_i$) se define como $c_i = 1/(d_i + 1)$. De esta manera se trata de tomar grupos que maximicen ambos valores: $c_i$ y $p_i$. Para encontrar grupos que tengan un buen balance de ambos indicadores se plantea maximizar una fun-

ción ($g$) que calcule la media armónica entre estos dos valores. Para poder comparar los indicadores en la misma escala se los divide por el máximo de cada escala ($\max(c_i)$ y $\max(p_i)$) para llevarlos al rango $[0, 1]$. Luego la función $g$ queda definida como sigue:

$$g(c_i, p_i) = 2 \times \frac{c_i/\max(c_i) \times p_i/\max(p_i)}{c_i/\max(c_i) + p_i/\max(p_i)} \tag{4.9}$$

Se analizaron diferentes umbrales ($u$) para la función $g$ ($g(c_i, p_i) > u$) con $u \in [0,10;\ 0,40]$ con un paso de $0,01$. Por ejemplo para los valores de $u$ de $0,10$, $0,20$ y $0,30$, solo 374, 85 y 28 grupos, respectivamente, obtuvieron un valor de $g$ por encima del umbral fijado. Al buscar solo los grupos que estén por encima del umbral se prioriza al mismo tiempo (y por igual) la cohesividad y el tamaño del grupo de forma tal que los pequeños o poco cohesivos son descartados. Utilizando esta estrategia se realizó un filtrado de grupos de tal forma de obtener un conjunto reducido de grupos (20 o menos) altamente relevantes sobre los cuales elegir manualmente los más interesantes. Para obtener estos 20 grupos se buscó el mínimo umbral a partir del cual solo 20 instancias o menos son seleccionadas (umbral $u = 0,35$).

Al igual que para el conjunto de datos de términos, se considera que la herramienta debe hacer un trabajo automático para filtrar variables relevantes de un gran conjunto hasta obtener una cantidad manejable por un usuario (en este caso 20), y que sea el usuario quien elija las variables finales que quiere incluir en el modelo. Para este trabajo como caso de estudio se analizan manualmente los veinte grupos resultantes y se seleccionan seis, los cuales se identifican con las etiquetas: `C109`, `C165`, `C201`, `C249`, `C269` y `C550`. Las etiquetas de los grupos son creadas utilizando el número de grupo obtenido del algoritmo *KMeans* anteponiendo la letra C a dicho número (resultando en las siguientes posibles etiquetas: `C000`, `C001`, ..., `C998` y `C999`).

En la Figura 4.8 se pueden observar los 1.000 grupos en el plano donde en el eje horizontal se tienen la cohesión y en el eje vertical se tiene el tamaño de los grupos. Se dibuja la línea negra que representa el umbral $u = 0,35$ y en rojo los 20 grupos que quedan por encima del umbral. Los grupos representados con una cruz roja forman parte de los seis manualmente elegidos mientras que los marcados con círculos rojos son los que están por encima del umbral pero no fueron seleccionados. Se etiqueta cada grupo por encima del umbral con su nombre de grupo (en gris los no seleccionados y negro los seleccionados). Los 980 puntos azules por debajo de la línea son los grupos descartados automáticamente por no tener una buena combinación de cohesión y tamaño (están por debajo del umbral).

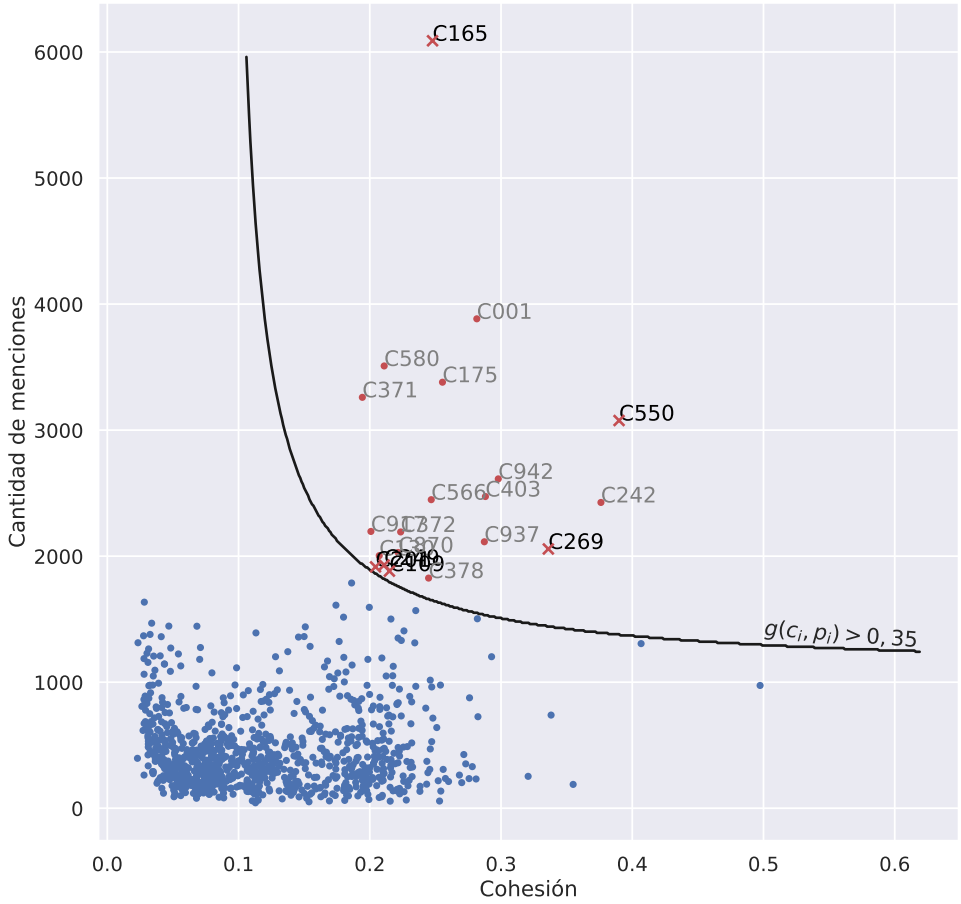



Figura 4.8: Gráfico de dispersión donde se visualizan los 1.000 grupos obtenidos con *KMeans* con $k = 1.000$. Se comparan los grupos en términos de cohesión y cantidad de menciones (instancias) dentro del grupo. La cohesión de cada grupo representa la similitud promedio de cada instancia del grupo con respecto a su centroide. Cuanto más grande el valor de cohesión mayor similitud entre las instancias dentro del grupo. Se buscan los mejores veinte grupos en términos de mayor cohesión y mayor cantidad de instancias. Para obtener los grupos que tengan un buen balance de estas dos métricas se computa la función $g$ que representa la media armónica de los valores normalizados de estas métricas para cada instancia. Se mueve el umbral hasta encontrar exactamente 20 grupos por encima del umbral establecido para la función $g$ (línea sólida negra), los 20 grupos elegidos se marcan en rojo (con cruz y círculo). De los veinte grupos se seleccionan manualmente seis (marcados con una cruz) para el análisis causal posterior. Para los veinte grupos en rojo se muestra la etiqueta del grupo, en negro para los seis manualmente seleccionados y en gris para los demás. Los puntos azules son los otros 980 grupos que **no** tienen un valor por encima del umbral ($g \leq 0,35$).

Para la selección manual de los grupos, como no es posible analizarlos en su totalidad (porque cada uno de los 20 grupos tiene cientos de menciones de eventos), se utilizó una representación visual de textos clásica: nubes de palabras. Se construyó una nube

de palabras para cada uno de los veinte grupos utilizando los mismos coeficientes de penalización usados para calcular los EPER (Ecuación 4.8). Esto es, para cada grupo se tomó cada mención de cada evento involucrado y se incluyó la totalidad de las palabras de la oración que contiene el evento, pero penalizando su importancia de acuerdo a qué tan lejos está del *event-trigger* de la mención. El *event-trigger* siempre es incluido sin penalización y cada palabra del contexto es incluida también en la nube de palabras, pero penalizada de manera cuadrática de acuerdo a la distancia al *event-trigger*. Las representaciones resultantes para los seis grupos elegidos se pueden ver en la Figura 4.9. En la Tabla 4.7 se reportan valores de estadística descriptiva del conjunto de eventos creado.

Como se puede observar en las nubes de palabras, se eligieron grupos que tuvieran una semántica clara y definida, y que puedan asociarse a eventos del mundo real. Por ejemplo, el grupo `C109` parece corresponderse con reportes de muertes durante la Guerra de Irak, tanto soldados (de ambos bandos) como civiles. Por otra parte, el grupo `C165` parece corresponderse con reportes de oposición a la guerra con Irak. El grupo `C201` parece tratarse principalmente de reportes de ataques terroristas, el grupo `C249` parece concentrarse en ataques o acciones militares (aparentemente por parte de Estados Unidos). El grupo `C269` trata sobre la invasión a Kuwait por parte de Irak, mientras que el grupo `550` reúne menciones de la guerra en Irak en general. Una descripción de los grupos es presentada en la Tabla 4.5.

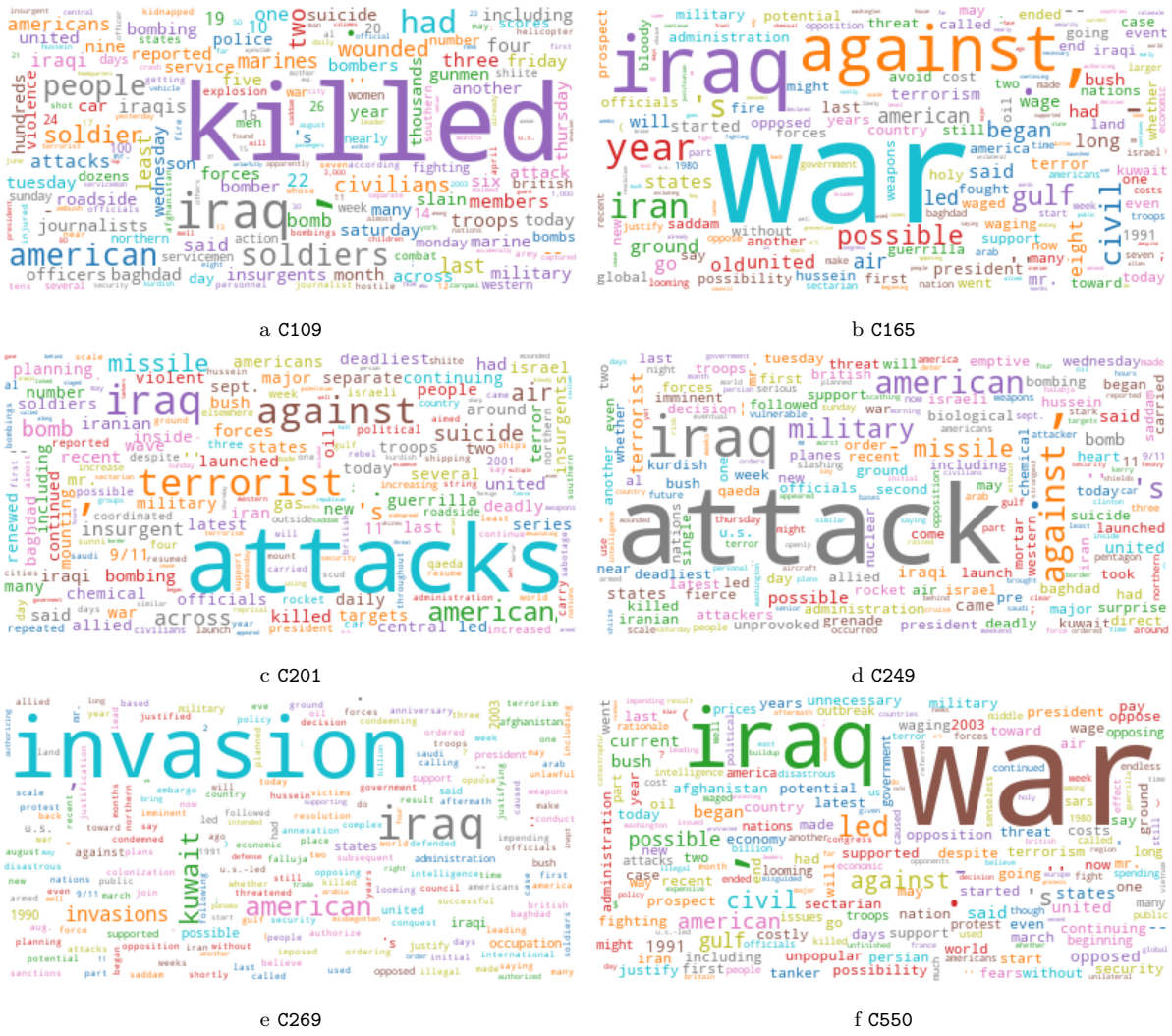
**Conjunto de datos de final.** Finalmente, luego de utilizar la técnica de pesaje de términos del Capítulo 2 para construir el conjunto de datos de términos (de dimensiones $246 \times 10$) y el modelo de detección de eventos en curso del Capítulo 3 para construir un conjunto de datos de eventos (de dimensiones $246 \times 6$) se construye un conjunto de datos final. Dado que el conjunto de datos de términos y el conjunto de datos de eventos en curso comparten la frecuencia (mensual) y el periodo de tiempo (enero-1987 a junio-2007), solo hay que juntar las variables en un solo conjunto de datos final (de dimensiones $246 \times 16$). Vale la pena mencionar que el propósito de utilizar las herramientas de los Capítulos 2 y 3 no es llegar a un único conjunto de datos con 16 variables, sino dar a un usuario potencial un conjunto manejable de variables, para que este usuario pueda seleccionar manualmente las variables que quiere utilizar para el aprendizaje de estructura causal. Si bien se pueden incorporar muchas más variables, es necesario que cada usuario elija las variables que considere relevantes para que la estructura causal resultante sea analizable de manera

| Etiqueta de grupo | Descripción |
|---|---|
| C109 | Reportes de muertes de soldados y civiles<br>(términos salientes: killed, Iraq, american, soldiers, civilians) |
| C165 | Menciones de la guerra de Irak, con una componente negativa<br>(términos salientes: against, war, iraq) |
| C201 | Reportes de ataques terroristas<br>(términos salientes: attacks, terrorist, iraq) |
| C249 | Menciones de acciones militares<br>(términos salientes: attack, iraq, military, missile, against |
| C269 | Menciones de la invasión de Kuwait<br>(términos salientes: invasion, iraq, kuwait, invasions, american) |
| C550 | Menciones de la Guerra de Irak<br>(términos salientes: war, iraq, led, 2003) |

Tabla 4.5: Descripción de seis eventos extraídos del corpus del *New York Times* para ser usados como variables del *framework* de recuperación de estructuras causales a partir de textos. Cada uno de estos eventos (o variables) consiste de múltiples menciones en diferentes textos del mismo evento. Por ejemplo, el grupo C109 consiste de múltiples menciones del mismo tipo de evento: reportes de muertes de soldados o civiles. Estos grupos son construidos a través de una técnica de agrupamiento (*KMeans*) la cual agrupó menciones de eventos semánticamente similares en el mismo grupo. Se toman seis de los 1.000 grupos formados, priorizando grupos con alta cohesión y mayor cantidad de menciones. Esto es, se necesitan variables que tengan una semántica bien definida y con muchas menciones.

sencilla y no esté superpoblada con demasiados nodos. En este trabajo en particular se redujo un gran vocabulario a 30 posibles términos. A la vez, se redujeron 1.000 eventos (grupos) a solo 20. En el presente trabajo se toman esas 50 variables sugeridas y se eligen 16 variables relevantes para mostrar un posible caso de uso de la herramienta.

Todos los conjuntos de datos sintéticos obtenidas de las fuentes *TETRAD* y *CauseMe*, como así también el conjunto de datos de demanda de energía eléctrica en el GBA (origen *CAMMESA*) se utilizan para evaluar las diferentes técnicas de aprendizaje de estructuras causales con el objetivo de elegir las mejores para ser usadas luego en los conjuntos de datos creados a partir de los textos del *New York Times*. Primero se presentan los resultados del análisis en los datos sintéticos en la Sección 4.4. Posteriormente se presenta el análisis sobre el conjunto de datos de *CAMMESA* en la Sección 4.5. Por último, en la Sección 4.6 tanto el conjunto de datos de términos ($246 \times 10$) como el de eventos ($246 \times 6$) como la




a C109 b C165

c C201 d C249

e C269 f C550

Figura 4.9: Seis nubes de palabras que describen a los seis grupos (*clusters*) elegidos manualmente a partir de las veinte opciones que fueron seleccionadas por tener mejor balance de cohesión y cantidad de menciones (los 20 por encima de la línea negra en la Figura 4.8). Estos seis grupos son los grupos Nro. 109, 165, 201, 249, 269 y 550 del total de 1.000 grupos construidos con *KMeans* ($K = 1.000$) (numerados desde el Nro. 0 hasta el 999). Como se puede observar existe una semántica identificable en cada grupo: (a) reporte de muertes debido a la Guerra en Irak (tanto civiles como soldados), (b) reportes de sentimientos en contra de la Guerra, (c) ataques terroristas, (d) ataques militares americanos contra Irak, (e) invasión de Kuwait por parte de Irak y (f) menciones de la Guerra de Irak.

unión de ambos ($246 \times 16$) son utilizados para aprendizaje de estructura causal usando las herramientas que parecen más apropiadas para el dominio. De esta manera se demuestra el prototipo completo de una herramienta causal que parte de información textual, extrae variables de interés y muestra vínculos causales entre ellas.

| | ('United', 'States') | ('war', 'Iraq') | ('Saddam', 'Hussein') | ('Bush', 'administration') | ('weapons', 'mass', 'destruction') | ('Persian', 'Gulf', 'war') | ('United', 'Nations', 'Security') | ('Iraq', 'invasion', 'Kuwait') | ('chemical', 'biological', 'weapons') | ('military', 'action', 'Iraq') |
|---|---|---|---|---|---|---|---|---|---|---|
| **mean** | 2699,73 | 49,37 | 98,07 | 144,33 | 21,80 | 33,76 | 20,39 | 4,05 | 7,33 | 2,84 |
| **std** | 562,88 | 105,56 | 159,61 | 214,57 | 35,47 | 48,38 | 18,38 | 13,94 | 12,12 | 8,34 |
| **min** | 1757 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **25 %** | 2291,25 | 2 | 15 | 0 | 4 | 9,25 | 8 | 0 | 1 | 0 |
| **50 %** | 2579 | 7 | 37 | 1 | 10 | 19 | 16 | 1 | 3 | 0 |
| **75 %** | 3008,75 | 59 | 104,5 | 365,75 | 25 | 35 | 25 | 3 | 8 | 2 |
| **max** | 4643 | 1007 | 1015 | 727 | 213 | 377 | 117 | 164 | 87 | 65 |

Tabla 4.6: Estadística descriptiva del conjunto de datos de términos (unigramas, bigramas y trigramas) generado. Se reporta (de abajo para arriba): el promedio, el desvío estándar, el valor mínimo, los valores de los percentiles 25 %, 50 % y 75 %, y el valor máximo. El conjunto consiste de diez variables (términos) con una longitud de serie de 246 meses (frecuencia mensual). Para cada variable en cada instante de tiempo se reporta la cantidad de menciones de dicha variable en ese instante de tiempo (mes).

| | **C109** | **C165** | **C201** | **C249** | **C269** | **C550** |
|---|---|---|---|---|---|---|
| **mean** | 7,65 | 24,75 | 7,78 | 7,85 | 8,36 | 12,50 |
| **std** | 13,62 | 53,79 | 12,16 | 14,68 | 14,17 | 36,48 |
| **min** | 0 | 0 | 0 | 0 | 0 | 0 |
| **25 %** | 0 | 3 | 0 | 0 | 0 | 0 |
| **50 %** | 1 | 8,5 | 2 | 2 | 1 | 1 |
| **75 %** | 5 | 24,75 | 11,75 | 9 | 13 | 9,75 |
| **max** | 77 | 534 | 67 | 118 | 89 | 357 |

Tabla 4.7: Estadística descriptiva del conjunto de datos de eventos en curso generado. Se reporta (de abajo hacia arriba): el promedio, el desvío estándar, el valor mínimo, los valores de los percentiles 25 %, 50 % y 75 %, y el valor máximo. El conjunto consiste de seis variables (eventos) con una longitud de serie de 246 meses (frecuencia mensual). Para cada variable en cada instante de tiempo se reporta la cantidad de menciones de dicha variable en ese instante de tiempo (mes).

