Análisis del Descubrimiento Causal en Series de Tiempo

Julio César Muñoz Benítez, Luis Enrique Sucar Succar

Reporte Técnico No. CCC-21-002
Abril, 2021

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.

Resumen

El Descubrimiento Causal en series de tiempo es un punto clave para la generación de conocimiento causal en diversas disciplinas, fenómenos o eventos de interés de acuerdo al contexto de la investigación. Esto se logra mediante la identificación de las relaciones causales, y la fuerza de éstos enlaces, en las variables que conforman la serie de tiempo. Sin embargo, existen retos que deben ser abordados en el análisis causal de las series de tiempo. Uno de los principales retos es el submuestreo, es decir, la velocidad con la que son observadas las variables puede no concordar con la velocidad en la que surgen las relaciones causales. De esta forma, como parte de este trabajo de investigación, se realizó un análisis de los trabajos relacionados con el análisis del descubrimiento causal en series de tiempo, detectando trabajos en la reconstrucción y análisis causal de series de tiempo, así como en la generación de datos sintéticos basados en modelos estructurales causales . Asimismo, se configuraron y realizaron diversos escenarios para analizar los resultados de estos trabajos. Estos experimentos sirvieron para analizar la afectación que los diversos cambios en los retrasos de tiempo tienen en el análisis de las relaciones causales de la serie de tiempo. Esto último representa un campo de oportunidad para la presente investigación.

Palabras clave— Descubrimiento Causal, Análisis Causal, Series de Tiempo, Submuestreo, Gráficos Causales.

Índice

1. Introducción

El principal objetivo del Descubrimiento Causal es identificar las relaciones existente en las variables de datos observados de algún fenómeno o situación. Recientemente, el interés en identificar y analizar los efectos causales en series de tiempo se ha visto incrementado, esto debido que la relación de los datos con el tiempo permite modelar diversos fenómenos en varias áreas de investigación. Esta relación puede llegar a facilitar la distinción entre la causa y el efecto. Además, dependiendo de la forma con la que se obtienen estos datos, se pueden proveer mediciones precisas en lapsos de tiempo regulares.

Sin embargo, la velocidad con la que estos datos son medidos puede no concordar con la velocidad con la que las relaciones causales pueden aparecer en el sistema complejo y dinámico que es observado. Esto puede conducir a análisis erróneos de la información causal obtenida. De esta forma, el análisis causal presenta desafíos que deben ser atendidos para descubrir las relaciones causales verdaderas de las variables que son observadas, permitiendo obtener información de interés e importancia en el contexto de la investigación que se esté llevando acabo.

Para esto, uno de los enfoques propuestos, en la literatura , es la reconstrucción de la estructura causal del conjunto de datos observados. En este sentido, se puede tener una reconstrucción de la serie de tiempo especificando los enlaces causales, así como su dirección, fuerza causal y el instante de tiempo en que aparecen. Sin embargo, para analizar y evaluar este tipo de enfoques muchas veces se requiere de conjuntos de datos que son complicado de obtener, esto debido al tipo de fenómeno de interés, a la dimensionalidad de los datos, al extenso tiempo de muestreo o al alto costo computacional y de recursos necesarios para obtener las mediciones. Para esto, la generación de datos sintéticos puede resultar útil en la composición de series de tiempo, con la cantidad de observaciones necesarias, para realizar las pruebas y evaluaciones en el descubrimiento causal en series de tiempo.

En este sentido, como parte de la presente investigación, se realizó la lectura de trabajos relacionados con el descubrimiento causal en series de tiempo, esto con el objetivo de detectar los desafíos que se tienen, así como los enfoques utilizados en esta área de investigación, esto con el objetivo de identificar áreas de oportunidad para desarrollar una propuesta de investigación. Dentro de los problemas identificados se encuentran la alta dimensionalidad de los datos, así como la falta de datos de acuerdo al área de investigación o de interés. Además, los sistemas complejos cuentan con la interacción de múltiples variables que no presentan una relación causal entre sí, pero si pueden presentar una correlación estadística. Aunado a esto, la mayoría de los autores mencionan que uno de los principales factores a tomar en cuenta es la relación entre la escala del tiempo de observación y la escala de tiempo causal, es decir, que las relaciones causales entre las variables estén presentes durante el tiempo de observación de la serie de tiempo. Esto último representa un área de oportunidad dentro del Descubrimiento Causal en series de tiempo, por lo que es necesario un análisis profundo para determinar las relaciones causales reales entre las variables observadas.

El presente documento está organizado de la siguiente manera: en la sección 2 se realizó una revisión preliminar de los trabajos relacionados con el descubrimiento causal, detectando los retos en el análisis causal de series de tiempo. La sección 3 presenta los resultados de los primeros experimentos en la reconstrucción de series de tiempo sintéticas, generadas bajo diversas configuraciones, esto con el objetivo de analizar los resultados del análisis de las relaciones causales de las variables, que conforman estas series de tiempo, con diversas configuraciones en el retraso en que los enlaces causales entre variables pueden suceder. En la sección 4 se mencionan brevemente las conclusiones de los experimentos realizados, así como del análisis preliminar de los trabajos relacionados. Por último, en la sección 5 se menciona el trabajo futuro que se llevará acabo en la presente investigación en curso.

2. Trabajos Relacionados

El conocimiento causal es un elemento crítico para responder diversos cuestionamientos e identificar las relaciones causales que permitan analizar sistemas complejos, y dinámicos, en escenarios y aplicaciones del mundo real. En este sentido, uno de los objetivos del Descubrimiento Causal es identificar las relaciones entre las variables de los datos

observados, donde el interés radica en identificar los efectos causales de las observaciones con respecto al tiempo.

Los datos que conforman series de tiempo permiten modelar información observada de diversos fenómenos dinámicos del mundo real, teniendo aplicaciones en diversas áreas de la ciencia como: economía, neurociencia, análisis climático, por mencionar algunas (Runge, Nowack, Kretschmer, Flaxman, y Sejdinovic, 2019). Uno de los principales beneficios de analizar una serie de tiempo es que el orden puede ayudar a distinguir la causa del efecto, esto debido a que el futuro no puede afectar el pasado y se puede identificar la variable que ocurre primero. Comparado con el análisis de datos independientes e idénticamente distribuidos (iid), el análisis de la causalidad en series de tiempo presenta diversos desafíos que deben ser atendidos (Lawrence, Kaiser, Sampaio, y Sipos, 2020).