## 4.4. Aplicación a Datos Sintéticos

En la presente sección se presentan los resultados y discusiones obtenidas del análisis comparativo realizado de las técnicas de descubrimiento causal aplicadas a datos de origen sintéticos. Estos son, los datos generados a partir de la herramienta *TETRAD* y los obtenidos de la plataforma *CauseMe*. Primero, las nueve técnicas del estado del arte para descubrimiento causal, mencionadas en la Sección 4.2, son analizadas en los 56 conjuntos de datos generados con *TETRAD*. Cada uno de estos conjuntos tienen diferentes características porque son generados usando diferentes configuraciones de la herramienta: variando cantidad de variables, longitud de la serie, cantidad de variables ocultas, cantidad de rezagos en el modelo causal real y variando la estrategia de generación de grafos (*RFDAG* o *SFDAG*). Los resultados y discusiones de aplicar estas nueve técnicas sobre los conjuntos de datos generados con *TETRAD* se reportan en la Sección 4.4.1. La gran cantidad y diversidad de los conjuntos de datos generados a partir de esta herramienta permitieron sacar conclusiones fuertemente fundamentadas sobre el desempeño de las nueve técnicas ante distintos escenarios. Posteriormente, utilizando estos experimentos se seleccionan las cuatro mejores técnicas para ser usadas en los experimentos realizados sobre el resto de los conjuntos de datos sintéticos: los obtenidos de la plataforma *CauseMe*. Los datos obtenidos de dicha plataforma agregan aún más diversidad al análisis al incluir relaciones causales no lineales. Los resultados de aplicar las cuatro mejores técnicas de descubrimiento causal sobre las ocho bases de datos obtenidas de *CauseMe* son reportados en la Sección 4.4.2. En la misma sección se presenta una discusión sobre dichos resultados.

### 4.4.1. Análisis Comparativo en *TETRAD*

En la presente sección se reportan los resultados de aplicar las nueve técnicas del estado del arte reportadas en la Sección 4.2 (*BigVAR*, *Direct-LiNGAM*, *ICA-LiNGAM*, *Lasso-Granger*, *PC*, *PCMCI*, *SIMoNe*, *Transfer Entropy* y *VAR*) sobre el conjunto de datos sintético originado con la herramienta de simulación de datos *TETRAD*. Utilizando dicha herramienta de simulación se crean 56 conjuntos de datos diferentes con diferentes configuraciones de acuerdo a lo descrito en la Sección 4.3.1 para ser utilizados como parte del presente análisis comparativo para estudiar la diferencia entre las técnicas de causalidad presentadas. Se agrega una décima técnica de causalidad que agrega arcos de manera aleatoria. Esta técnica de referencia, a la que se denomina *Random*, se incorpora

para poder establecer un desempeño mínimo a superar por las demás técnicas. Para construir los grafos causales la técnica *Random* analiza cada posible arco de un grafo totalmente conexo y aleatoriamente con proporción 50-50 decide si incorpora el arco al modelo causal resultante o no.

Como se mencionó en la Sección 4.3.1, los 56 conjuntos de datos se pueden dividir en cuatro escenarios de acuerdo a la configuración utilizada para generarlos. A su vez cada escenario pudo haber sido generado utilizando una de dos posibles técnicas de construcción de grafos acíclicos dirigidos (DAG por sus siglas en inglés): *scale-free DAG* (*SFDAG*) o *random forward DAG* (*RFDAG*). Finalmente, los conjuntos de datos se dividen en ocho categorías: escenarios 1 a 4 para *SFDAG* y escenarios 1 a 4 para *RFDAG*. Se reportan los resultados para cada una de esas ocho categorías en la presente sección.

En la Figura 4.10 se muestran los resultados para cuatro categorías: escenarios 1 y 2 tanto para *SFDAG* (derecha) como para *RFDAG* (izquierda). En la Figura 4.11 se muestran los resultados de las otras cuatro categorías: escenarios 3 y 4 tanto para *SFDAG* (derecha) como para *RFDAG* (izquierda). Debido a la gran cantidad de resultados a reportar, de este análisis se excluye la precisión y la cobertura (que son analizados posteriormente). Solo se reporta para cada una de las ocho categorías el F1-score obtenido por las diez técnicas para cada una de las diferentes variaciones de parámetros. Por ejemplo, para el escenario 1 usando *RFDAG* se reportan los nueve valores de F1-score obtenidos (uno por cada $N$ utilizado) para un total de diez técnicas (noventa valores de F1-score son reportados en total). Estos 90 valores son reportados en el gráfico en la esquina superior izquierda de la Figura 4.10 ("Escenario 1 - *RFDAG*").

A continuación, en la Figura 4.12, se incluye al presente análisis una comparación de las técnicas en términos de las métricas precisión, cobertura y F1-score promedio para cada escenario para cada configuración de DAG (*SFDAG* (derecha) y *RDDAG* (izquierda)). En la figura se puede observar seis gráficos de barra. Se reportan en la primera fila los gráficos de barra con los valores de precisión para todas las técnicas en los cuatro escenarios usando *RFDAG* y luego los valores de precisión para todas las técnicas en los cuatro escenarios para *SFDAG*. Análogamente, en la segunda fila se reportan los valores promedios de cobertura, primero para *RFDAG* y luego para *SFDAG*. Finalmente, en la última fila se reportan los valores promedios de F1-score para ambas estrategias de construcción de DAG (usando el mismo orden que para las dos anteriores). Además de los valores promedio se reportan los intervalos de confianza usando nivel de confianza de 95 %. Por ejemplo,

se computan los valores de precisión, cobertura y F1-score para cada uno de los nueve valores de $N$ usando la técnica *BigVAR* para el escenario 1 usando *RFDAG*, se promedian los nueve valores de precisión, cobertura y F1-score y se los reporta en la primera barra de los gráficos que están en la columna de la izquierda. Estos tres valores son representados con tres barras, una por gráfico, y se les agrega a cada una la visualización del intervalo de confianza.

Por último, en la Figura 4.13, se reporta nuevamente la precisión, cobertura y F1-score promedio pero en este caso sin distinguir entre escenarios. Esto es, se reporta en la columna izquierda la precisión, cobertura y F1-score promedio para cada técnica en la totalidad de los conjuntos de datos *RFDAG* (28 conjuntos). Análogamente, en el lado derecho se reportan las mismas métricas promediadas para los 28 conjuntos *SFDAG*. Además de los valores promedio se reportan los intervalos de confianza usando nivel de confianza de 95 %. Por ejemplo, se computan los valores de precisión, cobertura y F1-score usando la técnica *BigVAR* en todos los escenarios usando *RFDAG*, se promedian los valores de precisión, cobertura y F1-score y se los reporta en la primera barra de los gráficos que están en la columna de la izquierda. Estos tres valores son representados con tres barras, una por gráfico, y se les agrega a cada una la visualización del intervalo de confianza.

Para el cómputo de todas las métricas (precisión, cobertura y F1-score) no se utiliza como *ground truth* la totalidad de los arcos de la estructura real, ya que esta puede contener vínculos causales contemporáneos o vínculos con distancia mayor a un intervalo de tiempo. Estos dos tipos de arcos escapan al análisis aquí presentado. Esto es, los arcos contemporáneos no son encontrados por varias de las técnicas de causalidad, por ende son descartados del análisis y por simplicidad (sin disminuir la complejidad de la tarea) solo se buscan vínculos causales de distancia uno. Esto no simplifica la tarea, ya que vínculos con mayor distancia existen (en el escenario 4) y pueden complicar la tarea de encontrar los vínculos de distancia uno correctos.

**Discusión de los resultados obtenidos sobre los conjuntos de datos generados con *TETRAD*.** Como se puede ver en la Figura 4.10, para **el escenario 1** (donde se varía la cantidad de nodos ($N$)) se puede ver una diferencia marcada entre las cinco técnicas con mejor desempeño (*BigVAR*, *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*) y las cuatro técnicas con peor desempeño (*ICA-LiNGAM*, *Lasso-Granger*, *SIMoNe* y *Transfer Entropy*). Se puede observar que para pocos nodos algunas técnicas tuvieron un desempeño muy distinto al obtenido para muchos nodos. Por ejemplo, para ***RFDAG*** (izquierda)

con $N = 6$, *ICA-LiNGAM* obtuvo F1-score= 0, mientras que, para el mismo $N$ y mismo DAG, *SIMoNe* tuvo un desempeño inusualmente bueno comparado con su desempeño para otros valores de $N$. Esto se puede explicar debido al impacto del azar. Al tener menos nodos, y por ende menos arcos, cada decisión de la técnica (cada error o acierto) impacta mucho más en el desempeño que cuando el $N$ es más grande (porque, al haber pocos nodos, cada arco representa una proporción mayor del total). Por este motivo se considera que los valores de F1-score para $N < 12$ pueden ser ruidosos y no son buenos representantes del desempeño general de la técnica.

Aunque los valores de $N$ pequeños no son buenos representantes del desempeño, también existen técnicas que reportan valores poco estables de desempeño para todos los valores de $N$. Por ejemplo podemos ver que *ICA-LiNGAM* para el escenario 1 con *RFDAG* comienza con F1-score= 0 (peor que *Random*) y luego tiene su mejor desempeño en $N = 15$ para luego presentar una caída en el desempeño para $N > 15$. De manera similar, esa misma técnica comienza con un mal desempeño para *SFDAG*, mejora apenas por encima de *Random* y luego su desempeño vuelve a bajar por debajo de *Random* para $N = 15$ (el valor de $N$ que para el otro tipo de DAG obtuvo su mejor desempeño). Se puede concluir que *ICA-LiNGAM* no solo tiene un mal desempeño global, sino que también es poco consistente.

**Para el caso del escenario 1 con *RFDAG*** se puede ver que, en su mayoría, las técnicas tienen buena estabilidad en el desempeño ante variaciones en el valor del $N$. Esto se observa en general, excepto para para valores pequeños de $N$ y para las técnicas *PC* y *BigVAR* que para $N = 30$ muestran una clara caída en el valor del F1-score. Por otro lado **para el caso del escenario 1 con *SFDAG*** se puede observar que las cinco mejores técnicas siguen siendo las mismas pero con menos diferencia en desempeño con las demás, y esta diferencia se hace aún menor para $N = 6$ y $N = 9$ (por lo previamente discutido). Se puede ver nuevamente que *BigVAR* presenta una caída en el desempeño para $N = 30$, mientras que las demás técnicas mantienen mayor estabilidad en el valor de F1-score ante variaciones en el $N$.

**Para el escenario 2 usando *RFDAG*** (Figura 4.10) se puede observar una diferencia menos pronunciada entre las mejores técnicas y las peores. Nuevamente las técnicas *Direct-Lingam*, *PCMCI*, *PC* y *VAR* se ubican entre las mejores. La técnica *BigVAR* pasó de tener un buen desempeño para el escenario 1, a tener un desempeño muy poco estable y peor que *Random* para algunos valores de $N$ ($N = 30$ y $N = 27$). Se puede observar

que las técnicas *SIMoNe*, *Transfer Entropy* y *Lasso-Granger* tienen un desempeño bajo, comparable a la técnica *Random*. **Para el caso de *SFDAG*** se puede observar una mayor diferencia entre las mismas cuatro mejores técnicas y las demás. *BigVAR* nuevamente presenta un desempeño poco consistente, comenzando con buen desempeño, cayendo a un mínimo para $T = 1.000$ y volviendo a subir casi hasta obtener el mismo F1-score que las mejores técnicas. Se puede ver que en algunas ocasiones esta técnica alcanza buenos desempeños, pero no de manera consistente. Como no es posible identificar un conjunto de características que deban cumplir los datos para que *BigVAR* tenga un desempeño consistentemente bueno, no es una técnica que se considere como entre las mejores, sino que es considerada de bajo desempeño como *ICA-LiNGAM*, *Lasso-Granger*, *Transfer Entropy* y *SIMoNe*. En términos generales, el tamaño de la serie de datos (T) no parece tener un impacto significativo (ni positivo ni negativo) en el desempeño de las técnicas en general.

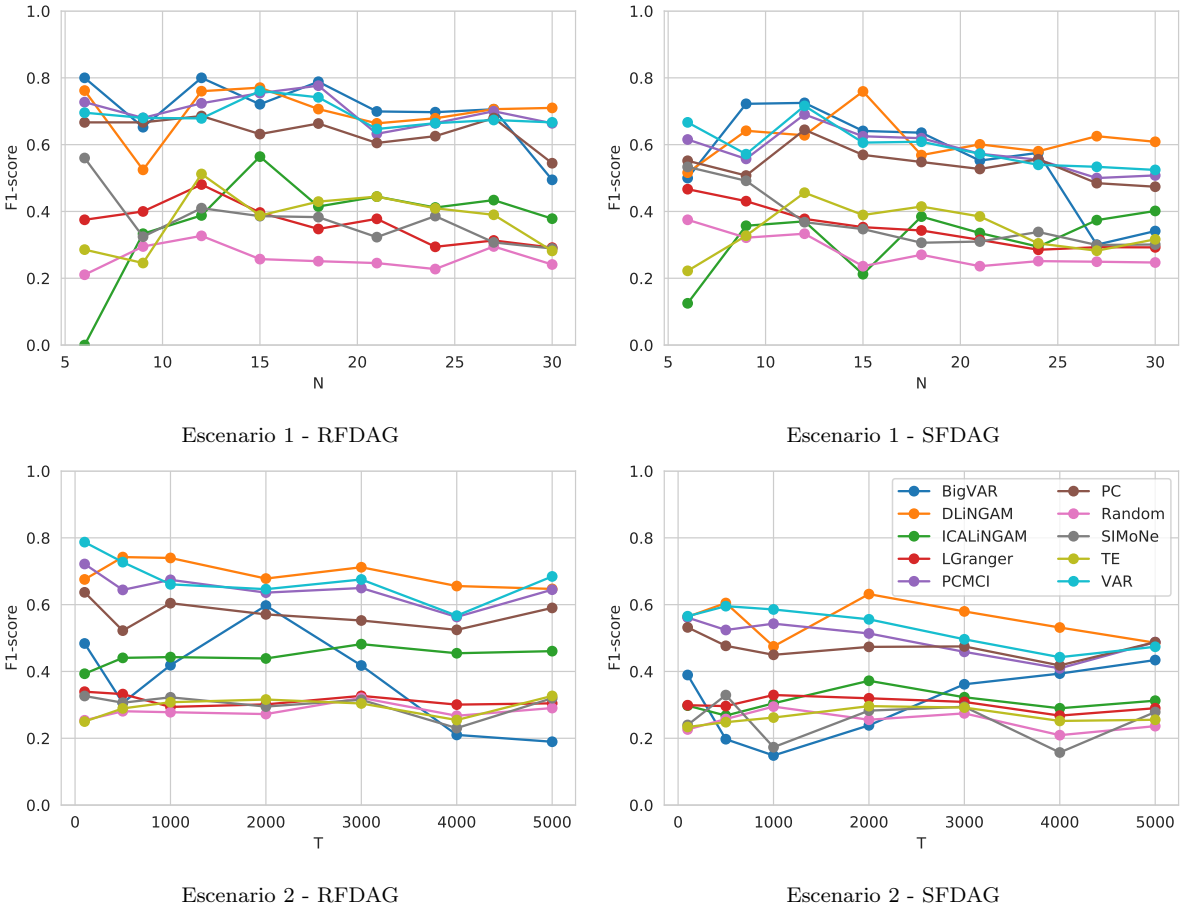
Se reportan los resultados **para el escenario 3** en la Figura 4.11, donde se agregan por primera vez variables no observadas ($H$), variando la cantidad de las mismas. **Para el caso de *RFDAG*** se puede observar que *Transfer Entropy* tiene un desempeño peor que *Random*. También se puede ver, una vez más, la inestabilidad de *BigVAR* y como las mismas cuatro técnicas siguen siendo las que tienen mejor desempeño. Los resultados sugieren que existe una cierta estabilidad de las técnicas hasta $H = 8$, punto a partir del cual todas las técnicas (excepto *BigVAR*) comienzan a tener peor desempeño. **Para el caso de *SFDAG*** se puede observar valores de F1-score menores que para el caso de *RFDAG*, sugiriendo que los conjuntos construidos con *SFDAG* son más complejos de resolver para las técnicas. Por otro lado, no se observa una tendencia clara de parte de todas las técnicas de pérdida de desempeño al aumentar el valor de $H$ (como era el caso para *RFDAG*). Sin embargo, se observa una tendencia a la baja para algunas técnicas, como ser el caso de *Direct-LiNGAM* o *VAR* que tienen una tendencia a la baja de desempeño al aumentar $H$. Nuevamente *BigVAR* presenta un desempeño poco consistente pero esta vez con muchos valores peores que *Random* (para $H \in \{2, 4, 8, 10\}$). Las cuatro mismas mejores técnicas siguen teniendo una diferencia por encima de las otras cinco, pero no tan marcada en este escenario para *SFDAG*.