En la literatura existen algunos métodos que tratan de encontrar y definir las relaciones causales en series de tiempo. Esta no es una tarea trivial debido a que el tiempo con el que se adquieren los datos puede ser más lento que el tiempo en el que aparecen las relaciones causales, puede existir algún error en la medición o el sistema puede no ser estacionario, es decir, puede ser un sistema dinámico cuya distribución de probabilidades de las variables puede cambiar, e incluso, generar nuevas relaciones causales. De esta forma, existen diversas estrategias para analizar series de tiempo. Una de estas es asumir o estimar un determinado número de efectos retrasados y tratar las mediciones separadas por nomás que ese número de retrasos como una unidad de análisis de las mediciones u observaciones. Para esto, Granger (1969) propuso una forma práctica para definir la causalidad con base en la predicción. La idea principal consiste en cuantificar si X causa Y ya que esto implica que existe información en X relevante para Y que no está contenida en el pasado de Y como información observada. La medición de la predicción puede ser implementada de diversas maneras, siendo los modelos de vectores autoregresivos (VAR) la más común. Una de las principales desventajas de esto es limitar el análisis a aplicaciones bivariadas, lo que puede dejar de lado enlaces causales directos en conjuntos de datos con alta dimensionalidad. Además, otra de las desventajas de la causalidad de Granger es su sensibilidad al tiempo, esto es, si los datos son submuestreados o agregados temporalmente es problabe que las relaciones causales reales de la serie de tiempo no puedan ser analizadas correctamente.

En este sentido, los datos que conforman una serie de tiempo proveen, en muchos casos, mediciones precisas en períodos de tiempo regulares pero las interacciones causales que pueden surgir por estas mediciones se pueden presentar en períodos de tiempo más rápidos que las mediciones observadas (Hyttinen, Plis, Järvisalo, Eberhardt, y Danks, 2017) o incluso las dependencias causales pueden aparecer fuera del tiempo de observación (Runge et al., 2019). De esta forma, la información contenida en el orden del tiempo puede simplificar el análisis causal, ya que puede proveer la direccionalidad en las relaciones causales, pero si no se realiza un muestreo adecuado se puede perder información relevante que guíe a análisis erróneos acerca de las conexiones causales reales de los datos contenidos en la serie de tiempo. A esto se le conoce como submuestreo o undersamplling (Hyttinen et al., 2017).

El submuestreo es uno de los principales desafíos en el descubrimiento causal en series de tiempo (Danks y Plis, 2014), donde se asume que la escala de tiempo en las observaciones concuerda con la escala de tiempo causal, es decir, se asume que se están observando los datos con el tiempo adecuado para observar las relaciones causales directas. Las estructuras causales temporales pueden representadas usando versiones dinámicas de los modelos causales gráficos, conocidas como redes dinámicas Bayesianas (DBN, dynamic Bayesian networks), de esta forma el tiempo es modelado en incrementos discretos. Para esto, se asume que los nodos que conforman las observaciones están conectados mediante la condición causal de Markov y causal faithfulness.

En la Figura 1(a) se muestra una serie de tiempo con las relaciones causales detectadas para cada una de las variables (X, Y, Z) para cada muestreo de tiempo (t), por otro lado, la Figura 1b muestra la estructura causal correspondiente a la misma serie de tiempo pero tomando muestras cada dos instantes de tiempo (t-2).

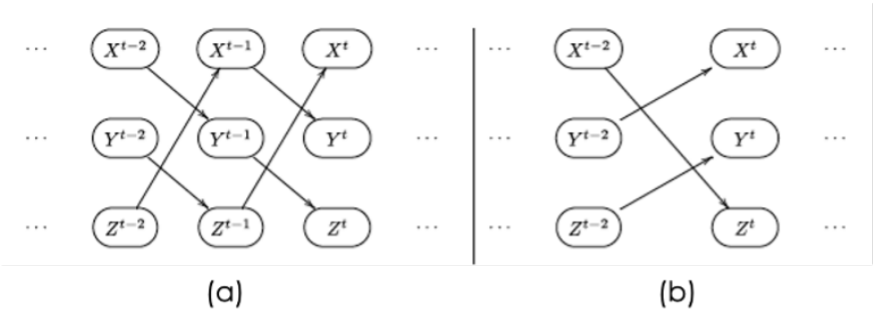

Figura 1. (a) La estructura causal de la serie de tiempo original. (b) La estructura causal de la serie de tiempo sí solo se contempla la observación cada dos muestreos de tiempo (Hyttinen et al., 2017).

Como se puede observar, la estructura causal en ambos casos es diferente, ya que si se realiza el muestreo cada dos instantes de tiempo los enlaces causales en *t-1* son discriminados. De esta forma, si no se toma en consideración la posibilidad de realizar submuestreo se puede asumir que la estructura mostrada en la Figura 1b es la correcta, perdiendo así la información de los enlaces causales reales, lo que puede llevar a un análisis erróneo donde Z puede influir en Y, siendo esta última realmente mediada por X.

Para esto, Danks y Plis (2014) proponen el análisis del cambio de la estructura causal a diferentes rangos de submuestreo para identificar las características de la verdadera estructura. Esto se logra mediante la noción de componente fuertemente conectado (SCC, strongly connected component). Este se define como un conjunto máximo de nodos $\mathbf{S} \subseteq \mathbf{V}$ tales que para cada X, $Y \in \mathbf{S}$ existe un enlace directo de X a Y. Si el gráfico no cuenta con ningún SSC se puede inferir que el submuestreo destruyó los enlaces directos. De esta forma, el submuestreo puede incrementarse para analizar si se cuenta con algún SSC, incluso si se cuenta con un lazo al mismo nodo, es decir, una variable que es causa de su propio valor en el siguiente lapso de tiempo. En este sentido, el submuestreo puede incrementarse hasta converger en una gráfica vacía. Esta convergencia destruye toda la información acerca de la estructura interna del conjunto de datos. Sin embargo, si no ha convergido hasta este nivel, se puede obtener conocimiento sobre la existencia de un lazo hacia el mismo nodo.