Se reportan los resultados **para el escenario 4** en la Figura 4.11, donde el modelo causal además de tener relaciones causales de un instante al siguiente (distancia uno ($L = 1$)), como todos los escenarios anteriores, ahora se agregan vínculos causales a mayor distancia en el tiempo ($L \in \{1, 2, 3, 4, 5\}$). **Para el caso de *RFDAG*** se puede observar

nuevamente que para valores bajos de $L$, especialmente para $L = 1$ (el valor usado en los escenarios anteriores), existe una diferencia entre las mejores cuatro técnicas y las demás (a excepción de *BigVAR* que en este caso tuvo desempeño comparable al de las cuatro mejores técnicas). Sin embargo, a medida que aumenta el $L$ se observa una clara pérdida de desempeño por parte de las mejores técnicas, alcanzándose un desempeño relativamente similar para todas las técnicas. Por último, **para el caso de *SFDAG***, se puede observar cómo las cuatro mejores técnicas comienzan con valores casi equivalentes de F1-score para $L = 1$, y su desempeño cae rápidamente al crecer el valor de $L$. Una vez más se observa que *BigVAR*, que suele tener esporádicos buenos desempeños para algunas configuraciones en algunos conjuntos de datos, presenta un desempeño inconsistente y peor que *Random* para varios valores de $L$. Para este escenario y tipo de DAG se puede observar que hubo un peor desempeño en términos generales, donde varias técnicas tuvieron desempeños cercanos o peores a *Random* (incluso *PC* y *Direct-LiNGAM* que pertenecen al grupo de las cuatro mejores técnicas).

**En resumen**, en términos generales se concluye que las técnicas que tuvieron mejor desempeño son *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*, siendo de estas cuatro *PC* la que tiene peor desempeño en términos generales. De las otras tres no hay una técnica que sea superior para todos los escenarios y los parámetros. En términos generales no parece haber un gran impacto del valor de $N$ y de $T$ excepto para algunos casos puntuales. Por otro lado, al crecer el valor de $H$ para *RFDAG*, hay evidencia que sugiere una caída de desempeño. En el caso de *SFDAG* esta caída existe para algunas técnicas pero dicha caída no es pronunciada. Para el caso de $L$ en *RFDAG* se puede observar una caída para el valor de $L = 5$ con respecto a todos los demás. Mientras que este no es el caso para *SFDAG*. En términos globales las técnicas tuvieron peor desempeño para *SFDAG* que para *RFDAG*, más aún el escenario más difícil fue el 4 con la configuración *SFDAG*. *BigVAR* fue la técnica con peor consistencia, alcanzado de forma alternada los mejores y peores resultados, sin respetar ningún patrón evidente de características que marcasen un mejor o peor desempeño.

Un análisis similar se desprende de analizar **el desempeño en términos de las métricas precisión, cobertura y F1-score** disponibles en la Figura 4.12. Por ejemplo, se puede ver que **para el caso de *RFDAG* los mejores valores de precisión** son alcanzados por las mejores cuatro técnicas y *BigVAR*. Este valor alto de precisión para *BigVAR* es esperable ya que al ser un modelo VAR penalizado se espera que tenga mayor-




Escenario 1 - RFDAG

Escenario 1 - SFDAG

Escenario 2 - RFDAG

Escenario 2 - SFDAG

Figura 4.10: Desempeño, medido en términos de F1-score, de las nueve técnicas del estado del arte y el modelo de referencia (*baseline*) *Random* sobre 32 conjuntos de datos sintéticos generados con la herramienta de simulación de *TETRAD*. Estos conjuntos de datos se dividen en cuatro: (i) escenario 1 construido usando el enfoque *RFDAG*, (ii) escenario 1 construido usando el enfoque *SFDAG*, (iii) escenario 2 usando el enfoque *RFDAG* y (iv) escenario 2 usando el enfoque *SFDAG*. Estas cuatro categorías (i, ii, iii, iv) tienen 9, 9, 7 y 7 conjuntos de datos cada una, respectivamente. El escenario 1 mantiene la misma configuración, pero modificando la cantidad de nodos ($N$). El escenario 2 mantiene la misma configuración, pero variando la longitud de la serie ($T$). Se puede observar que para la métrica considerada las mejores técnicas son *Direct-LiNGAM* (DLiNGAM), *PCMCI*, *PC* y *VAR*. Mientras que las peores técnicas son: *Lasso-Granger* (LGranger), *SIMoNe*, *Transfer Entropy* (TE), e *ICA-LiNGAM*. Se puede ver que la técnica *BigVAR* tiene un desempeño poco consistente. En términos de los parámetros considerados ($N$ y $T$) no se observan diferencias significativas en desempeño. Solo se observa desempeños menos consistentes para $N$ pequeño en *SIMoNe* (con un desempeño inusualmente alto con $N = 6$), *ICA-LiNGAM* (con F1-score cero para $N = 6$) y *Direct-LiNGAM* (con desempeño inusualmente pequeño para $N = 6$).

mente coeficientes cero (penalizados) y solo unos pocos coeficientes significativos distintos de cero (muy precisos). Se puede observar que *ICA-LiNGAM* tiene buenos valores de

precisión promedio, aunque inferiores a los de las mejores cuatro técnicas y con intervalos de confianza mayores en la mayoría de los casos. Finalmente se puede observar que el escenario cuatro es el más difícil, ya que todas las técnicas obtienen valores de precisión más bajos en comparación y hay menos distancia entre las mejores y las peores técnicas. Un escenario similar se da para los valores de **precisión obtenidos con *SFDAG***, donde las mismas cuatro técnicas y *BigVAR* sobresalen en precisión, y donde se puede ver un escenario 4 con menores valores de precisión indicando la dificultad de dicho escenario. Se observa también, en líneas generales, peor desempeño en términos de precisión para la tarea usando *SFDAG* comparado con el desempeño usando *RFDAG*. La técnica con los peores valores de precisión es *Random*, aunque es igualada en mal desempeño por varias técnicas en varias configuraciones.

Para el caso de la **cobertura promedio**, se puede observar que para el **escenario 1 en *RFDAG*** existe una clara diferencia entre las cuatro mejores técnicas y las demás. En este caso se puede ver el bajo desempeño de *BigVAR* (lo cual es esperable ya que por construcción prioriza precisión) y además se puede observar que tiene un intervalo de confianza grande, indicando baja consistencia en sus resultados. Se puede ver, por ejemplo, por parte de técnicas como *Lasso-Granger*, que si bien mostraron mal desempeño en términos de F1-score, para el caso de cobertura el desempeño es mejor. Dicha técnica alcanzó valores de cobertura similares a los de las cuatro mejores técnicas, confirmando que es una técnica que prioriza la cobertura obteniendo una mala precisión. **Para el caso de *SFDAG*** se observa un comportamiento similar de las técnicas en términos de cobertura promedio: las cuatro mejores técnicas alcanzaron los mejores valores de cobertura, la técnica *BigVAR* obtuvo valores considerablemente bajos, y la técnica *Lasso-Granger* una vez más tuvo valores de cobertura altos, casi comparables con las cuatro mejores técnicas. Otra técnica que sobresale para algunos escenarios es *SIMoNe*, que también obtiene buenos valores de cobertura promedio en varias configuraciones. La técnica que tiene los peores valores de cobertura es *ICA-LiNGAM*.

Por último, de los **valores de F1-score promedio** reportados en la Figura 4.12 se extraen conclusiones similares a las obtenidas de los valores de F1-score no promediados reportados en 4.10 y 4.11. Esto es, se puede ver el desempeño superior de las cuatro técnicas **para el caso de *RFDAG***. También se puede ver un desempeño bueno por parte de *BigVAR* pero con intervalos de confianza mayores (debido a su inconsistencia). Adicionalmente se puede ver que *ICA-LiNGAM* obtiene, en términos generales, desempeños

malos y presenta intervalos de confianza grandes. Se puede ver que la técnica *Random* es la que tiene peor desempeño excepto para el escenario 3 (donde *Transfer Entropy* tiene el peor desempeño). Adicionalmente, si bien *Random* es la peor técnica, se puede observar que no tiene una diferencia significativa con algunas técnicas en algunos escenarios. Por ejemplo, para el escenario 4 debido al gran intervalo de confianza de *ICA-LiNGAM* no se puede asegurar que sus desempeños sean estadísticamente distintos, indicando que no hay una diferencia significativa entre *Random* e ICA-LiNGAM para este escenario con *RFDAG*. Por último, **para el caso de *SFDAG*** en el escenario 4, se puede ver que las mejores cuatro técnicas siguen manteniendo resultados consistentes mientras *BigVAR* muestra una gran caída de desempeño y un gran aumento en los intervalos de confianza. La dificultad de este escenario se vuelve evidente ya que la diferencia entre las mejores y las peores técnicas se hace menor. Finalmente, para las técnicas con bajo desempeño se puede observar un valor de F1-score promedio muy cercano a *Random* (o peor para algunas técnicas en algunos escenarios).

**Un tercer análisis** se desprende de la Figura 4.13 donde se reporta nuevamente la **precisión, cobertura y F1-score promedio pero en este caso sin distinguir entre escenarios**. Se puede observar que tanto **para *RFDAG* como para *SFDAG*** se tiene que las mejores cinco técnicas en términos de precisión (de mejor a peor) son: *BigVAR*, *Direct-LiNGAM*, *PCMCI*, *VAR* y *PC*. En términos de cobertura se puede ver que **para *RFDAG*** las mejores cuatro técnicas son: *VAR*, *PCMCI*, *Direct-LiNGAM* y *PC*. **Para el caso de *SFDAG*** las cuatro mejores técnicas son las mismas, solo invirtiendo el orden de *Direct-LiNGAM* con *PC*. Por último, en términos de F1-score se puede apreciar que, **tanto para *RFDAG* como para *SFDAG***, las mejores cuatro técnicas son (de mejor a peor): *Direct-LiNGAM*, *PCMCI*, *VAR* y *PC*. En esta Figura se puede apreciar un intervalo de confianza muy grande para la técnica *BigVAR* dando, nuevamente, indicios de sus inconsistencias.

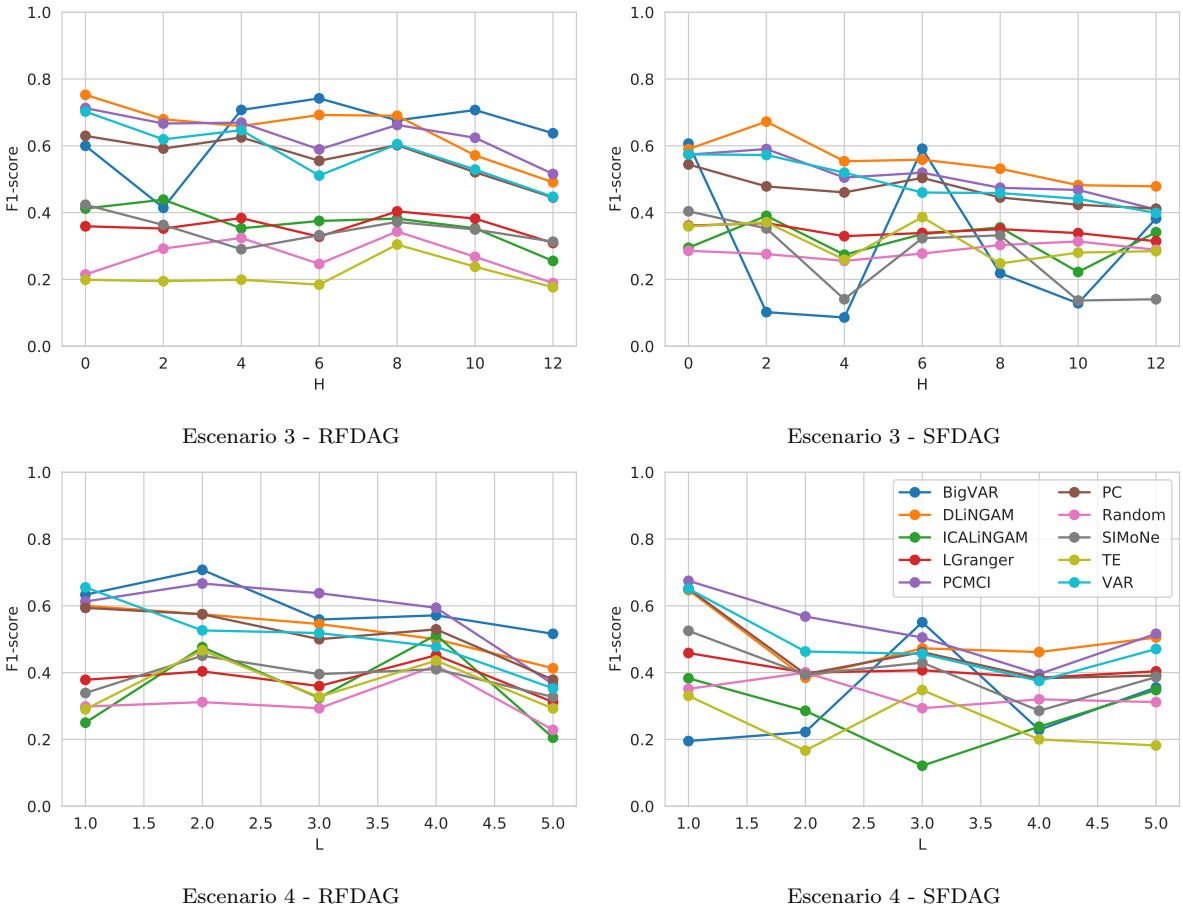
También se puede observar que, si bien *Random* fue la peor técnica para ambas configuraciones (***RFDAG* y *SFDAG***), las técnicas con peor desempeño (*BigVAR*, *ICA-LiNGAM*, *Lasso-Granger*, *SIMoNe* y *Transfer Entropy*) **en la configuración *SFDAG*** no tuvieron un desempeño significativamente mejor que este modelo de referencia (*baseline*). Esto da indicios de que las técnicas no tienen buen desempeño en general y de que el escenario *SFDAG* es más difícil que el *RFDAG*, esto se vuelve especialmente evidente al notar que la diferencia entre las peores y las mejores técnicas es menor para esta

configuración.

**Para el caso de *RFDAG***, la técnica *Random* no fue significativamente distinta a *ICA-LiNGAM*, *Lasso-Granger*, *SIMoNe* y *Transfer Entropy*. El bajo desempeño por parte de estas cuatro técnicas y *BigVAR* puede ser explicado por diferentes motivos. Por ejemplo, para el caso de *ICA-LiNGAM*, una serie de limitaciones fueron descritas por los autores en publicaciones posteriores donde propusieron a *Direct-LiNGAM* como una mejora a ICA-LiNGAM [SIS+11]. Estas limitaciones son mencionadas en la Sección 4.2). Para el caso de *Transfer Entropy*, la técnica puede estar limitada por tratarse de una técnica que analiza causalidad de a pares (lo cual deja fuera del modelo variables potencialmente relevantes para el descubrimiento causal). Similarmente, para el caso de *Lasso-Granger*, si bien se aplica una etapa de selección de variables con la técnica lasso, el proceso posterior de descubrimiento causal es utilizando la técnica de causalidad de Granger de a pares, que puede tener las mismas limitaciones que *Transfer Entropy*. Por último, las técnicas basadas en modelos VAR penalizados, *BigVAR* y *SIMoNe*, tienen valores de cobertura bajos con respecto a las mejores técnicas, probablemente por la componente de penalización que afecta negativamente a esta métrica. Aunque ambas técnicas tienen valores de cobertura bajos, la diferencia entre estas técnicas es en su desempeño en términos de precisión. *BigVAR* tiene valores altos de precisión y bajos de cobertura, consistentes con una técnica que encuentra mayormente arcos correctos y penaliza fuertemente a los que sean probablemente negativos. Mientras que *SIMoNe* tiene valores bajos para ambas métricas, esto quiere decir que la técnica recupera muchos arcos incorrectos (baja precisión) y el proceso de penalización afecta a la cobertura dando como resultado una técnica que no resulta útil para el descubrimiento causal.

La variedad y cantidad de los datos generados a partir de la herramienta *TETRAD* permite sacar conclusiones fundamentadas sobre cuáles técnicas aportan al descubrimiento causal en este dominio y para este tipo de conjuntos de datos (y cuales técnicas tienen un desempeño cercano a *Random*). A partir de este estudio, como se pudo observar un desempeño consistentemente bueno por parte de las técnicas *Direct-LiNGAM*, *PCMCI*, *PC* y *VAR*, esas cuatro técnicas son seleccionadas para continuar el análisis comparativo de herramientas causales. Para sumar a este análisis se incorporan los conjuntos de datos obtenidos de la plataforma *CauseMe* y de la compañía *CAMMESA*. Sobre esos conjuntos de datos se comparan las cuatro mejores técnicas para sacar conclusiones sobre su desempeño. El objetivo es elegir la o las mejores para ser utilizadas como parte del *framework*

de recuperación de estructuras causales a partir de textos. Esto es, poder utilizar la o las mejores técnicas sobre los conjuntos de datos originadas a partir del *New York Times* que son descritos en 4.3.4.




Escenario 3 - RFDAG

Escenario 3 - SFDAG

Escenario 4 - RFDAG

Escenario 4 - SFDAG

Figura 4.11: Desempeño, medido en términos de F1-score, de las nueve técnicas del estado del arte más el modelo de referencia (*baseline*) *Random* sobre 24 conjuntos de datos sintéticos generados con la herramienta de simulación de *TETRAD*. Estos conjuntos de datos se dividen en cuatro: (i) escenario 3 usando el enfoque *RFDAG*, (ii) escenario 3 usando el enfoque *SFDAG*, (iii) escenario 4 usando el enfoque *RFDAG* y (iv) escenario 4 usando el enfoque *SFDAG*. Estas cuatro categorías (i, ii, iii, iv) tienen 7, 7, 5 y 5 conjuntos de datos cada una, respectivamente. El escenario 3 mantiene la misma configuración, pero modificando la cantidad de variables ocultas ($H$). El escenario 4 mantiene la misma configuración, pero variando la cantidad de variables rezagadas del modelo causal real ($L$). Para el escenario 3 se puede ver que las mejores cuatro técnicas siguen siendo las mismas: *Direct-LiNGAM* (DLiNGAM), *PCMCI*, *PC* y *VAR*. Por otra parte, para el mismo escenario se observa que las cuatro peores técnicas son *Lasso-Granger* (LGranger), *SIMoNe*, *Transfer Entropy* (TE) e *ICA-LiNGAM*. Se puede ver que la técnica *BigVAR* tiene un desempeño poco consistente, especialmente para *SFDAG*. Para el escenario 4 *RFDAG* y *SFDAG* las cuatro mejores técnicas comienzan con mejor desempeño, pero la diferencia disminuye al aumentar el $L$ o el $H$. Para el escenario 4 con *SFDAG* la diferencia entre las mejores y peores técnicas es mucho menor. Se observa que, en general, *SFDAG* es una tarea más difícil que *RFDAG*, donde hay menos diferencia entre las mejores y peores técnicas. También se puede observar que los parámetros analizados ($H$ y $L$) también tienen un impacto en el desempeño (a mayor valor menor desempeño).