La principal aportación de este trabajo es proponer una base para el aprendizaje causal dado el desconocimiento del grado de submuestreo. Un enfoque propuesto es el desarrollo de un algoritmo basado en puntuación que busque las posibles estructuras causales itinerando entre diferentes grados de submuestreo. Así, se pueden diseñar múltiples algoritmos que permitan extraer algunas características de la verdadera estructura causal de la serie de tiempo submuestreada.

Otro de los trabajos relacionados con la reconstrucción de la estructura gráfica en series de tiempo y la detección de asociaciones causales es propuesto por (Runge et al., 2019) donde se desarrolló un módulo en Python para el análisis del descubrimiento causal basado en la reconstrucción de series de tiempo con alta dimensionalidad de datos. La propuesta es la implementación de pruebas lineales y no lineales de independencia causal junto con un algoritmo de descubrimiento para estimar las redes causales de los datos que conforman la serie de tiempo. De esta forma se busca estimar la intensidad de la relación causal así como la dirección de los enlaces entre las variables.

Este método de descubrimiento causal implementa la estimación de independencias condicionales a través de las muestras de la serie de tiempo, las cuales son truncadas al alcanzar un retraso máximo de tiempo. Este retraso máximo depende de la aplicación o puede ser especificado de acuerdo con el retraso causal máximo esperado en el sistema complejo, o puede estar basado en el retraso más prolongado con la correlación más significativa. Los autores definen

esto como prueba de independencia condicional completa (full conditional independence, FullCI).

De esta forma, al usar las gráficas de las series de tiempo se puede proveer una vista para comprender mejor los enlaces causales. En este sentido, los nodos en la gráfica de tiempo representan las variables en los diferentes retrasos de tiempo (Figura 2A). El objetivo del descubrimiento causal es estimar los padres causales en la serie de tiempo. Sin embargo, la prueba FullCI presenta deficiencias al manejar series de tiempo con alta dimensionalidad, ya que FullCI prueba directamente los enlaces condicionados de cada una de las variables, esto reduce drásticamente el poder de detección de los enlaces causales al incrementar el número de variables en la serie de tiempo. Por otro lado, si sólo se condiciona un conjunto de variables que al menos incluyan los padres de una de las variables de la serie de tiempo puede ser suficiente para identificar enlaces erroneos. Por lo que, los algoritmos de descubrimiento de Markov, tales como, el algoritmo PC, permite detectar estos padres y puede ser implementado con diversos modos de prueba de independencia condicional para funciones de dependencia no lineales o variables continuas o discretas. Debido a esto, los autores definen un método basado en el uso del algoritmo PC, para identificar las condiciones relevantes de los padres para todas las variables en la serie de tiempo, y una prueba de independencia condicional momentaria (MCI, momentary conditional independence) para analizar la fuerza de los enlaces causales y así detectar falsos enlaces causales en la serie de tiempo. El método propuesto es conocido como PCMCI, esto es, en la primera parte se remueven las condiciones irrelevantes para cada una de las variables que conforman la serie de tiempo iterando pruebas de independencia (Figura 2B, colores rojos y azules). En la primera iteración las variables que no cuentan con una dependencia incondicional son removidas (los colores más claros de rojo y azul), durante la segunda iteración las variables que se volvieron condicional independiente con la dependencia más grande en la iteración previa son removidas. De esta forma, el algoritmo PC continua la iteración hasta que solo queden unas pocas condiciones relevantes (colores oscuros de rojo y azul), esto incluye los padres causales con alta probabilidad y falsos positivos potenciales (denotados con una estrella en la Figura 2B).

Durante la segunda etapa, se realiza la prueba MCI, la cual ayuda a estimar si las condiciones causales de los padres detectados en la etapa anterior (cuadros azules en la Figura 2C) son suficientes para establecer independencia condicional, esto es, identificar enlaces causales indirectos y comunes. Además, las condiciones adicionales de los padres, en otros retrasos de tiempo (cuadros rojos en la Figura 2C), ayudan a determinar la autocorrelación, con el objetivo de controlar los falsos positivos. La fuerza de cada enlace puede ser evaluado de acuerdo a los valores de probabilidad de la prueba MCI. Esto puede ser ajustado de acuerdo a la proporción del descubrimiento de enlaces falsos.

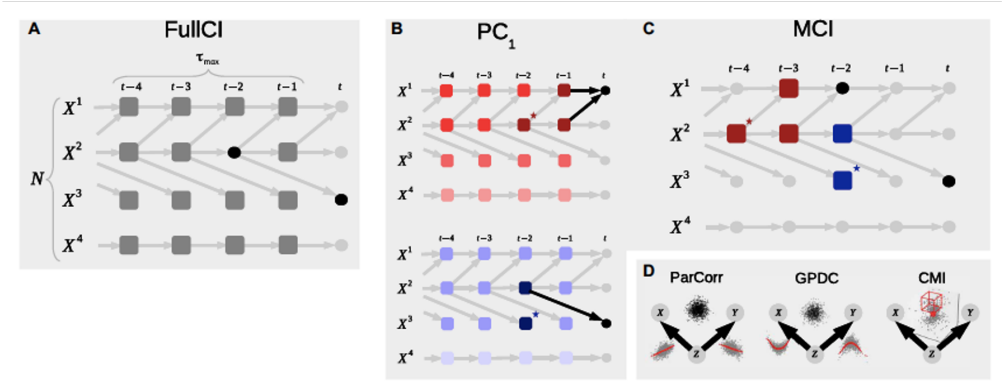

Figura 2. Método de descubrimiento causal propuesto por Runge et al. (2019).

Por último, en la Figura 2D, tanto las etapas PC como MCI pueden ser combinadas con pruebas de independencia lineal (ParCorr) o no lineal (GPDC y CMI). ParCorr asume modelos de ruido aditivo y GPDC sólo aditivo, las gráficas de dispersión grises en la Figura 2D muestran la regresión de X y Y en Z, mientras que las gráficas de dispersión negras muestran los residuos.

Como resultado, el método propuesto arroja un análisis robusto de las condiciones causales entre variables de la serie de tiempo, teniendo un una mejora significativa comparados con algoritmos convencionales como PC o causalidad de Granger. Esto permite estimar la intensidad de los enlaces de estas variables. De esta forma, se pueden identificar los enlaces causales más fuertes que contengan información de interés de acuerdo con el contexto de la investigación. Esto mismo puede servir para identificar rutas de mediación causal para modificar la influencia causal de variables individuales.

Sin embargo, una de las deficiencias detectadas por los autores es la falta de conjuntos de datos que cumplan con ciertas características para la correcta evaluación del desempeño de nuevos algoritmos de descubrimiento causal. Muchas veces no se cuenta con datos reales de experimentos o escenarios del mundo real. Sumado a esto, los datos disponibles pueden no estar ajustados al período de interés para la evaluación y análisis causal.

En este sentido, uno de las opciones es utilizar datos sintéticos que permitan la correcta evaluación de los modelos de descubrimiento causal para medir su desempeño. Así, para el análisis y evaluación del descubrimiento causal se utilizó la generación de datos sintéticos propuesto por (Lawrence et al., 2020), quienes establecen que para analizar el descubrimiento causal, en casos del mundo real, en ciertas ocasiones es imposible contar u observar todas las variables involucradas para asegurar la relación causal, es decir, debido a las limitantes como el procesamiento, las fuentes de datos o el tiempo de muestreo las observaciones disponibles no pudieran ser suficientes para determinar una relación causal real. De esta forma, si se considera un sistema con una serie de tiempo altamente correlacionada podría ser difícil identificar si un efecto observado proviene de una sola serie de tiempo, de un subconjunto o de todo el conjunto de series de tiempo (Lawrence et al., 2020). Debido a esto, los autores proponen un framework para generar grandes cantidades de datos, con diferentes propiedades, incluyendo el número de observaciones y el número de variables (tanto observadas como latentes u ocultas), esto para realizar el análisis de una metodología o caso específico o incluso analizar cómo el sistema escala en complejidad con el número de series de tiempo o la adición de otras variables.

Para esto los autores utilizan modelos causales estructurales (SCM, structural causal models) los cuales asumen que los nodos en una gráfica causal tienen una función de dependencia con sus padres. Esto es, dado un conjunto de variables, cada variable puede ser representada en términos de una función en conjunto con sus padres.

El proceso general sigue tres pasos: especificar los elementos que conforman el gráfico causal de la serie de tiempo; especificar y generar el modelo causal estructural; y especificar el ruido y la configuración del tiempo de ejecución para generar el conjunto de datos sintéticos que conforman la serie de tiempo. En la Figura 3 se muestra el proceso propuesto por los autores, donde se especifican las características básicas del modelo gráfico, el cual es generado de forma aleatoria, así como el modelo causal estructural.

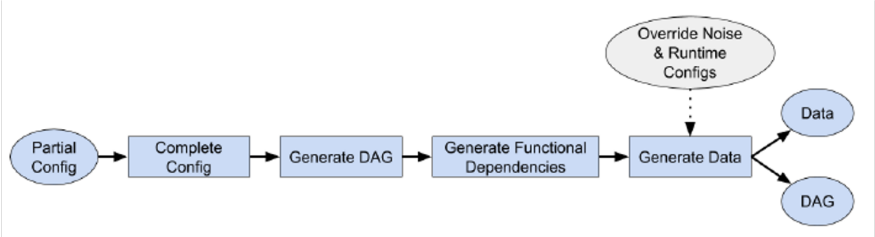

Figura 3. Proceso de generación de datos sintéticos propuesta por Lawrence et al. (2020).

Dada una configuración parcial, o completa, se genera aleatoriamente una gráfica causal de la serie de tiempo. De esta forma, para cada nodo y enlace de la DAG, la dependencia causal es generada de forma aleatoria, resultando a su vez en un modelo causal estructural aleatorio. Una vez definido este modelo estructural se generan los datos. De

un solo modelo estructural se pueden generar múltiples conjuntos de datos, esto con el objetivo de evaluar diversos escenarios. El usuario puede modificar las distribuciones del ruido o volver a generar el conjunto de datos para obtener datos de variables no observadas.

Los autores exponen cuatro tipos de variables: objetivos, características, latentes y ruido. De esta forma se puede configurar la estructura para casos específicos de acuerdo al contexto del análisis. Un ejemplo de los datos generados se puede apreciar en la Figura 4.

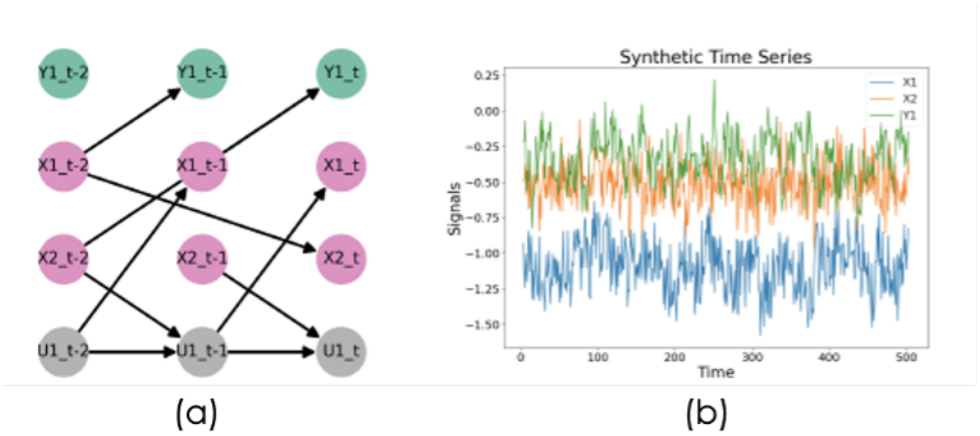

Figura 4. Gráfica causal y datos sínteticos de la serie de tiempo (Lawrence et al., 2020).

La Figura 4a muestra la gráfica causal de la serie de tiempo, donde se muestran los enlaces causales de cada una de las variables observadas (X), la variable objetivo (Y) y la variable latente o no observada (U), de esta forma se puede comprobar que no se cuenta con suficiencia causal ya que existe una variable no observada.