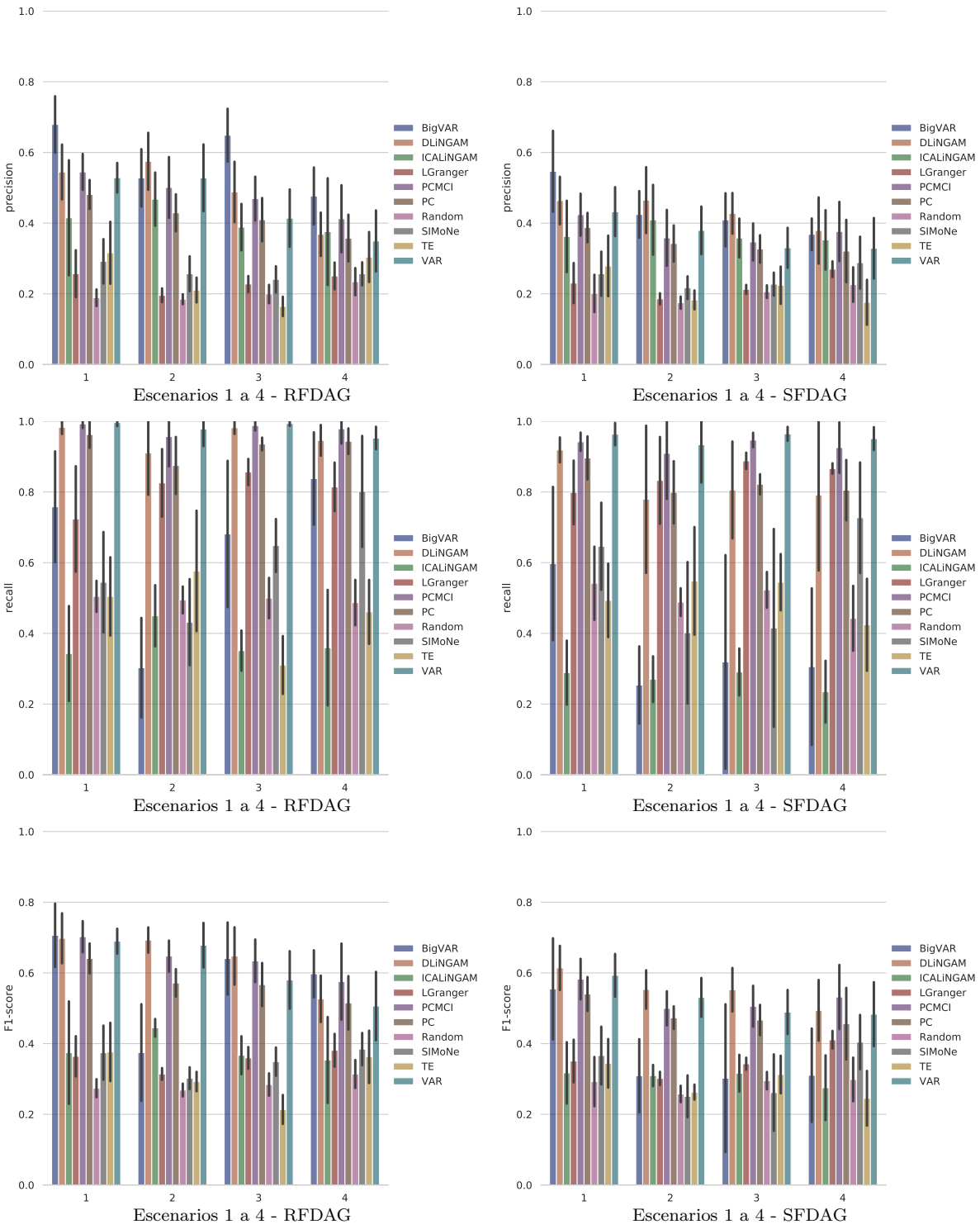



Figura 4.12: Desempeño en términos de precisión (arriba), cobertura (medio) y F1-score (abajo) de las nueve técnicas del estado del arte analizadas y el modelo de referencia (*baseline*) *Random* sobre los cuatro escenarios generados con *TETRAD*. Los conjuntos de datos se separan en 28 conjuntos de datos generados usando *RFDAG* (columna izquierda) y 28 conjuntos de datos usando *SFDAG* (columna derecha). En el eje horizontal de cada gráfico de barras se separan los resultados para cada uno de los cuatro escenarios. Se reporta para cada métrica de cada escenario el promedio de la métrica para ese escenario y el intervalo de confianza para el mismo (con un nivel de confianza de 95 %).

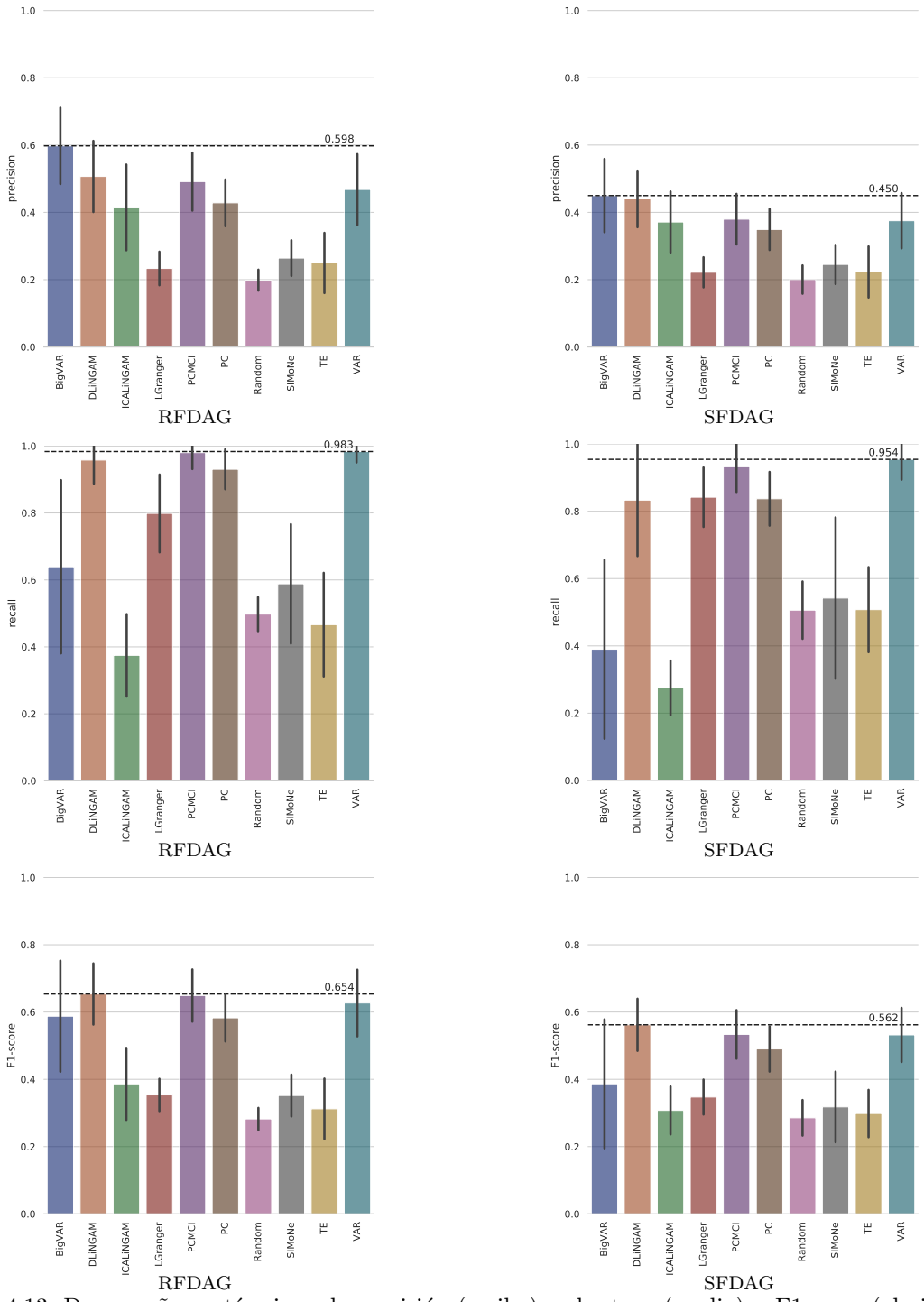



Figura 4.13: Desempeño en términos de precisión (arriba), cobertura (medio) y F1-score (abajo) de las nueve técnicas del estado del arte analizadas y el modelo de referencia (*baseline*) *Random* sobre todos los conjuntos de datos generados con *TETRAD*. Para computar cada una de las tres métricas se promedian los desempeños obtenidos para cada técnica promediando para todos los escenarios (1 al 4) para *RFDAG* (izquierda) y lo mismo para *SFDAG* (derecha). Entonces para cada métrica usando *RFDAG* se calcula un promedio de 28 conjuntos de datos, y otros 28 para *SFDAG*. Se reportan junto con los promedios, los intervalos de confianza con nivel de confianza de 95 %.

### 4.4.2. Aplicación a *CauseMe*

En la presente sección se reportan los resultados de aplicar cuatro técnicas de recuperación causal sobre ocho conjuntos de datos recuperados de la plataforma de *benchmarking* *CauseMe*. Se seleccionan cuatro del total de nueve técnicas del estado del arte basándose en los experimentos realizados sobre los conjuntos de datos generados con *TETRAD*, los métodos seleccionados son: *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*. Adicionalmente, se reporta la técnica *Random* (construida de la misma forma que para *TETRAD*) a modo de modelo de referencia (*baseline*). Estas cinco técnicas son aplicadas sobre los ocho conjuntos de datos correspondientes al conjunto *nonlinear-VAR* que se encuentra disponible en *CauseMe*. Estos ocho experimentos están construidos de manera similar, pero variando el tamaño de la longitud de la serie ($T \in \{300, 600\}$) y la cantidad de nodos ($N \in \{3, 5, 10, 20\}$). Una descripción completa de los ocho conjuntos de datos se presenta en la Sección 4.3.2. La *ground truth* de los experimentos no se encuentra disponible para descargar de la plataforma, en su lugar es necesario subir el resultado de las técnicas a la plataforma y la misma retorna las métricas de desempeño (tasa de falsos positivos, tasa de verdaderos positivos y F-measure con $\beta = 0,5$). Utilizando las métricas provistas se computa y reportan las mismas métricas que para los experimentos en *TETRAD*: precisión, cobertura y F1-score.

En la Figura 4.14 se reportan los valores de precisión, cobertura y F1-score para los ocho escenarios. La columna de la izquierda contiene los resultados de las métricas para los cuatro conjuntos de datos con longitud de serie trescientos ($T = 300$), mientras que la columna de la derecha contiene los resultados de las métricas para los cuatro conjuntos de datos con longitud seiscientos ($T = 600$). Para ambas columnas primero se muestra la precisión, en la siguiente fila se muestra la cobertura y en la última el F1-score. Para cada subfigura se muestra el resultado obtenido por cada una de las cinco técnicas para cada uno de los cuatro valores de $N$.

**Discusión de los resultados sobre los conjuntos de datos obtenidos de *CauseMe*.** Se puede observar, tanto para $T = 300$ como para $T = 600$, una caída en la precisión por parte de todas las técnicas al aumentar el $N$ (incluso por parte de *Random*), aunque la pendiente de esta caída varía de técnica a técnica. *Direct-LiNGAM* es la única técnica que al crecer el $N$ mantiene un nivel de precisión con menos caída. Esto habla positivamente de la estabilidad de los arcos encontrados por esta herramienta, sugiriendo que es una buena técnica si se tiene por objetivo solo encontrar arcos correctos

(maximizar precisión). Se puede observar que la técnica *PC* es la que tiene peor precisión seguida por *PCMCI*. Como esta última técnica construye los arcos a partir de *PC* es esperable que tenga mejor precisión (toma los arcos de *PC* como el conjunto de padres potenciales y los refina con la técnica *MCI*). Aunque no hay mucha diferencia en términos de precisión para $T = 300$ comparado con $T = 600$, algunas técnicas parecen tener una leve mejoría al aumentar el $T$, aunque no hay suficientes datos como para saber si es una diferencia significativa. Se puede observar que tanto para $T = 300$ como para $T = 600$ las técnicas se ubican igual en términos de precisión: primero *Direct-LiNGAM*, luego *VAR*, luego *PCMCI* y por último *PC*.

**En términos de cobertura**, se observa que la técnica *Random* obtiene los resultados más altos junto con *PC*. Por la forma en la que está construida, la técnica *Random* va a tener en promedio la mitad de arcos del grafo totalmente conexo (ya que por cada posible arco del grafo completo aleatoriamente elige si considerarlo causal o no usando una proporción 50-50). Por este motivo, si el grafo causal real consiste de unos pocos arcos y la técnica *Random* toma arcos como correctos con una proporción alta, se va a maximizar la cobertura y minimizar la precisión. Se puede observar que tanto para $T = 300$ como para $T = 600$ las técnicas se ubican igual en términos de cobertura: primero *PC*, luego *VAR*, luego *Direct-LiNGAM* y por último *PCMCI*. A diferencia de la precisión, la cobertura no se ve afectada por el aumento en el $N$.

**En términos de F1-score**, se puede ver que tanto para $T = 300$ como para $T = 600$ las técnicas comienzan ordenadas de mejor a peor de la siguiente manera: *PC*, *VAR*, *Random*, *Direct-LiNGAM* y *PCMCI*. Al aumentar el $N$ el desempeño de las técnicas es afectado de formas distintas, siendo *Random* la más afectada en términos de desempeño y *Direct-LiNGAM* la menos afectada. Para $N = 20$ se puede ver que *Direct-LiNGAM* pasó a ser la mejor, seguida por *VAR*, *PC*, *PCMCI* y en último lugar *Random*.

Vale la pena mencionar que los desempeños absolutos en términos de F1-score son diferentes a los obtenidos en *TETRAD* porque las métricas de desempeño son computadas de maneras distintas. Para el caso de *CauseMe* se computan las métricas de desempeño usando todos los arcos de la *ground truth*, esto incluye arcos contemporáneos (si es que hay) y arcos con distancias en el tiempo mayores a uno (si es que hay). Estos dos tipos de arcos no son los objetivos del presente estudio y no son considerados y, por ende, para computar las métricas de los experimentos de *TETRAD* no son considerados. Como la plataforma *CauseMe* no proporciona la *ground truth* sino que proporciona las métricas

obtenidas por los métodos, no es posible computar las métricas como están computadas en *TETRAD*. Por este motivo se reportan las métricas obtenidas a partir de la plataforma (que tienen en consideración todos los arcos). Por este motivo, por ejemplo, una cobertura perfecta $(1, 0)$ no es posible, ya que las técnicas propuestas solo se concentran en una porción de los arcos que existen en la estructura real (los arcos no contemporáneos con distancia uno).

Del presente análisis se puede observar que la tarea de recuperación causal tiene muchas dimensiones a tener en cuenta, y que dependiendo del dominio trabajado y el objetivo que se tenga, el análisis y las conclusiones pueden variar. Por ejemplo, que la técnica *Random* haya tenido mejor F1-score que *Direct-LiNGAM* y *PCMCI* para $N = 3$ da indicio que hay que ser cuidadoso al momento de elegir las métricas y la forma en la que se estudian las técnicas. *Random* obtuvo mejor F1-score al maximizar cobertura (agregando arcos con proporción 50-50) y aunque tuvo la peor precisión que todas las demás técnicas, obtuvo valores de F1-score mejores que técnicas que hacían análisis de causalidad reales (no que elegían arcos aleatoriamente). Entonces, dependiendo del dominio, la métrica F1-score podría no ser la mejor debiendo optar por la técnica F-measure con un $\beta$ que priorice la precisión. Más aún, en lugar de utilizar métricas de recuperación de información como falsos positivos, falsos negativos, precisión u otros, es importante conocer y plantear el uso (en caso de ser necesario) de otras métricas como distancia estructural de Hamming [AdC03] o distancia estructural de intervenciones [PB15].

Estos resultados muestran que no existe una única técnica mejor que todas, ya que dependiendo el dominio y las características de los datos una técnica que era la mejor puede dejar de serlo. Por ejemplo *PCMCI* tuvo el segundo mejor F1-score promedio en los conjuntos de *TETRAD* mientras que en *CauseMe* tuvo consistentemente peores resultados que las otras tres técnicas para casi todos los escenarios con todas las configuraciones (excepto para $N = 3$ donde fue mejor que *Direct-LiNGAM* por un pequeño margen). Análogamente, *Direct-LiNGAM* fue la peor técnica en términos de F1-score para $T = 300$ y $T = 600$ con $N = 3$. Sin embargo, para $N = 20$ pasó a ser la mejor técnica de las cuatro. Por este motivo, para seguir profundizando en el análisis de las técnicas, se incluye un tercer dominio que es el conjunto de datos de demanda de energía eléctrica provisto por *CAMMESA*. Sobre este conjunto de datos se estudian nuevamente las misma cuatro mejores técnicas: *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*.

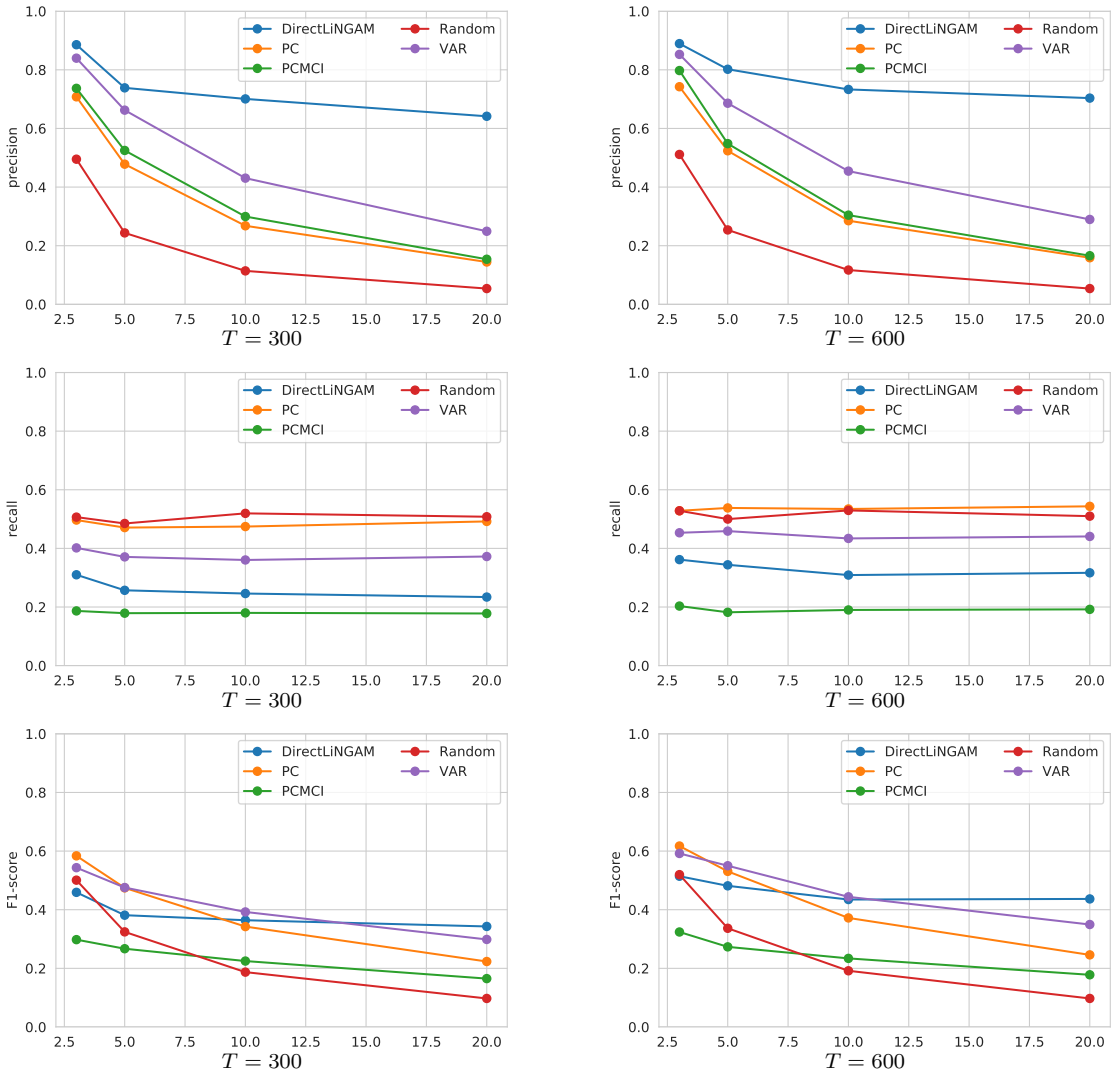



Figura 4.14: Resultado de aplicar técnicas seleccionadas del estado del arte (*Direct-LiNGAM*, *PCMCI*, *PC*, y *VAR*), y el modelo de referencia (*baseline*) *Random*, sobre los 8 conjuntos de datos obtenidos de la plataforma de *CauseMe*. Estos conjuntos de datos se corresponden con ocho experimentos del conjunto de datos *nonlinear-VAR* detallado en la plataforma antes mencionada. Estos experimentos varían la longitud de la serie (T), con $T = 300$ en la primera columna y $T = 600$ en la segunda. Dentro de estas dos longitudes de series se varía la cantidad de nodos ($N \in \{3, 5, 10, 20\}$) (eje horizontal). Para todas estas técnicas y conjuntos de datos se reporta, de arriba para abajo, la precisión, la cobertura y la media armónica de estos dos valores (F1-score) con respecto a los arcos reales. Se puede observar que para $N = 20$ y para ambos valores de $T$, las técnicas se ordenan en desempeño en términos de F1-score de la siguiente manera (de mejor a peor): *Direct-LiNGAM*, *VAR*, *PC* y *PCMCI*. Para $N = 3$, donde hay menos oportunidad para demostrar la capacidad de cada técnica (solo hay que decidir sobre unos pocos arcos) y donde cada error cuenta más (es una proporción mayor del total de arcos), las técnicas *PCMCI* y *Direct-LiNGAM* no demuestran ser mejores que *Random*. Sin embargo, *Direct-LiNGAM* muestra una mayor estabilidad en desempeño al aumentar el $N$.

## 4.5. Aplicación a Datos de Demanda Eléctrica

En la presente sección se reportan los resultados de aplicar las cuatro mejores técnicas de recuperación causal sobre el conjunto de datos de demanda de energía eléctrica provista por la compañía administradora del mercado mayorista eléctrico de argentina (*CAMMESA*). Una descripción detallada del conjunto de datos se puede encontrar en la Sección 4.3.3. Al igual que para *CauseMe*, se seleccionan cuatro del total de nueve técnicas del estado del arte basándose en los experimentos realizados sobre los conjuntos de datos de *TETRAD*. Los métodos seleccionados son: *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*.

Los resultados de aplicar estas cuatro técnicas al conjunto de datos de *CAMMESA* se pueden ver en la primera fila de la Figura 4.15. De izquierda a derecha se ven los grafos resultantes de aplicar *Direct-LiNGAM*, *VAR*, *PCMCI* y *PC* sobre dicho conjunto de datos. Se representan en rojo los arcos incorrectos y en negro los correctos de acuerdo a la *ground truth* discutida en la Sección 4.3.3. Como se puede observar las técnicas tuvieron un mal desempeño en términos de precisión. De los 7, 9, 8, 8 arcos recuperados por las técnicas *Direct-LiNGAM*, *VAR*, *PCMCI*, *PC*, respectivamente, se recuperaron 3, 5, 4 y 4 arcos incorrectos (respectivamente). Esto da como resultado los siguientes valores de precisión para las técnicas *Direct-LiNGAM*, *VAR*, *PCMCI* y *PC* respectivamente: $0,571$; $0,444$; $0,500$ y $0,500$. En términos de cobertura se puede ver que las cuatro técnicas recuperaron los mismos cuatro arcos correctos (del total de cinco), llegando a una cobertura de $0,800$. Se puede notar que en todos los casos fallaron en encontrar el arco Hum $\rightarrow$ Temp.

Los bajos niveles de precisión obtenidos por las cuatro técnicas en este conjunto de datos pueden ser explicados por la naturaleza cíclica de los datos. Al tratarse de datos de demanda eléctrica y variables climatológicas de varios años con frecuencia diaria se presentan varios ciclos en los datos. Primero, la temperatura presenta un ciclo anual que se corresponde con las estaciones del año (presentando, cada año, temperaturas más altas en verano y más bajas en invierno). Segundo, la serie de la demanda de energía eléctrica (que se vincula fuertemente con la temperatura) presenta un ciclo similar anual en respuesta al ciclo de la temperatura (temperaturas muy altas o muy bajas incentivan el consumo, generando un pico en invierno y otro en verano todos los años). Tercero, la demanda tiene una componente cíclica semanal. Esto es, cada siete días se puede esperar que se repita un patrón (la demanda de un lunes es similar a la del lunes anterior y la demanda de un domingo es similar a la del domingo anterior). Se tiene la hipótesis de que, debido a la presencia de esos ciclos (que agregan correlaciones entre diferentes variables

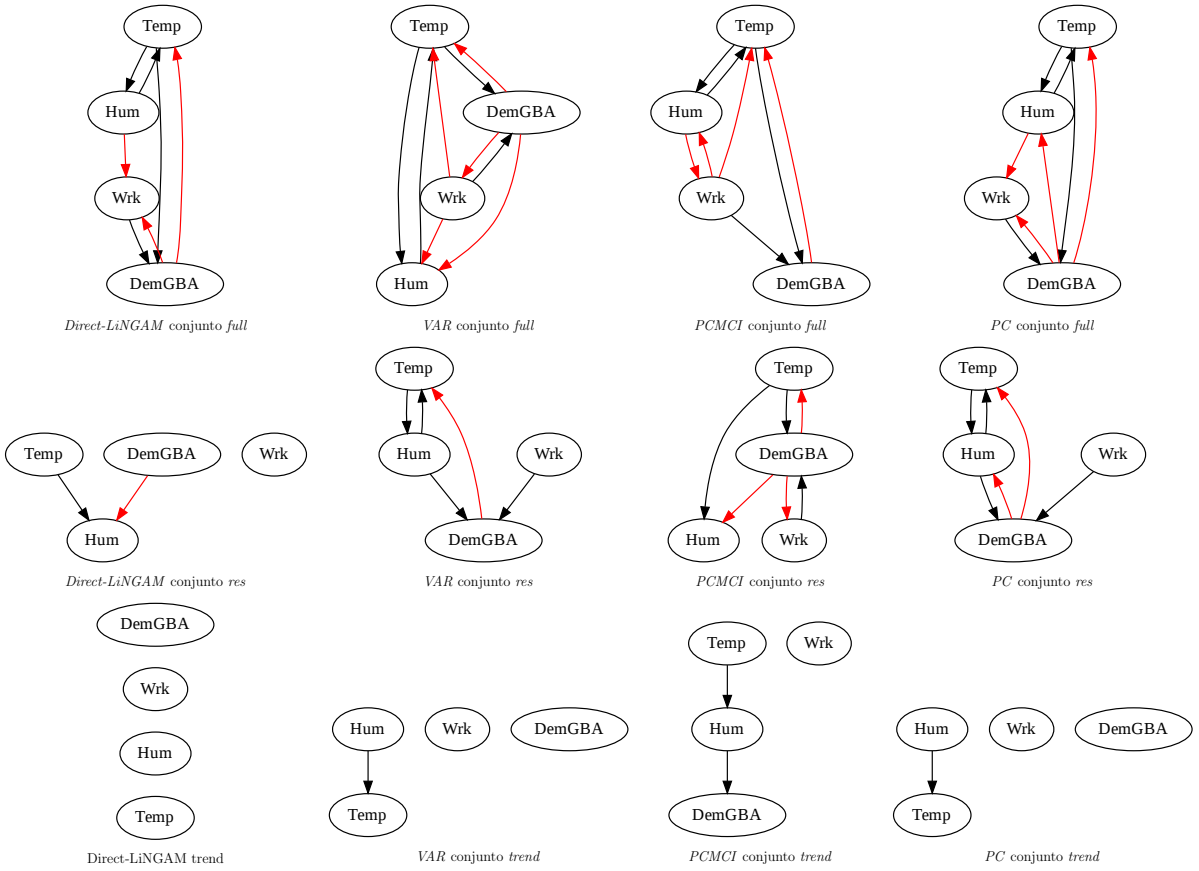



Figura 4.15: Grafos causales resultantes de aplicar las técnicas *Direct-LiNGAM*, *VAR*, *PCMCI* y *PC* sobre el conjunto de datos provisto por *CAMMESA* (*full*) (arriba) y dos transformaciones del mismo: *res* (medio) y *trend* (abajo). La primera transformación consiste de reemplazar las variables Temp, Hum y DemGBA por los residuos de modelar esas mismas variables como una regresión lineal de 20 variables que dan información de la estación, mes del año y día de la semana. Al modelar las variables como lo no explicado por esas 20 variables (los residuos de la regresión), se eliminan componentes cíclicos. El conjunto *trend*, consiste de las mismas variables, pero luego de aplicarles un filtro de descomposición estacional mediante promedios móviles para solo mantener la componente de la tendencia. Estas dos transformaciones de los datos son aplicadas a las variables reales, no a la variable binaria Wrk. Se puede ver un gran nivel de precisión para el conjunto *trend*, pero comprometiendo el desempeño en términos de cobertura. Se observa un grafo poco informativo para *full*, dónde se encuentra casi el grafo totalmente conectado, obteniendo alta cobertura, pero mala precisión. Por otro lado, el conjunto *res* resulta un buen intermedio a los escenarios obtenidos por los otros dos conjuntos de datos.

en diferentes instantes de tiempo) la capacidad de extracción causal se puede ver afectada debido a que se vuelve posible predecir valores futuros de variables a partir de valores pasados de variables no causales. Esta capacidad adicional de predicción puede hacer que las técnicas de extracción de causalidad detecten vínculos causales entre variables que no

los tienen, generando así una pérdida en precisión.

Para contrarrestar este fenómeno se aplicaron dos técnicas distintas para la eliminación de los componentes cíclicos en los datos. Primero se aplica el filtro de descomposición estacional del paquete *statsmodels*[8] versión 0.10.2 usando frecuencia siete, para eliminar el ciclo semanal. De esta descomposición solo se tomó la componente de la pendiente (descartando el ciclo). A este conjunto de datos se lo denomina *trend*. Como segunda técnica de filtrado de ciclo se construyeron veinte variables auxiliares binarias para capturar las componentes cíclicas del conjunto de datos. Estas variables son: *primavera*, *verano*, *otoño*, *lunes*, *martes*, *miércoles*, *jueves*, *viernes*, *sábado*, *febrero*, *marzo*, *abril*, *mayo*, *junio*, *julio*, *agosto*, *septiembre*, *octubre*, *noviembre* y *diciembre*. Se realiza la regresión de cada una de las variables continuas (Hum, Temp y DemGBA) con respecto a las variables auxiliares.

$$\begin{aligned} \text{Hum} = &\quad \beta_0 + \beta_1 \text{primavera} + \beta_2 \text{verano} + \cdots + beta_{20} \text{diciembre} + \varepsilon_{\text{Hum}} \\ \text{Temp} = &\quad \beta_0 + \beta_1 \text{primavera} + \beta_2 \text{verano} + \cdots + beta_{20} \text{diciembre} + \varepsilon_{\text{Temp}} \\ \text{DemGBA} = &\quad \beta_0 + \beta_1 \text{primavera} + \beta_2 \text{verano} + \cdots + beta_{20} \text{diciembre} + \varepsilon_{\text{DemGBA}} \end{aligned} \tag{4.10}$$

El objetivo es modelar las componentes cíclicas de cada variable en término de estas variables auxiliares. Finalmente, cada una de las tres variables es reemplazada por los residuos de la regresión lineal de sí misma contra las variables auxiliares. Ya que los residuos resultantes capturan todo lo no explicado por las variables que representan el momento del año, se espera que los residuos contengan la información de la tendencia de la serie (y no de sus componentes cíclicas). Utilizando esta técnica se construye el conjunto de datos filtrado *res*.

En la segunda y tercera fila de la Figura 4.15 se pueden ver los resultados de aplicar las cuatro técnicas a los conjuntos de datos *res* y *trend*, respectivamente. Se puede observar para ambos conjuntos una gran caída en cantidad de arcos incorrectos recuperados, en muchos casos aumentando notablemente la precisión. Por ejemplo, para la última fila, correspondiente al conjunto *trend*, se puede apreciar que al no encontrar ningún arco incorrecto se tiene una precisión perfecta (todos los arcos marcados como relevantes son correctos, es decir, no hay falsos positivos). Por otra parte, para *res* se encuentran algunos arcos incorrectos, pero aun así se tiene un incremento de la precisión para las técnicas *VAR* y *PC*, que pasaron de $0,444$ y $0,500$ de precisión a tener $0,800$ y $0,667$, respectivamente. Respecto a la cobertura, se puede observar que para el conjunto de datos *trend* todas

[8] https://www.statsmodels.org/stable/generated/statsmodels.tsa.seasonal.seasonal_decompose.html

las técnicas empeoraron los valores para esa métrica. Por otra parte, para el conjunto *res* se mantuvo constante para *PC* y *VAR*, pero para las otras dos técnicas la cobertura empeoró.

Un resumen de los valores de precisión, cobertura y F1-score se pueden ver en la Tabla 4.8. Como se puede apreciar, considerando todos los conjuntos de datos, la técnica con mejor precisión fue *VAR*. En términos de cobertura, teniendo en cuenta todos los conjuntos de datos, *PCMCI*, *PC* y *VAR* obtuvieron el mismo valor de cobertura promedio. Por último, considerando todos los conjuntos de datos, la técnica *PCMCI* tuvo el mejor F1-score promedio.

En estos experimentos se puede ver, nuevamente, que de acuerdo a las características de los datos y del dominio, y de los objetivos que se persiguen, la metodología, las métricas y las técnicas deben ser ajustadas adecuadamente. No hay una sola técnica ni métrica mejor para todos los dominios. Por ejemplo, la técnica *Direct-LiNGAM* obtuvo el mejor F1-score para el conjunto de datos *full* (0,667) pero el peor para los datos *res* (0,286). La elección de las técnicas y la metodología va a depender de las características de los datos y de si el objetivo es obtener un conjunto pequeño de arcos correctos o si el objetivo es obtener buena cobertura para un posterior filtrado de falsos positivos. Estos resultados también muestran que, dependiendo de la naturaleza de los datos y los objetivos, un paso de preprocesamiento de los datos puede o no ser necesario.

Por último, como los conjuntos de datos obtenidos de *CAMMESA* eran pocos (uno más dos transformaciones de datos) y con pocas variables (solo cuatro variables consideradas) se pudo visualizar los grafos causales resultantes (algo que no fue posible para los datos sintéticos por la gran cantidad de conjuntos de datos y de variables). Del análisis de los grafos surge nuevamente la importancia de las métricas: mientras el conjunto *full* tuvo el mejor desempeño en términos de F1-score promedio, resulta evidente que no se trata de un conjunto informativo para ese conjunto de datos (ya que los grafos son casi totalmente conectados, y muchos arcos son incorrectos). Por otro lado, al aplicar las técnicas al conjunto *res*, se obtuvo peor desempeño en términos de F1-score promedio, pero los grafos causales resultantes son más informativos.

Guiado por estas conclusiones, no se elige una única técnica para ser usada como parte del *framework* de extracción de relaciones causales a partir del texto, sino que las cuatro técnicas son utilizadas en la Sección 4.6 utilizando una estrategia de votación por consenso (*ensemble* de técnicas) para maximizar la precisión de los arcos extraídos.

| | Precisión | | | | Cobertura | | | | F1-score | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ***full*** | ***trend*** | ***res*** | **promedio** | ***full*** | ***trend*** | ***res*** | **promedio** | ***full*** | ***trend*** | ***res*** | **promedio** |
| ***PCMCI*** | 0,500 | 1,000 | 0,500 | 0,667 | 0,800 | 0,400 | 0,600 | **0,600** | 0,615 | 0,571 | 0,545 | **0,577** |
| ***PC*** | 0,500 | 1,000 | 0,667 | 0,722 | 0,800 | 0,200 | 0,800 | **0,600** | 0,615 | 0,333 | 0,727 | 0,559 |
| ***DLiNGAM*** | 0,571 | 0,000[9] | 0,500 | 0,357 | 0,800 | 0,000 | 0,200 | 0,333 | 0,667 | 0,000 | 0,286 | 0,317 |
| ***VAR*** | 0,444 | 1,000 | 0,800 | **0,748** | 0,800 | 0,200 | 0,800 | **0,600** | 0,571 | 0,333 | 0,800 | 0,568 |
| **Promedio** | 0,504 | **0,750** | 0,617 | 0,624 | **0,800** | 0,200 | 0,600 | 0,533 | **0,617** | 0,310 | 0,590 | 0,505 |

Tabla 4.8: Desempeños en términos de precisión, cobertura y F1-score de aplicar las técnicas *Direct-LiNGAM* (*DLiNGAM*), *VAR*, *PCMCI* y *PC* sobre el conjunto de datos provisto por *CAMMESA* (*full*) y dos transformaciones del mismo: *res* y *trend*. La primera transformación consiste de reemplazar las variables Temp, Hum y DemGBA por los residuos de modelar esas mismas variables como una regresión lineal de 20 variables que dan información de la estación, mes del año y día de la semana. Al modelar las variables como lo no explicado por esas 20 variables (los residuos de la regresión), se eliminan componentes cíclicos. El conjunto *trend*, consiste de las mismas variables, pero luego de aplicarles un filtro de descomposición estacional mediante promedios móviles para solo mantener la componente de la tendencia. Estas dos transformaciones de los datos son aplicadas a las variables reales, no a la variable binaria Wrk. Se puede observar la alta cobertura obtenida con el conjunto *full* a costa de una mala precisión. Se observa el escenario inverso para *trend* (alta precisión, baja cobertura). Por otra parte, *res* representa un escenario intermedio entre los otros dos. Se puede observar que, si bien el conjunto *full* parece el mejor en términos de F1-score, no necesariamente indica que este sea el mejor conjunto de datos. Como se ve en la Figura 4.15, el grafo obtenido con *full* no es informativo ya que es casi el grafo totalmente conectado. Debido a la gran cantidad de arcos en la *ground truth* de *CAMMESA*, una representación de los datos que maximiza la cobertura se ve beneficiada, esto no implica que esa representación sea la más indicada para esos datos.