Por otro lado, la Figura 4b muestra la gráfica de los datos que componen la serie de tiempo sintética. Así, se pueden generar varios conjuntos de datos, que estén basados en el mismo modelos estructural causal, modificando la distribución del ruido.

Una de las principales ventajas de generar una serie de tiempo sintética para el análisis del desempeño de los algoritmos de descubrimiento causal es que se pueden comparar los resultados de los enlaces causales descubiertos y los enlaces originales propuestos en la gráfica causal y el modelo causal estructural. De esta forma, se pueden realizar las adecuaciones necesarias para analizar diversos escenarios controlados y, posteriormente, analizar escenarios reales.

3. Reconstrucción de Series de Tiempo

Como parte de este análisis preliminar se configuraron ambos modelos para la generación de series de tiempo sintéticas y, posteriormente, analizar su reconstrucción y el descubrimiento causal, esto con el objetivo de entender los conceptos del descubrimiento causal en series de tiempo y así detectar áreas de oportunidad para la propuesta de investigación.

Como un trabajo inicial se realizó el análisis de la reconstrucción de una serie de tiempo con base en datos sintéticos generados en Phyton utilizando el trabajo propuesto por Lawrence et al. (2020). De esta forma, se pudo observar el resultado del descubrimiento causal y comparar ambas gráficas. La Tabla 1 muestra la configuración básica para la generación de la serie de tiempo[1].

[1]La configuración completa se muestra en el Anexo I.

Tabla 1. Configuración básica para la generación de la serie de tiempo.

Atributo	Valor
Lag mínimo	1
Lag máximo	2
Número de objetivos	1
Número de características	2
Número de variables latentes	1
Varianza del ruido	0.01 – 0.1
Observaciones	1,000

Esta configuración sirvió para generar los datos sintéticos que forman la serie de tiempo. La Figura 5 muestra la gráfica causal resultante, donde se especifican los enlaces causales y el tiempo en el que estos son generados.

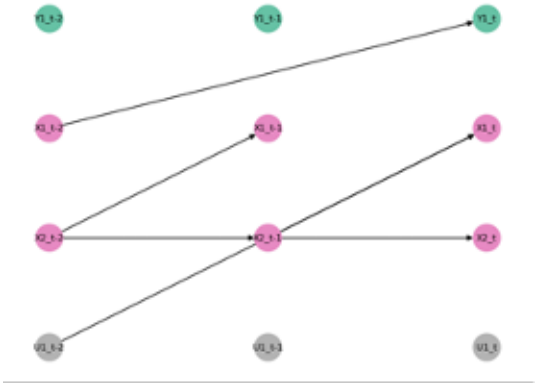

Figura 5. Gráfica causal generada de forma aleatoria

Asimismo, se generaron las gráficas correspondientes de la correlación entre la variable Y1 y X1 en el lag 2, es decir, en el tiempo en el que se puede observar el enlace causal (Figura 6a), así como la gráfica de los valores para cada una de las variables: X1, X2 y Y1(Figura 6b).

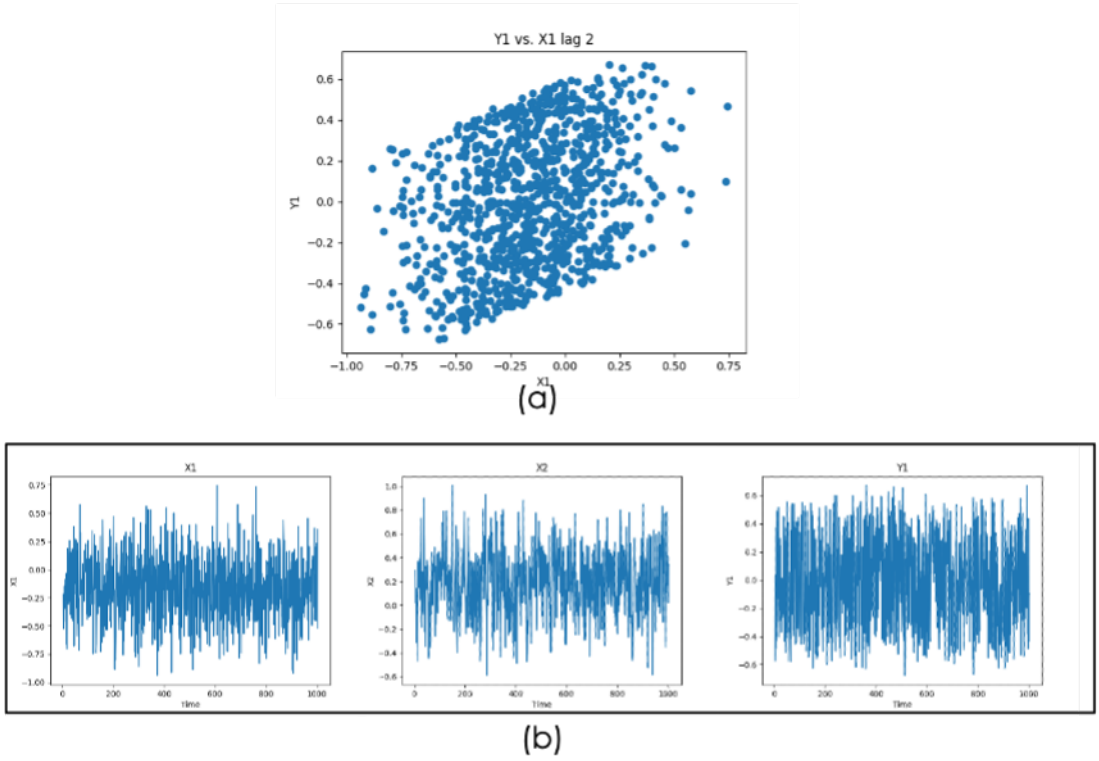

Figura 6. Gráfica de correlación entre Y1 y X1; y gráfica de los valores generados para las tres variables.

Los datos generados fueron almacenados en un archivo csv para ser utilizados en la reconstrucción de la serie de tiempo. Una muestra de estos datos se observa en la Tabla 2.

Tabla 2. Muestra de los datos generados.