## 4.6. Aplicación a Textos de Artículos Periodísticos

En la presente Sección se reportan los resultados de aplicar las cuatro mejores técnicas de recuperación causal sobre los tres conjuntos de datos de textos extraídos del *New York Times*: (1) el conjunto de datos de términos extraídos usando la técnica de pesaje de términos $\text{FDD}_\beta$, (2) el conjunto de datos de eventos en curso detectados de los textos y (3) el conjunto de dato que combina los dos anteriores. Una descripción detallada de los conjuntos de datos se puede encontrar en la Sección 4.3.4. Al igual que para los experimentos sobre *CauseMe* y *CAMMESA*, se seleccionan cuatro del total de nueve

[9]La técnica *Direct-LiNGAM* (*DLiNGAM*) para el conjunto de datos obtenido de *CAMMESA* con la transformación *trend* encontró cero arcos causales, por ende el valor de la precisión queda indefinido (0/0). Para evitar usar este valor indefinido, a esa técnica con ese conjunto, se le asigna el valor de precisión cero.

técnicas del estado del arte basándose en los experimentos previos realizados. Los métodos seleccionados son: *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*. Estas cuatro técnicas son usadas en conjunto para crear una única técnica que consiste de la aplicación de las cuatro anteriores con una política de votación unánime. Esto es, la técnica solo considera un arco como correcto si ese arco fue detectado simultáneamente por las cuatro técnicas utilizadas. A esta técnica, en este trabajo, se la denomina *ensemble*.

Los resultados de la aplicación de la técnica *ensemble* al conjunto de datos de términos (1) y al conjunto de datos de eventos en curso (2) se pueden ver en la Figura 4.16 (arriba a la izquierda y arriba a la derecha, respectivamente). En la misma figura (abajo) se reporta el grafo resultante de aplicar la técnica *ensemble* al conjunto de datos que combina eventos en curso y términos (3).

**De analizar el grafo resultante de aplicar el *ensemble* al conjunto de datos de términos (Figura 4.16, arriba a la izquierda), se desprenden las siguientes conclusiones**. Algunas relaciones causales resultan de interpretación directa y se pueden presuponer correctas. Por ejemplo, las menciones de acciones militares sobre Irak que hayan precedido a la guerra en dicho país. Otros arcos no tienen una interpretación tan directa como el caso de United Nations Security (Council) y su relación causal con la guerra en Irak, con Saddam Hussein o con la guerra del golfo pérsico. Sin embargo, teniendo en cuenta que en este análisis causal la causa precede al efecto, el arco ('United', 'Nations', 'Security') → ('war', 'Iraq') puede estar capturando los esfuerzos por parte de esa entidad para que "cumplan con sus obligaciones de desarme" a través de resoluciones como la 1441, esfuerzos que precedieron a la guerra. Otros vínculos como ('weapons', 'mass', 'destruction') → ('chemical', 'biological', 'weapons') son más difíciles de interpretar, pero es posible que se deban a que cada vez que surgía el tema de posibles programas de armas de destrucción masiva que funcionaban en Irak en ese momento, se hablaba después de posibles programas de armas biológicas. De igual manera, el vínculo ('United', 'Nations', 'Security') → ('Persian', 'Gulf', 'war'), no puede interpretarse de manera directa (ya que las causas de una guerra no se pueden resumir a una sola interacción con una sola entidad o institución), pero se puede considerar que se refieren a la resolución 678 del *United Nations Security Council*, en la cual se le da una fecha límite a Irak para retirarse de Kuwait, evento que precedió a la Guerra del Golfo.

**De analizar el grafo resultante de aplicar el *ensemble* al conjunto de datos de eventos en curso (Figura 4.16, arriba a la derecha), se desprenden las**

**siguientes conclusiones**. Las variables `C550` y `C165` están vinculadas con una relación causal en ambas direcciones (`C550` $\leftrightarrow$ `C165`). Ambas variables están compuestas mayoritariamente por menciones de la guerra de Irak, pero la primera tiene una fuerte componente de menciones de sentimientos en contra de la guerra (against) de acuerdo a las nubes de palabras de la Figura 4.9 (para ver la descripción completa de los grupos ver Tabla 4.5). Que eventos relacionados a la guerra reportados en los artículos periodísticos causen sentimientos en contra de la guerra es un vínculo causal esperable. El sentido contrario es más difícil de analizar, pero puede tener que ver con la dinámica de la forma en la que se reportaban noticias de la guerra. Por ejemplo, es posible que cuando la conversación sobre la guerra (y el estar en contra o no) está más presente en lo cotidiano, se vuelve más probable que los medios reporten noticias relacionadas a la guerra. Aunque esta es una posible teoría, también puede ser que esto se deba simplemente a lo similares que son los eventos y que suelen ser mencionados en conjunto, lo cual indicaría que es una limitación de las técnicas al confundir co-ocurrencia con causalidad. Los vínculos causales `C269` $\rightarrow$ `C109` $\leftarrow$ `C201` son de interpretación más sencilla. Las variables representan la invasión a Kuwait por parte de Irak, reportes de muertes de soldados y civiles por parte de diferentes países y reportes de ataques terroristas, respectivamente. El hecho de que la invasión a Kuwait y los ataques terroristas causen posteriores reportes de muertes de civiles y soldados es una relación causal fácil de interpretar y que se puede suponer correcta.

Por último, **de analizar el grafo resultante de aplicar el *ensemble* a la combinación de los conjuntos de datos de eventos en curso y de términos (Figura 4.16, abajo), se desprenden las siguientes conclusiones**. Primero, se puede ver cómo variables de distintos tipos (eventos y términos) se vinculan causalmente en el grafo resultante de aplicar la técnica *ensemble*, aun siendo que estas variables se construyeron de maneras distintas y presentando diferentes características (diferentes frecuencias medias y desvíos estándares, ver Tablas 4.7 y 4.6). Esto daría evidencia para afirmar que el *framework* de recuperación de estructuras causales puede manejar diferentes tipos de variables extraídas del texto e incluso variables exógenas al texto que se pueden incluir si están en el mismo periodo de tiempo y frecuencia (como, por ejemplo, agregar precios de acciones de la bolsa, o precios de materias primas o indicadores socioeconómicos de países, entre otros).

Respecto a las relaciones causales extraídas, se puede observar vínculos causales con una semántica bien definida. Por ejemplo, el vínculo entre la posible existencia de armas

de destrucción masiva en Irak como uno de los posibles justificativos para iniciar acciones militares ('weapons', 'mass', 'destruction') → ('military', 'action','Iraq') dando inicio a la guerra en Irak ('military', 'action','Iraq') → ('war', 'Iraq'). Este tipo de relaciones, si bien no deberían ser interpretadas automáticamente como una relación causal real, brinda información sobre la forma en la que son reportados ciertos eventos en los textos de artículos de noticias utilizados. Su extracción puede ofrecer información sobre pares de eventos con fuerte co-ocurrencia donde uno precede al otro, y permitiría un análisis sobre el contexto a la luz de esta información. Por otro lado, siendo que la variable C109 representa reportes de muertes de civiles y soldados, el vínculo causal ('war', 'Iraq') → C109, es una relación causal que se puede presuponer correcta, ya que es razonable considerar la existencia de la guerra como la causa de los reportes de bajas de civiles y soldados. Los vínculos causales 201 → C109 ← 269, que relacionan causalmente los ataques terroristas y la invasión a Kuwait con reportes de civiles y soldados muertos, representan un vínculo que también se puede presuponer correcto. Este vínculo causal ya había sido encontrado por el *ensemble* al ser aplicado en el conjunto de datos de eventos. Análogamente, el vínculo ('weapons', 'mass', 'destruction') → ('chemical', 'biological', 'weapons'), ya había sido encontrado y discutido durante los experimentos con el *ensemble* sobre el conjunto de datos de términos.

Como se mencionó previamente, no todos los arcos encontrados son necesariamente correctos, pero proveen a un posible experto con información sobre qué variables están fuertemente interconectadas (y posiblemente causalmente vinculadas) en un dominio. Esta herramienta puede ser de gran utilidad a un experto que está tratando de entender un dominio que puede ser muy complejo y contener una gran cantidad de variables y relaciones. Adicionalmente, al transformar los textos de lenguaje natural a información estructurada de variables relevantes (series de tiempo), se provee una gran flexibilidad para combinar información del mundo real reportada en las noticias con diferentes variables que ya tienen formato de serie de tiempo (precios de materias primas, precios de acciones de la bolsa, indicadores socioeconómicos, entre otros), lo cual permitiría enriquecer el dominio y dar información extra a los expertos.

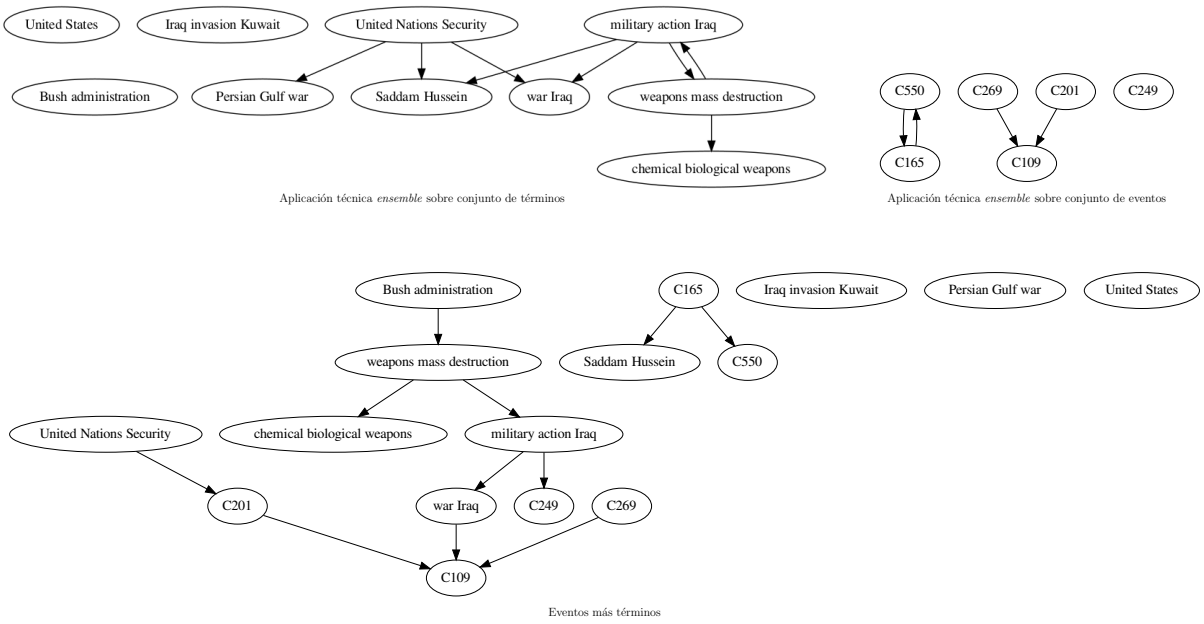



Figura 4.16: Grafos causales resultantes de aplicar la técnica *ensemble* sobre el conjunto de datos de términos (arriba a la izquierda), eventos en curso (arriba a la derecha) y el conjunto de datos construido combinando los anteriores (abajo). Ambos conjuntos de datos (el de términos y el de eventos) son extraídos del corpus del *New York Times*. La técnica *ensemble* consiste de la aplicación de las cuatro técnicas *Direct-LiNGAM*, *PCMCI*, *PC* y *VAR*, decidiendo sobre cada arco con una política de votación unánime. Los nodos del conjunto de datos de eventos están identificados con el número de grupo (cluster). Cada grupo contiene menciones (instancias) de un mismo evento (semántica similar). El grupo `C109` se corresponde con reportes de muertes durante la Guerra de Irak, tanto soldados (de ambos bandos) como civiles. El grupo `C165` se corresponde con reportes de opinión sobre la guerra de Irak. El grupo `C201` son menciones de eventos de ataques terroristas. El grupo `C249` son menciones en ataques o acciones militares. El grupo `C269` trata sobre la invasión a Kuwait. Finalmente, el grupo `550` reúne menciones de la guerra en Irak en general. Una descripción de estos grupos es presentada en la Tabla 4.5. Se puede ver la capacidad del *framework* de recuperar variables relevantes al dominio, y la capacidad del mismo de encontrar vínculos causales interpretables entre las mismas. A partir de estos vínculos se puede estudiar la precedencia y el impacto de una variable hacia otra.

## 4.7. Conclusiones

En el presente capítulo se presentó un extenso análisis de nueve técnicas de descubrimiento causal, utilizando 64 conjuntos de datos sintéticos (56 creados con *TETRAD*, 8 descargados de *CauseMe*) y dos reales (los conjuntos de datos provistos por la empresa *CAMMESA* y los creados a partir de los textos del *New York Times*). Las nueve técnicas son puestas a prueba en los 56 conjuntos de datos creados con *TETRAD*, posteriormente, se seleccionan las mejores cuatro para continuar en los demás conjuntos, siendo las

cuatro mejores técnicas: *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*. El *framework* de descubrimiento causal a partir de textos de artículos periodísticos es formalmente definido en este capítulo y un caso de uso es presentado usando los datos generados a partir del *New York Times*. Dicho *framework* consiste de una etapa de selección de variables de textos en la cual dos tipos distintos de variables son extraídos (términos y eventos en curso). Esta primera etapa de selección de variables es llevada a cabo con las herramientas propuestas en los Capítulos 2 y 3. Posteriormente una técnica de descubrimiento causal es aplicada sobre dichas variables, la técnica es una combinación de las cuatro mejores siguiendo una política de votación unánime, a la cual se la denomina *ensemble*.

**En los experimentos realizados sobre los conjuntos de datos generados con *TETRAD***, se puede observar una gran disparidad de desempeños. Se puede ver que hay tres técnicas que en muchos casos tienen un desempeño no mucho mejor, e incluso a veces peor, que *Random*. Estas tres técnicas son *Lasso-Granger*, *Transfer entropy* e *ICA-LiNGAM*, las cuales no tuvieron una diferencia significativa en términos de F1-score comparado con *Random*. Para el caso de *Transfer Entropy* se puede explicar debido a que esta técnica tiene la limitante de analizar causalidad de a pares (potencialmente dejando variables relevantes fuera del análisis, generando problemas de variables ocultas). Por otra parte, *ICA-LiNGAM* tiene limitaciones que fueron descritas por los mismos autores en publicaciones posteriores: problemas de convergencia a la solución correcta si el estado inicial no es el adecuado y sensibilidad a la escala de los datos. Para el caso de *Lasso-Granger*, si bien dicha técnica tiene un proceso previo de selección de variables usando lasso, sufre de la misma limitación que *Transfer Entropy*, esto es, solo considera causalidad de a pares (usando la técnica de causalidad de Granger de a pares de variables). Por las limitaciones de estas técnicas al ser aplicadas a estos contextos multivariados, las mismas no fueron consideradas para los experimentos posteriores en los demás conjuntos de datos (*CauseMe*, *CAMMESA*, *New York Times*).

Por otro lado, las técnicas que construyen un modelo VAR penalizado (*BigVAR* y *SIMoNe*), obtuvieron resultados poco consistentes (grandes intervalos de confianza) y en muchos casos, la técnica *SIMoNe* tuvo desempeño peor que la técnica *Random*. Más aún *SIMoNe*, en términos promedio, no tuvo un desempeño significativamente distinto a *Random*. En contraposición, la técnica *VAR* (sin penalización) obtuvo desempeños que la ubican entre las mejores cuatro técnicas. Esto demuestra que los procesos de penalización no fueron adecuados para el proceso de descubrimiento causal. Para el caso de

*BigVAR*, aumentó el desempeño en términos de precisión, pero obtuvo un mal desempeño en términos de cobertura. Dependiendo de los objetivos de la tarea, una técnica de este estilo puede ser considerada, aunque su desempeño fue en términos generales poco consistente. Por otro lado, *SIMoNe* obtuvo bajos valores de cobertura, probablemente debido a la penalización, pero también un desempeño bajo en términos de precisión. Por lo cual tanto la inferencia como la penalización de esta técnica no fueron buenas, resultando en muchos falsos negativos (tal vez por problemas de penalización) y falsos positivos (por limitaciones de la técnica para el descubrimiento de vínculos causales correctos). Ninguna de las dos técnicas basadas en modelos VAR penalizados fueron consideradas para los experimentos posteriores. Finalmente, las cuatro técnicas seleccionadas para los experimentos en los demás conjuntos de datos (*CauseMe*, *CAMMESA* y *New York Times*) son *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*.

Se puede observar que, estas cuatro mejores técnicas, se ordenan de mayor a menor en términos de F1-score de la siguiente manera: primero *Direct-LiNGAM* con el mejor F1-score, luego *PCMCI*, *VAR* y por último *PC*. Aunque las técnicas compartan este orden en términos promedio, se puede observar de los intervalos de confianza que no hay una diferencia significativa en sus desempeños para la métrica F1-score.

**En los experimentos realizados sobre los conjuntos de datos obtenidos de *CauseMe***, se puede observar nuevamente, que varias técnicas en ciertas configuraciones no pudieron superar a *Random*. Para el caso de pocas variables, donde cada arco equivocado tiene proporcionalmente un impacto mayor en las métricas, se puede ver que la técnica *Random* fue mejor en término de F1-score, que *PCMCI* (para $N = 3$ y $N = 5$) y que *Direct-LiNGAM* (para $N = 3$). Estos resultados se mantienen tanto para $T = 300$ como para $T = 600$. La técnica *Random*, logró valores altos de cobertura con respecto a las otras técnicas a través de seleccionar aleatoriamente qué arcos incluir con una proporción de aceptación alta (50-50). De esta manera, esta técnica logró superar en términos de F1-score a otras sin aportar información causal real. Esto pone en evidencia la importancia de las métricas en el descubrimiento causal. Siendo que se puede obtener cobertura perfecta con una técnica que siempre da como resultado el grafo totalmente conectado, en muchos casos hay que priorizar la precisión, u obtener un balance de ambas métricas pero dándole mayor ponderación a la precisión (por ejemplo, usando F-measure con un $\beta < 1$). También se pudo ver que técnicas que fueron las mejores en otros conjuntos de datos pasaron a tener desempeños peores que *Random* en algunas configuraciones. Por ejemplo,

*PCMCI* pasó de ser la segunda mejor técnica en *TETRAD* a ser peor que *Random* en términos de F1-score para $N = 3$ y $N = 5$, tanto con $T = 300$ como con $T = 600$.

Los experimentos en los conjuntos de datos obtenidos de *CauseMe* permiten llegar a conclusiones similares a las obtenidas de *TETRAD*. Primero, las cuatro mejores técnicas siguen teniendo desempeños similares, no existe una técnica que haya tenido un desempeño notablemente mejor o peor que otras. También se pudo ver que dependiendo del conjunto de datos puede que las técnicas tengan pequeñas variaciones en los desempeños, dando evidencia para suponer que no hay una única técnica que sea mejor para todos los conjuntos de datos y todas las configuraciones. Se puso en evidencia la importancia de usar las métricas correctas. De otro modo se pueden juzgar técnicas que no aportan información causal como *Random* (que no realiza descubrimiento causal real), como mejores que otras que sí han demostrado capacidad de descubrir relaciones causales. Teniendo en cuenta que muchas veces el desempeño de las técnicas depende del conjunto de datos y que no necesariamente hay una mejor, se continúa con las cuatro técnicas para los experimentos en los conjuntos de datos reales: el provisto por la empresa *CAMMESA* y el generado a partir de los textos del *New York Times*.

**Los experimentos realizados sobre los conjuntos de datos provistos por la empresa *CAMMESA*** tenían por objetivo agregar perspectivas diferentes al análisis. Primero, se agrega al análisis los resultados de aplicar las técnicas a un conjunto de datos real con formas funcionales, ruidos y características desconocidas. Adicionalmente, a diferencia de los conjuntos de datos sintéticos de los cuales se tenían 64 conjuntos de datos, para *CAMMESA* solo se tenía un único conjunto de datos, al cual se le aplicaron dos transformaciones resultando en tres conjuntos de datos basados en una misma fuente. Tener pocas variables y pocos conjuntos de datos permitió, por primera vez, reportar e inspeccionar en detalle los grafos resultantes (no solo concentrarse en las métricas de desempeño). Analizar el mismo conjunto de datos, pero sometido a diferentes transformaciones (filtrados de ciclos), permitió incorporar una dimensión más al análisis: la importancia del correcto preprocesamiento de los datos.

Se puede observar que, para el conjunto de datos sin transformar, se obtienen, para todas las técnicas, altos valores de cobertura, pero unos valores de precisión muy bajos, con una gran cantidad de arcos incorrectos. Un escenario de este estilo puede ser beneficioso si se busca maximizar la cobertura para un filtrado de precisión posterior (tal vez con un usuario involucrado en esta segunda etapa). Pero si se busca maximizar la precisión, los

otros dos conjuntos de datos son mejores (los dos conjuntos que se obtienen de aplicar transformaciones a los datos). En el otro extremo, la transformación que surge de aplicar un filtro de descomposición para obtener la tendencia de los datos (*trend*), obtiene precisión perfecta reportando cero arcos incorrectos encontrados, pero a costa de tener valores de cobertura muy bajos. Nuevamente se vuelve evidente la importancia de las métricas, observándose que en términos de F1-score promedio, el conjunto de datos sin transformar tiene los mejores resultados. Sin embargo, al analizar en detalle los grafos resultantes se puede observar que estos grafos no aportan mucha información causal ya que son grafos casi totalmente conectados. Estos grafos tienen poco valor para un usuario que tendría que analizar un grafo casi totalmente conectado para tratar de sacar conclusiones causales. En este caso, al guiarse por la métrica F1-score, uno puede llegar a la conclusión de que no transformar los datos es la mejor opción, pero depende mucho del objetivo que se persigue (alta cobertura o alta precisión).

Una vez más se puede apreciar que no hay una técnica que sea consistentemente superadora en términos de F1-score para todos los conjuntos de datos. En promedio la técnica con peor F1-score es *Direct-LiNGAM*, la cual tuvo el mejor desempeño en términos de F1-score tanto para los conjuntos generados con *TETRAD* como los obtenidos de *CauseMe*. Esto puede deberse a que las fuertes restricciones que impone la técnica (linealidad con ruidos no gaussianos) no se cumplen en un conjunto de datos real como el de *CAMMESA*. Al no haber una técnica superadora se plantea el uso de las cuatro técnicas para el conjunto de datos generados del *New York Times*. Como se trata de un conjunto de datos en el que la *ground truth* no es conocida, y que se tiene un gran número de variables, se prioriza la precisión por encima de la cobertura. Por este motivo, en lugar de utilizar las cuatro técnicas por separado y analizar su desempeño, se propone la creación de una única técnica que consiste en aplicar las cuatro mejores técnicas con una política de votación unánime. A esta técnica se la denomina *ensemble*.

**Para el caso de uso del *framework* realizado sobre el conjunto de datos construido a partir del *New York Times*** se puede observar que a pesar de haber aplicado la técnica *ensemble*, que maximiza la precisión, se tiene un gran número de arcos resultantes para analizar. De los resultados y el análisis presentado en la Sección 4.6 se puede ver el potencial del *framework* para capturar variables altamente relevantes a un dominio, las cuales en muchos casos capturaban eventos del mundo real reportados en las noticias que son de suma importancia para el dominio. Adicionalmente, se vio

la capacidad de la técnica de proporcionar vínculos entre estas variables que, con una correcta interpretación por parte de un experto, le permitirían a este mismo entender mejor los sucesos relacionados al dominio que acontecieron (o están aconteciendo).

Adicionalmente, se vio cómo la técnica de recuperación causal fue capaz de encontrar relaciones significativas entre variables que fueron construidas de formas distintas. Esto es, relaciones entre variables tipo evento en curso (construidas a partir de un modelo de aprendizaje automático implementado para la detección de dichas variables) y variables tipo términos (extraídas a partir de la herramienta de pesaje de términos propuesta, $\text{FDD}_\beta$). Estos dos tipos de variables fueron construidas de formas distintas y presentan escalas distintas (por ejemplo, diferente media y desvío estándar), e incluso ante variables heterogéneas como estas, se pudieron reconstruir relaciones causales relevantes. Esto da indicios para asumir que el *framework* propuesto posee una gran flexibilidad, ya que podría incorporar variables de distintos tipos (siempre y cuando todas las variables utilizadas puedan ser representadas como series de tiempo comprendidas dentro del mismo periodo de tiempo y con igual frecuencia). Esto permitiría, por ejemplo, incorporar variables financieras como el precio de acciones de empresas relevantes, o precios de materias primas importantes para el dominio, o indicadores socioeconómicos del país, o los países, relevantes al dominio.

El desarrollo completo del presente caso de uso del *framework* demostró la capacidad del mismo para obtener representaciones resumidas de contextos complejos a través de la representación de variables y relaciones relevantes entre las mismas. El *framework*, demostró un gran potencial para ser usado como parte de una herramienta de visualización interactiva, donde diferentes parámetros se pueden configurar para seleccionar más o menos variables. Por ejemplo cambiar los valores de $\beta$ de la técnica $\text{FDD}_\beta$ para seleccionar variables más descriptivas o más discriminativas. Así como también es posible configurar diferentes aspectos de la recuperación de eventos (diferentes valores de $K$, o diferentes umbrales mínimos de cohesión y cantidad de instancias por grupo) para obtener mayor o menor cantidad de variables de tipo evento sobre las cuales elegir cuáles se visualizan y cuáles no. Adicionalmente, se puede incorporar interacción para el paso de descubrimiento causal, cambiando la política de votación *ensemble*, o agregando o sacando técnicas del mismo.

Si bien la propuesta demostró ser viable, aún existen algunos puntos que se pueden mejorar para obtener representaciones más útiles, completas y fáciles de visualizar. Por

ejemplo, la representación de los grupos de menciones del mismo evento en el grafo se realizó a través de la etiqueta de identificación del grupo (`CXXX`). Para poder interpretar cada grupo se procedió a la construcción de nubes de palabras para cada evento, de tal forma que la representación se concentró principalmente en el *event-trigger* y en menor medida en el resto de los términos de la oración, ponderando como más relevantes a aquellos términos más cercanos al *trigger*. Esta representación de nube de palabras no resultó adecuada para incorporar al grafo causal (cada nodo evento tendría que haber sido una nube de palabras dificultando la legibilidad del grafo). Una posible dirección para mejorar la representación de grafo causal es cambiar la etiqueta de grupo (`CXXX`) por un resumen textual del grupo a través de alguna técnica de resumen automática (*Text summarization*).

Otra mejora posible consiste en complementar la etapa de recuperación causal con herramientas de NLP que recuperan menciones de causalidad explícita del texto. Estás técnicas se puede aplicar sin modificación al *framework*, ya que estas recuperan relaciones causales entre palabras (o conjuntos de palabras) dentro del mismo texto, y como las variables del *framework* son también conjuntos de palabras (unigramas, bigramas, trigramas o *event-triggers* de eventos en curso), las relaciones de causalidad podrían ser directamente agregadas al grafo causal o podrían ser usadas como un voto adicional para el *ensemble* para incrementar la precisión del mismo.

En resumen, el caso de uso presentado fue efectivo en mostrar la viabilidad de la propuesta y en señalar posibles direcciones de trabajo futuro en las cuales se puede mejorar el mismo. Por ejemplo, a través de los trabajos futuros previamente mencionados, que incluyen: (1) la incorporación de más variables de distintos tipos, (2) la construcción de una herramienta interactiva para modificar los diferentes parámetros y visualizar el grafo resultante, (3) mejorar la representación de los nodos mediante la implementación de una herramienta para resumir los grupos de menciones de eventos y (4) agregar la funcionalidad de extraer pares causales directamente del texto a través de herramientas de NLP.

# Bibliografía


[AdC03] Acid, S., and de Campos, L. M. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research 18*, 1 (May 2003), 445–490.

[AGZ+18] Abualsaud, M., Ghelani, N., Zhang, H., Smucker, M. D., Cormack, G. V., and Grossman, M. R. A system for efficient high-recall retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (New York, NY, USA, 2018), SIGIR '18, Association for Computing Machinery, p. 1317–1320.

[AH19] Alsmadi, I., and Hoon, G. K. Term weighting scheme for short-text classification: Twitter corpuses. *Neural Computing and Applications 31*, 8 (Aug 2019), 3819–3831.

[Ahe16] Ahelegbey, D. F. The econometrics of bayesian graphical models: a review with financial application. *Journal of Network Theory in Finance 2*, 2 (2016), 1–33.

[Ahn06] Ahn, D. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events* (USA, 2006), ARTE '06, Association for Computational Linguistics, p. 1–8.

[AIR96] Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*, 434 (1996), 444–455.

[AOGD+16] Adedoyin-Olowe, M., Gaber, M. M., Dancausa, C. M., Stahl, F., and Gomes, J. B. A rule dynamics approach to event detection in twitter

with its application to sports and politics. *Expert Systems with Applications 55* (2016), 351–360.

[APL98] Allan, J., Papka, R., and Lavrenko, V. On-line new event detection and tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 1998), SIGIR '98, Association for Computing Machinery, p. 37–45.

[AX20] Alshaher, H., and Xu, J. A new term weight scheme and ensemble technique for authorship identification. In *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis* (New York, NY, USA, 2020), ICCDA 2020, Association for Computing Machinery, p. 123–130.

[BBS09] Barnett, L., Barrett, A. B., and Seth, A. K. Granger causality and transfer entropy are equivalent for gaussian variables. *Phys. Rev. Lett. 103* (Dec 2009), 238701.

[BCFS19] Balashankar, A., Chakraborty, S., Fraiberger, S., and Subramanian, L. Identifying predictive causal factors from news streams. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China, November 2019), Association for Computational Linguistics, pp. 2338–2348.

[BDL+15] Bronstein, O., Dagan, I., Li, Q., Ji, H., and Frank, A. Seed-based event trigger labeling: How far can event descriptions get us? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (Beijing, China, July 2015), Association for Computational Linguistics, pp. 372–376.

[BGJM17] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics 5* (06 2017), 135–146.

[BHBL11] Bizer, C., Heath, T., and Berners-Lee, T. *Linked Data: The Story so Far*. Semantic Services, Interoperability and Web Applications: Emerging Concepts. IGI Global, Hershey, PA, USA, 2011, pp. 205–227.

[Bor18] Boroş, E. *Neural Methods for Event Extraction*. Theses, Université Paris Saclay (COmUE), September 2018.

[Car90] Card, D. The impact of the mariel boatlift on the miami labor market. *ILR Review 43*, 2 (1990), 245–257.

[CG09] Cecchini, Rocío L.Lorenzetti, C. M., and G., A. Evolving disjunctive and conjunctive topical queries based on multi-objective optimization criteria. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial* (2009).

[CK93] Card, D., and Krueger, A. B. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. Working Paper 4509, National Bureau of Economic Research, October 1993.

[CNL03] Chieu, H. L., Ng, H. T., and Lee, Y. K. Closing the gap: Learning-based information extraction rivaling knowledge-engineering methods. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (Sapporo, Japan, July 2003), Association for Computational Linguistics, pp. 216–223.

[CSG+08] Chiquet, J., Smith, A., Grasseau, G., Matias, C., and Ambroise, C. SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinformatics 25*, 3 (12 2008), 417–418.

[Cun21] Cunningham, S. *Causal Inference: The Mixtape*. Yale University Press, 2021.

[CXL+15] Chen, Y., Xu, L., Liu, K., Zeng, D., and Zhao, J. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Beijing, China, July 2015), Association for Computational Linguistics, pp. 167–176.

[CY13] Cooper, G. F., and Yoo, C. Causal discovery from a mixture of experimental and observational data. *arXiv preprint arXiv:1301.6686* (2013).

[CZLZ16] Chen, K., Zhang, Z., Long, J., and Zhang, H. Turning from tf-idf to tf-igm for term weighting in text classification. *Expert Systems with Applications 66* (2016), 245–260.

[CZSL18] Chen, X., Zhou, X., Sellis, T., and Li, X. Social event detection with retweeting behavior correlation. *Expert Systems with Applications 114* (2018), 516–523.

[DCLT18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[DGB+10] Deisy, C., Gowri, M., Baskar, S., Kalaiarasi, S., and Ramraj, N. A novel term weighting scheme midf for text categorization. *Journal of Engineering Science and Technology 5*, 1 (2010), 94–107.

[DH50] DOLL, R., and HILL, A. B. Smoking and carcinoma of the lung; preliminary report. *British medical journal 2*, 4682 (Sep 1950), 739–748. 14772469[pmid].

[DHZ17] Duan, S., He, R., and Zhao, W. Exploiting document level information to improve event detection via recurrent neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Taipei, Taiwan, November 2017), Asian Federation of Natural Language Processing, pp. 352–361.

[DLY14] Deng, Z.-H., Luo, K.-H., and Yu, H.-L. A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications 41*, 7 (2014), 3506–3513.

[DMP+04] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*

*(LREC'04)* (Lisbon, Portugal, May 2004), European Language Resources Association (ELRA).

[DMPS15] Domeniconi, G., Moro, G., Pasolini, R., and Sartori, C. A study on term weighting for text categorization: A novel supervised variant of tf.idf. In *Proceedings of 4th International Conference on Data Management Technologies and Applications* (Setubal, PRT, 2015), DATA 2015, SCITEPRESS - Science and Technology Publications, Lda, p. 26–37.

[DS04] Debole, F., and Sebastiani, F. Supervised term weighting for automated text categorization. In *Text Mining and its Applications* (Berlin, Heidelberg, 2004), S. Sirmakessis, Ed., Springer Berlin Heidelberg, pp. 81–97.

[DSDN18] Dasgupta, T., Saha, R., Dey, L., and Naskar, A. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue* (Melbourne, Australia, July 2018), Association for Computational Linguistics, pp. 306–316.

[DTY+02] Deng, Z.-H., Tang, S.-W., Yang, D.-Q., Zhang, M., Wu, X.-B., and Yang, M. A linear text classification algorithm based on category relevance factors. In *Digital Libraries: People, Knowledge, and Technology* (Berlin, Heidelberg, 2002), E.-P. Lim, S. Foo, C. Khoo, H. Chen, E. Fox, S. Urs, and T. Costantino, Eds., Springer Berlin Heidelberg, pp. 88–98.

[DU19a] Dogan, T., and Uysal, A. K. Improved inverse gravity moment term weighting for text classification. *Expert Systems with Applications 130* (2019), 45–59.

[DU19b] Dogan, T., and Uysal, A. K. On term frequency factor in supervised term weighting schemes for text classification. *Arabian Journal for Science and Engineering 44*, 11 (Nov 2019), 9545–9560.

[EM07] Eaton, D., and Murphy, K. Exact bayesian structure learning from uncertain interventions. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics* (San Juan, Puerto Rico, 21–24

Mar 2007), M. Meila and X. Shen, Eds., vol. 2 of *Proceedings of Machine Learning Research*, PMLR, pp. 107–114.

[ES19] Eckroth, J., and Schoen, E. A genetic algorithm for finding a small and diverse set of recent news stories on a given subject: How we generate aaai's ai-alert. *Proceedings of the AAAI Conference on Artificial Intelligence 33*, 01 (Jul. 2019), 9357–9364.

[FFSG20] Fan, B., Fan, W., Smith, C., and Garner, H. Adverse drug event detection and extraction from open data: A deep learning approach. *Information Processing & Management 57*, 1 (2020), 102131.

[FGM05] Finkel, J. R., Grenager, T., and Manning, C. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (Ann Arbor, Michigan, June 2005), Association for Computational Linguistics, pp. 363–370.

[FHK+20] Fischbach, J., Hauptmann, B., Konwitschny, L., Spies, D., and Vogelsang, A. Towards causality extraction from requirements. In *2020 IEEE 28th International Requirements Engineering Conference (RE)* (2020), pp. 388–393.

[FLSZ18] Feng, G., Li, S., Sun, T., and Zhang, B. A probabilistic model derived term weighting scheme for text classification. *Pattern Recognition Letters 110* (2018), 23–29.

[FQL18] Feng, X., Qin, B., and Liu, T. A language-independent neural network for event detection. *Science China Information Sciences 61*, 9 (Aug 2018), 092106.

[Fre98] Freitag, D. Information extraction from html: Application of a general machine learning approach. In *AAAI/IAAI* (1998), pp. 517–523.

[FS16] Fattah, M., and Sohrab, M. Combined term weighting scheme using ffnn, ga, mr, sum, & average for text classification. *International Journal of Scientific and Engineering Research 7*, 8 (2016), 2031–2040.

[Gar97] Garcia, D. Coatis, an nlp system to locate expressions of actions connected by causality links. In *Knowledge Acquisition, Modeling and Management* (Berlin, Heidelberg, 1997), E. Plaza and R. Benjamins, Eds., Springer Berlin Heidelberg, pp. 347–352.

[GCR16] Grossman, M. R., Cormack, G. V., and Roegiest, A. Trec 2016 total recall track overview. In *TREC* (2016).

[GM02] Girju, R., and Moldovan, D. I. Text mining for causal relations. In *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference* (2002), AAAI Press, p. 360–364.

[Gra69] Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica 37*, 3 (1969), 424–438.

[GSS00] Galavotti, L., Sebastiani, F., and Simi, M. Experiments on the use of feature selection and negative evidence in automated text categorization. In *Research and Advanced Technology for Digital Libraries* (Berlin, Heidelberg, 2000), J. Borbinha and T. Baker, Eds., Springer Berlin Heidelberg, pp. 59–68.

[HA10] Hirata, Y., and Aihara, K. Identifying hidden common causes from bivariate time series: A method using recurrence plots. *Phys. Rev. E 81* (Jan 2010), 016203.

[HAT+92] Hobbs, J. R., Appelt, D., Tyson, M., Bear, J., and Israel, D. SRI international: Description of the FASTUS system used for MUC-4. In *Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992* (1992).