X1	X2	Y1
-0.2188687	0.29500207	-0.5712596
-0.5234719	0.23379265	0.03595448
-0.4122958	0.0204329	-0.263006
-0.4568404	0.13210921	-0.4814605
-0.147364	0.13836464	0.37347069
-0.2161871	-0.1489183	-0.4271889
-0.3452627	-0.3147719	0.49109353

Al realizar la reconstrucción de la serie de tiempo se muestran los posibles enlaces causales con base en la probabilidad de que estos sean enlaces verdaderos, especificando la fuerza del enlace y si la correlación es positiva o negativa.

La figura 7 muestra la comparación de la gráfica causal original de los datos sintéticos y la gráfica resultante de la reconstrucción de la serie de tiempo. Como se observa, los enlaces resultantes del análisis del descubrimiento causal

concuerdan con los enlaces causales originados al generar los datos sintéticos. En el mismo sentido, se aprecia el instante de tiempo en el que estos enlaces aparecen y el tipo de relación causal que tienen las variables.

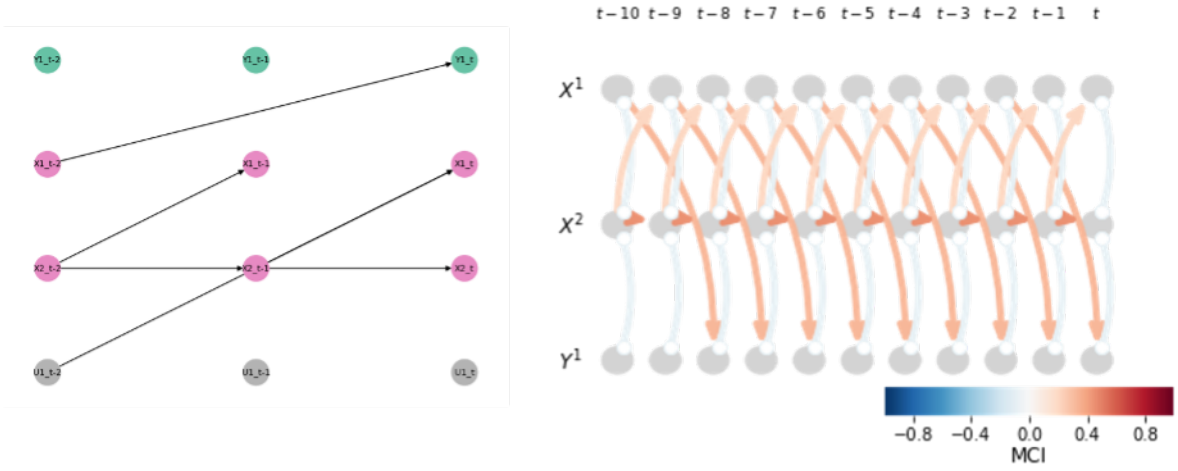

Figura 7. Comparación de la gráfica causal original y la gráfica obtenida mediante la reconstrucción y análisis de la serie de tiempo.

Al analizar la serie de tiempo se puede tener una idea de los enlaces causales que existen entre las variables. Sin embargo, como se puede notar en la Figura 7, los falsos positivos pueden surgir dependiendo de la cantidad de variables o el tiempo que se esté analizando. Existe una muy baja probabilidad de un enlace entre las variables X1 y X2, así como X2 y Y1. Esto puede indicar un falso positivo que debe tomarse en cuenta al momento de analizar la relación causal real de las variables.

Debido a esto, se generaron varios escenarios con el objetivo de analizar el resultado de la reconstrucción de la serie de tiempo, teniendo en cuenta el aumento del retraso, el submuestreo de los datos y los retrasos de tiempo prolongados.

3.1. Aumento del Retraso Máximo

En algunos escenarios, los enlaces causales se pueden presentar en tiempos prolongados, por lo que también se generaron datos sintéticos aumentando el retraso máximo con el que se pueden presentar estos enlaces. De esta forma se puede analizar el descubrimiento causal que puedo tardar más tiempo en presentarse en las variables observadas. Para esto, la Tabla 3 muestra la configuración aumentando el retraso máximo, mientras el retraso mínimo se mantiene con el mismo valor.

Tabla 3. Configuración aumentando el retraso máximo a 3.

Atributo	Valor
Lag mínimo	1
Lag máximo	3
Número de objetivos	1
Número de características	2
Número de variables latentes	1
Varianza del ruido	0.01 – 0.1
Observaciones	1,000

La Figura 8 muestra la gráfica causal resultante del cambio del retraso máximo y la reconstrucción de la serie de tiempo y el análisis causal.

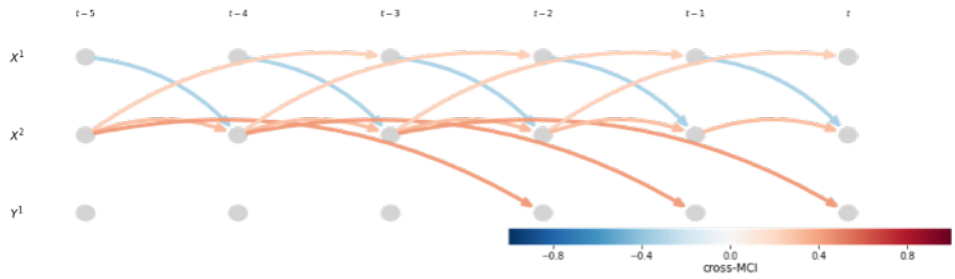

Figura 8. Análisis causal aumentando el retraso máximo a 3.

Se puede apreciar que el análisis causal de la serie de tiempo recuperar los enlaces generados al crear la serie de tiempo. Es decir, se aprecia que el enlace de X2 a Y1 aparece t-3. De igual forma, el enlace causal de X2 a X1 en t-2 se muestra en la gráfica resultante del análisis causal, así como el enlace de X1 a X2 en t-1. La gráfica muestra la fuerza de cada uno de los enlaces. Se comprobó que, aunque se aumentó el retraso máximo, al analizar los datos de la serie de tiempo los enlaces originales se mantuvieron presentes. Esto no ocurre cuando se incrementa nuevamente el retraso máximo a 4, para esto la Tabla 4 muestra la configuración de los atributos para generar la serie de tiempo con un retraso mínimo de 1 y máximo de 4.

Tabla 4. Configuración aumentando el retraso máximo a 4.