[HCJ+16] Huang, R., Cases, I., Jurafsky, D., Condoravdi, C., and Riloff, E. Distinguishing past, on-going, and future events: The EventStatus corpus. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, Texas, November 2016), Association for Computational Linguistics, pp. 44–54.

[HDPM18] Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant causal prediction for nonlinear models:. *Journal of Causal Inference 6*, 2 (2018), 20170016.

[HJCV17] Huang, L., Ji, H., Cho, K., and Voss, C. R. Zero-shot transfer learning for event extraction. *arXiv preprint arXiv:1707.01066* (2017).

[HLSP17] Harnack, D., Laminski, E., Schünemann, M., and Pawelzik, K. R. Topological causality in dynamical systems. *Phys. Rev. Lett. 119* (Sep 2017), 098301.

[Hma20] Hmamouche, Y. Nlints: An r package for causality detection in time series. *The R Journal 12* (06 2020), 21–.

[HR11] Huang, R., and Riloff, E. Peeling back the layers: Detecting event role fillers in secondary contexts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, Oregon, USA, June 2011), Association for Computational Linguistics, pp. 1137–1147.

[HZM+11] Hong, Y., Zhang, J., Ma, B., Yao, J., Zhou, G., and Zhu, Q. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, Oregon, USA, June 2011), Association for Computational Linguistics, pp. 1127–1136.

[JG08] Ji, H., and Grishman, R. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT* (Columbus, Ohio, June 2008), Association for Computational Linguistics, pp. 254–262.

[JLH18] Jacobs, G., Lefever, E., and Hoste, V. Economic event detection in company-specific news text. In *Proceedings of the First Workshop on Economics and Natural Language Processing* (Melbourne, Australia, July 2018), Association for Computational Linguistics, pp. 1–10.

[JY16] Jagannatha, A. N., and Yu, H. Bidirectional rnn for medical event detection in electronic health records. *Proceedings of the conference. As-*

*sociation for Computational Linguistics. North American Chapter. Meeting 2016* (Jun 2016), 473–482. 27885364[pmid].

[KBR91] Kaplan, R. M., and Berry-Rogghe, G. Knowledge-based acquisition of causal relationships in text. *Knowledge Acquisition 3*, 3 (1991), 317–337.

[KCN00] Khoo, C. S. G., Chan, S., and Niu, Y. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (USA, 2000), ACL '00, Association for Computational Linguistics, p. 336–343.

[KF09] Koller, D., and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[KJRI91] Krupka, G., Jacobs, P., Rau, L., and Iwanska, L. GE: Description of the NLToolset system as used for MUC-3. In *Third Message Uunderstanding Conference (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21-23, 1991* (1991).

[KS05] Kipper-Schuler, K. *VerbNet: a broad-coverage, comprehensive verb lexicon*. PhD thesis, Computer and Information Science Department, Universiy of Pennsylvania, Philadelphia, PA, 2005.

[kSM12] kwang Song, S., and Myaeng, S. H. A novel term weighting scheme based on discrimination power obtained from past retrieval results. *Information Processing & Management 48*, 5 (2012), 919–930. Large-Scale and Distributed Systems for Information Retrieval.

[KSv+14] Kemmeren, P., Sameith, K., van de Pasch, L., Benschop, J., Lenstra, T., Margaritis, T., O'Duibhir, E., Apweiler, E., van Wageningen, S., Ko, C., van Heesch, S., Kashani, M., Ampatziadis-Michailidis, G., Brok, M., Brabers, N., Miles, A., Bouwmeester, D., van Hooff, S., van Bakel, H., Sluiters, E., Bakker, L., Snel, B., Lijnzaad, P., van Leenen, D., Groot Koerkamp, M., and Holstege, F. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell 157*, 3 (2014), 740–752.

[LCJ03] Lee, C.-S., Chen, Y.-J., and Jian, Z.-W. Ontology-based fuzzy event extraction agent for chinese e-news summarization. *Expert Systems with Applications 25*, 3 (2003), 431–447.

[LCLZ18] Liu, J., Chen, Y., Liu, K., and Zhao, J. Event detection via gated multilingual attention mechanism. In *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[LG10] Liao, S., and Grishman, R. Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (Uppsala, Sweden, July 2010), Association for Computational Linguistics, pp. 789–797.

[Lin57] Lind, J. *A Treatise on the Scurvy*. A. Millar, 1757.

[LJH13] Li, Q., Ji, H., and Huang, L. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Sofia, Bulgaria, August 2013), Association for Computational Linguistics, pp. 73–82.

[LK02] Leopold, E., and Kindermann, J. Text categorization with support vector machines. how to represent texts in input space? *Machine Learning 46*, 1 (Jan 2002), 423–444.

[LLH18] Liu, X., Luo, Z., and Huang, H. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium, October-November 2018), Association for Computational Linguistics, pp. 1247–1256.

[Llo82] Lloyd, S. Least squares quantization in pcm. *IEEE Transactions on Information Theory 28*, 2 (1982), 129–137.

[LLS09] Liu, Y., Loh, H. T., and Sun, A. Imbalanced text classification: A term weighting approach. *Expert Systems with Applications 36*, 1 (2009), 690–701.

[LLZR21] Li, Z., Li, Q., Zou, X., and Ren, J. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *Neurocomputing 423* (2021), 207–219.

[LMG11] Largeron, C., Moulin, C., and Géry, M. Entropy based feature selection for text categorization. In *Proceedings of the 2011 ACM Symposium on Applied Computing* (New York, NY, USA, 2011), SAC '11, Association for Computing Machinery, p. 924–928.

[LML+16] Lorenzetti, C., Maguitman, A., Leake, D., Menczer, F., and Reichherzer, T. Mining for topics to suggest knowledge model extensions. *ACM Trans. Knowl. Discov. Data 11*, 2 (December 2016).

[LMR14] Leake, D., Maguitman, A., and Reichherzer, T. Experience-based support for human-centered knowledge modeling. *Knowledge-Based Systems 68* (2014), 77–87. Enhancing Experience Reuse and Learning.

[LS04] Liu, H., and Singh, P. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal 22*, 4 (Oct 2004), 211–226.

[LSLL09] Liu, D.-R., Shih, M.-J., Liau, C.-J., and Lai, C.-H. Mining the change of event trends for decision support in environmental scanning. *Expert Systems with Applications 36*, 2, Part 1 (2009), 972–984.

[LSLT05] Lan, M., Sung, S.-Y., Low, H.-B., and Tan, C.-L. A comparative study on term weighting schemes for text categorization. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* (2005), vol. 1, pp. 546–551 vol. 1.

[LTSL09] Lan, M., Tan, C. L., Su, J., and Lu, Y. Supervised and traditional term weighting methods for automatic text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 4 (2009), 721–735.

[MAC14] Ma, H., Aihara, K., and Chen, L. Detecting causality from nonlinear dynamics with short-term time series. *Scientific Reports 4*, 1 (Dec 2014), 7464.

[Mai19] Maisonnave, M. Detección de textos similares a través de una técnica de agrupamiento basada en densidad. In *Communication at the XV Dr. Antonio Monteiro Congress, Bahía Blanca, Argentina* (2019).

[MCCD13] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[MDT+20a] Maisonnave, M., Delbianco, F., Tohmé, F., Maguitman, A., and Milios, E. Event detection dataset. Mendeley Data, V1 - http://dx.doi.org/10.17632/7d54rvzxkr.1, 2020.

[MDT+20b] Maisonnave, M., Delbianco, F., Tohmé, F., Maguitman, A. G., and Milios, E. E. Assessing causality structures learned from digital text media. In *Proceedings of the ACM Symposium on Document Engineering 2020* (New York, NY, USA, 2020), DocEng '20, Association for Computing Machinery.

[MDT+21] Maisonnave, M., Delbianco, F., Tohmé, F., Maguitman, A., and Milios, E. Detecting ongoing events using contextual word and sentence embeddings. *arXiv preprint arXiv:2007.01379* (2021).

[MDTM18] Maisonnave, M., Delbianco, F., Tohmé, F. A., and Maguitman, A. G. A supervised term-weighting method and its application to variable extraction from digital media. In *XIX Simposio Argentino de Inteligencia Artificial (ASAI)-JAIIO 47 (CABA, 2018)* (2018), p. 40–53.

[MDTM19a] Maisonnave, M., Delbianco, F., Tohmé, F., and Maguitman, A. Economic relevant news from the guardian. Mendeley Data, V3 - http://dx.doi.org/10.17632/yt8j2f3hpp.3, 2019.

[MDTM19b] Maisonnave, M., Delbianco, F., Tohmé, F. A., and Maguitman, A. G. A flexible supervised term-weighting technique and its application to variable extraction and information retrieval. *Inteligencia Artificial 22*, 63 (Feb. 2019), 61–80.

[MDTM21] Maisonnave, M., Delbianco, F., Tohmé, F., and Maguitman, A. Assessing the behavior and performance of a supervised term-weighting tech-

nique for topic-based retrieval. *Information Processing & Management 58*, 3 (2021), 102483.

[Mil95] Miller, G. A. Wordnet: A lexical database for english. *Commun. ACM 38*, 11 (November 1995), 39–41.

[MLRM04] Maguitman, A., Leake, D., Reichherzer, T., and Menczer, F. Dynamic extraction topic descriptors and discriminators: Towards automatic context-based topic search. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management* (New York, NY, USA, 2004), CIKM '04, Association for Computing Machinery, p. 463–472.

[MSC+13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26* (10 2013), 3111–3119.

[NCG16] Nguyen, T. H., Cho, K., and Grishman, R. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, California, June 2016), Association for Computational Linguistics, pp. 300–309.

[NFCG16] Nguyen, T. H., Fu, L., Cho, K., and Grishman, R. A two-stage approach for extending event detection to new types via neural networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP* (Berlin, Germany, August 2016), Association for Computational Linguistics, pp. 158–165.

[NG15] Nguyen, T. H., and Grishman, R. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (Beijing, China, July 2015), Association for Computational Linguistics, pp. 365–371.

[NG16] Nguyen, T. H., and Grishman, R. Modeling skip-grams for event detection with convolutional neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, Texas, November 2016), Association for Computational Linguistics, pp. 886–891.

[NG18] Nguyen, T., and Grishman, R. Graph convolutional networks with argument-aware pooling for event detection. In *Thirty-second AAAI conference on artificial intelligence* (2018).

[NMB17] Nicholson, W., Matteson, D., and Bien, J. Bigvar: Tools for modeling sparse high-dimensional multivariate time series. *arXiv preprint arXiv:1702.07094* (2017).

[Nov90] Novak, J. D. Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching 27*, 10 (1990), 937–949.

[ONSH20] Ombadi, M., Nguyen, P., Sorooshian, S., and Hsu, K.-l. Evaluation of methods for causal discovery in hydrometeorological systems. *Water Resources Research 56*, 7 (2020), e2020WR027251. e2020WR027251 2020WR027251.

[PB15] Peters, J., and Bühlmann, P. Structural intervention distance for evaluating causal graphs. *Neural Computation 27*, 3 (2015), 771–799.

[PBM16] Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 78*, 5 (2016), 947–1012.

[PCI+03] Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering 3* (2003), 28–34.

[Pea85] Pearl, J. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA* (1985), pp. 15–17.

[Pea09] Pearl, J. *Causality*, 2 ed. Cambridge University Press, 2009.

[PFM18] Pinho, C., Franco, M., and Mendes, L. Web portals as tools to support information management in higher education institutions: A systematic literature review. *International Journal of Information Management 41* (2018), 80–92.

[PJS17] Peters, J., Janzing, D., and Schölkopf, B. *Elements of Causal Inference : Foundations and Learning Algorithms*. The MIT Press, 2017.

[PK07] Pechsiri, C., and Kawtrakul, A. Mining causality for explanation knowledge from text. *Journal of Computer Science and Technology 22*, 6 (2007), 877.

[PM18] Pearl, J., and Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*, 1st ed. Basic Books, Inc., USA, 2018.

[PNI+18] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[PR09] Patwardhan, S., and Riloff, E. A unified model of phrasal and sentential evidence for information extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore, August 2009), Association for Computational Linguistics, pp. 151–160.

[PSM14] Pennington, J., Socher, R., and Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar, October 2014), Association for Computational Linguistics, pp. 1532–1543.

[QWQ11] Quan, X., Wenyin, L., and Qiu, B. Term weighting schemes for question categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 5 (2011), 1009–1021.

[RBB+19] Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J. Inferring causation from time

series in earth system sciences. *Nature Communications 10*, 1 (Jun 2019), 2553.

[RCGC15] Roegiest, A., Cormack, G. V., Grossman, M. R., and Clarke, C. Trec 2015 total recall track overview. *Proc. TREC-2015* (2015).

[RDM12] Radinsky, K., Davidovich, S., and Markovitch, S. Learning causality for news events prediction. In *Proceedings of the 21st International Conference on World Wide Web* (New York, NY, USA, 2012), WWW '12, Association for Computing Machinery, p. 909–918.

[Rij79] Rijsbergen, C. J. V. *Information Retrieval*, 2nd ed. Butterworth-Heinemann, USA, 1979.

[Ril96a] Riloff, E. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2* (1996), AAAI'96, AAAI Press, p. 1044–1049.

[Ril96b] Riloff, E. An empirical study of automated dictionary construction for information extraction in three domains. *Artificial Intelligence 85*, 1 (1996), 101–134.

[RJ76] Robertson, S. E., and Jones, K. S. Relevance weighting of search terms. *Journal of the American Society for Information Science 27*, 3 (1976), 129–146.

[RNK+19] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances 5*, 11 (2019).

[Rob04] Robertson, S. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation 60*, 5 (Jan 2004), 503–520.

[San08] Sandhaus, E. The new york times annotated corpus LDC2008T19. *Linguistic Data Consortium, Philadelphia 6*, 12 (2008), e26752.

[SB88] Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management 24*, 5 (1988), 513–523.

[SBMM19] Samant, S. S., Bhanu Murthy, N. L., and Malapati, A. Improving term weighting schemes for short text classification in vector space model. *IEEE Access 7* (2019), 166578–166592.

[Sch00] Schreiber, T. Measuring information transfer. *Phys. Rev. Lett. 85* (Jul 2000), 461–464.

[Scu10] Sculley, D. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web* (New York, NY, USA, 2010), WWW '10, Association for Computing Machinery, p. 1177–1178.

[SG91] Spirtes, P., and Glymour, C. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review 9*, 1 (1991), 62–72.

[SGSH00] Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

[SHHK06] Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research 7*, 72 (2006), 2003–2030.

[Sim80] Sims, C. A. Macroeconomics and reality. *Econometrica 48*, 1 (1980), 1–48.

[SIS+11] Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *J. Mach. Learn. Res. 12*, null (July 2011), 1225–1248.

[SKW07] Suchanek, F. M., Kasneci, G., and Weikum, G. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web* (New York, NY, USA, 2007), WWW '07, Association for Computing Machinery, p. 697–706.

[SLB+21] Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE 109*, 5 (2021), 612–634.

[SMY+12] Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., and Munch, S. Detecting causality in complex ecosystems. *Science 338*, 6106 (2012), 496–500.

[Sno56] Snow, J. On the mode of communication of cholera. *Edinburgh medical journal 1*, 7 (Jan 1856), 668–670. 29647347[pmid].

[SPP+05] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science 308*, 5721 (2005), 523–529.

[SQCS18] Sha, L., Qian, F., Chang, B., and Sui, Z. Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction. *Proceedings of the AAAI Conference on Artificial Intelligence 32*, 1 (Apr. 2018).

[SSG+98] Scheines, R., Spirtes, P., Glymour, C., Meek, C., and Richardson, T. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research 33*, 1 (1998), 65–117. PMID: 26771754.

[STA06] Surdeanu, M., Turmo, J., and Ageno, A. A hybrid approach for the acquisition of information extraction patterns. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)* (2006).

[TC60] Thistlethwaite, D. L., and Campbell, D. T. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology 51*, 6 (1960), 309.

[Tho53] Thorndike, R. L. Who belongs in the family? *Psychometrika 18*, 4 (Dec 1953), 267–276.

[Tib96] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*, 1 (1996), 267–288.

[TLL20] Tang, Z., Li, W., and Li, Y. An improved term weighting scheme for text classification. *Concurrency and Computation: Practice and Experience 32*, 9 (2020), e5604. e5604 CPE-19-0287.R1.

[TM94] Tokunaga, T., and Makoto, I. Text categorization based on weighted inverse document frequency. In *Special Interest Groups and Information Process Society of Japan (SIG-IPSJ* (1994), pp. 33–39.

[TWC+20] Tong, M., Wang, S., Cao, Y., Xu, B., Li, J., Hou, L., and Chua, T.-S. Image enhanced event detection in news articles. *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 05 (Apr. 2020), 9040–9047.

[Var14] Varian, H. R. Big data: New tricks for econometrics. *Journal of Economic Perspectives 28*, 2 (May 2014), 3–28.

[VH99] Voorhees, E., and Harman, D. Overview of the eight text retrieval conference. In *Proc. TREC-8, the 8th text retrieval conference* (1999).

[vHP81] van Rijsbergen, C., Harper, D., and Porter, M. The selection of good search terms. *Information Processing & Management 17*, 2 (1981), 77–91.

[VSHK16] Verberne, S., Sappelli, M., Hiemstra, D., and Kraaij, W. Evaluation and analysis of term scoring methods for term extraction. *Information Retrieval Journal 19*, 5 (Oct 2016), 510–545.

[WBL+14] Wu, S., Bondugula, S., Luisier, F., Zhuang, X., and Natarajan, P. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2014).

[WGG17] Wu, H., Gu, X., and Gu, Y. Balancing between over-weighting and under-weighting in supervised term weighting. *Information Processing & Management 53*, 2 (2017), 547–557.

[WL21] Weng, J., and Lee, B.-S. Event detection in twitter. *Proceedings of the International AAAI Conference on Web and Social Media 5*, 1 (Aug. 2021), 401–408.

[WSMM06] Walker, C., Strassel, S., Medero, J., and Maeda, K. ACE 2005 multilingual training corpus LDC2006T06. *Linguistic Data Consortium, Philadelphia 57* (2006).

[WZ13] Wang, D., and Zhang, H. Inverse-category-frequency based supervised term weighting schemes for text categorization. *Journal of Information Science and Engineering 29*, 2 (2013), 209–225. cited By 28.

[YGTH00] Yangarber, R., Grishman, R., Tapanainen, P., and Huttunen, S. Automatic acquisition of domain knowledge for information extraction. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics* (2000).

[ZJS19] Zhang, T., Ji, H., and Sil, A. Joint Entity and Event Extraction with Generative Adversarial Imitation Learning. *Data Intelligence 1*, 2 (04 2019), 99–120.

[ZLZ+16] Zhao, S., Liu, T., Zhao, S., Chen, Y., and Nie, J.-Y. Event causality extraction based on connectives analysis. *Neurocomputing 173* (2016), 1943–1950.

[ZWM+17] Zhao, S., Wang, Q., Massung, S., Qin, B., Liu, T., Wang, B., and Zhai, C. Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2017), WSDM '17, Association for Computing Machinery, p. 335–344.