Atributo	Valor
Lag mínimo	1
Lag máximo	4
Número de objetivos	1
Número de características	2
Número de variables latentes	1
Varianza del ruido	0.01 – 0.1
Observaciones	1,000

La Figura 9 muestra la gráfica resultante del aumento del retraso máximo e 4 y los enlaces causales generados de forma aleatoria.

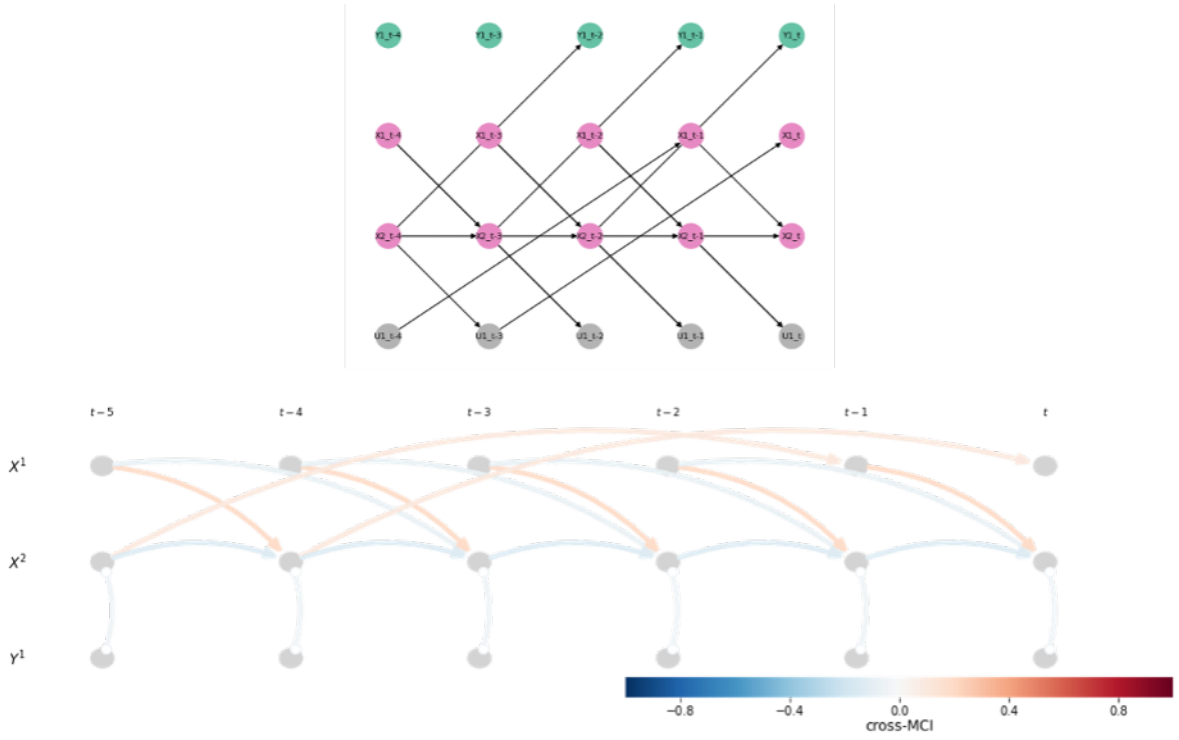

Figura 9. Análisis causal aumentando el retraso máximo a 4.

Como se puede apreciar en la Figura 9, al realizar el análisis causal de la serie de tiempo, cuyo retraso máximo se aumentó a 4, los enlaces causales no corresponden con aquellos que se generaron originalmente. Aunque se detecta la auto correlación de X2 y el enlace de X1 a X2 en t-1, estos se detectan con una intensidad (probabilidad) menor, lo que podría considerarse como falsos positivos. Sin embargo, no se detecta el enlace causal de X2 a Y1. En este sentido, se puede comprobar que el aumento del retraso representa un desafío que debe ser tomado en cuenta al momento de realizar el análisis causal.

3.2. Submuestreo

Como se mencionó previamente, el submuestreo es uno de los principales desafíos en el descubrimiento causal, aún más cuando no se cuenta con el conocimiento del grado de submuestreo con el que cuenta el conjunto de datos de la serie de tiempo. Es por esto, que se generaron escenarios con diversos grados de submuestreo para analizar como afectan las relaciones causales generadas. La configuración utilizada para la generación de la serie de tiempo se muestra en la Tabla 5. En este sentido, los datos generados tuvieron diversos grados de submuestreo para analizar como afectan la gráfica del análisis causal.

Tabla 5. Configuración inicial.

Atributo	Valor
Lag mínimo	1
Lag máximo	2
Número de objetivos	1
Número de características	2
Número de variables latentes	1
Varianza del ruido	0.01 – 0.1
Observaciones	1,000

La Figura 10 presenta la gráfica causal original de la serie de tiempo y la gráfica resultante del análisis causal cada dos retrasos de tiempo. Como se puede observar el submuestreo hace que se pierda cierta información contenida en el conjunto de datos de la serie de tiempo. Por ejemplo, en la gráfica original el enlace causa de X1 a Y1 sucede en t-2, sin embargo, en la gráfica resultante del análisis causal, este mismo enlace sucede en t-1. Asimismo, sucede con la auto correlación de X2. Aunado a esto, no se puede observar el enlace causal de X2 a X, además de la existencia de falsos positivos con una fuerza causal mínima.

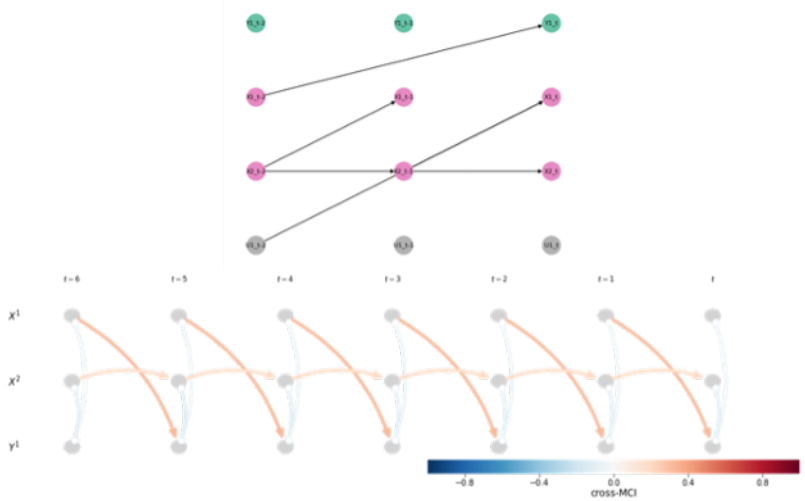

Figura 10. Análisis causal con submuestreo cada dos retrasos.

Asimismo, se incrementó el submuestreo para tener observaciones cada tres retrasos de tiempo, lo que afecta por completo la gráfica del análisis causal. Esto se observa en la Figura 11, donde se puede comparar los enlaces causales originales y el enlace descubierto por el análisis causal.

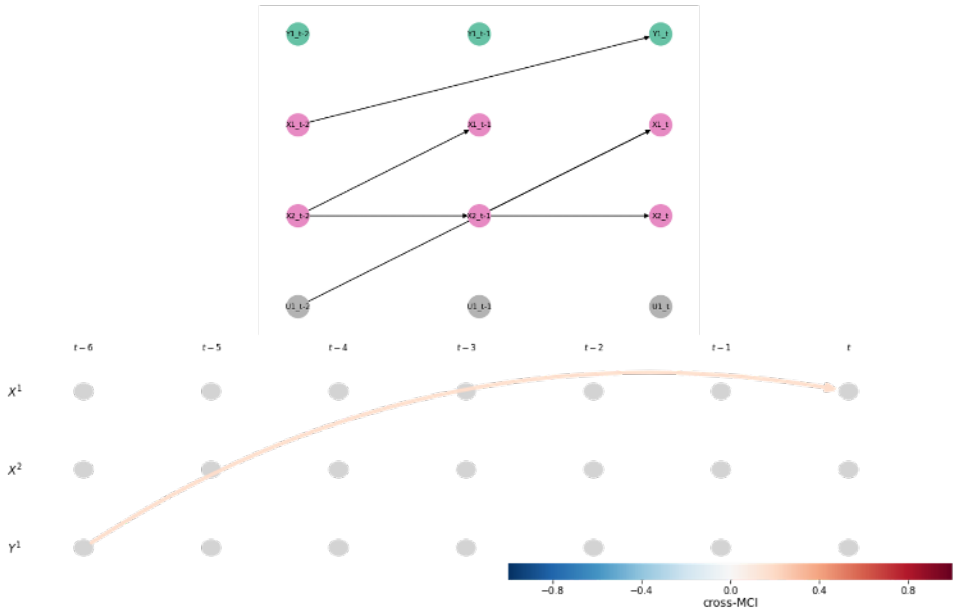

Figura 11. Análisis causal con submuestreo cada tres retrasos.

Como se aprecia, el aumento del submuestreo modifica por completo los enlaces causales originales, teniendo como resultado un único enlace de Y1 a X1, el cual aparece t-6. Esto, sin lugar a duda, demuestra el grado de afectación del descubrimiento causal al no conocer el grado de submuestreo de las observaciones. De esta forma, si se aumenta el submuestreo, por ejemplo, cada 4 retrasos se puede perder toda información causal contenida en el conjunto de datos que conforman la serie de tiempo, esto se aprecia en la Figura 12.

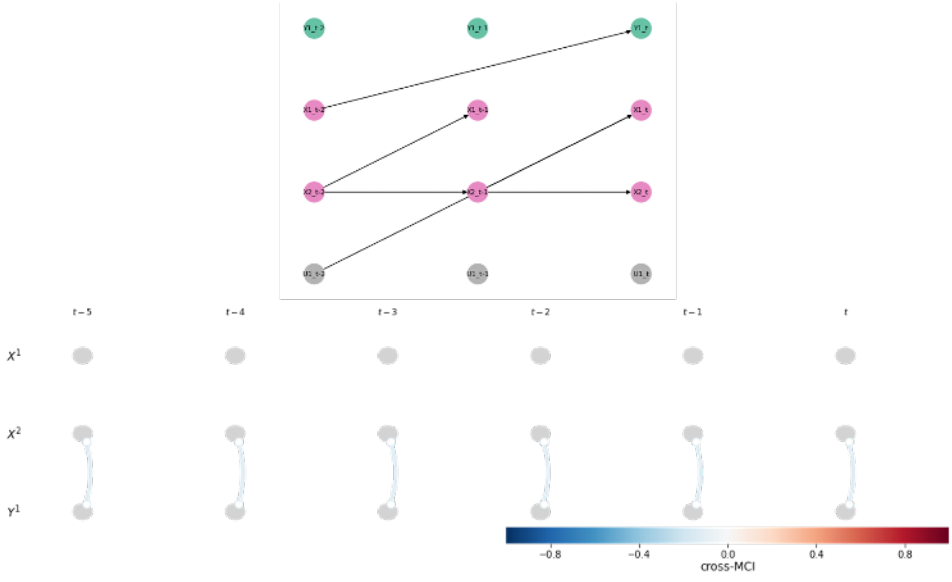

Figura 12. Análisis causal con submuestreo cada cuatro retrasos.

La gráfica causal resultante del análisis no demuestra la información causal de las variables de la serie de tiempo,

lo que puede llevar a conclusiones erróneas. Esto demuestra la importancia de tomar en consideración el submuestreo, al momento de realizar el análisis del descubrimiento causal de la serie de tiempo, ya que es un factor determinante para obtener un análisis completo y correcto de las relaciones causales, así como de su intensidad, entre las variables que conforman la serie de tiempo del evento, fenómeno o sistema de interés.

3.3. Sobremuestreo

Por último, como parte del trabajo inicial en el análisis del descubrimiento causal, se aumentó el retraso mínimo con el que podría aparecer una relación causal en la serie de tiempo. De este modo, se configuraron los atributos para la generación de los datos sintéticos. Esta nueva configuración se muestra en la Tabla 6.

Tabla 6. Configuración de la generación de datos sintéticos aumentando el retraso.

Atributo	Valor
Lag mínimo	2
Lag máximo	4
Número de objetivos	1
Número de características	2
Número de variables latentes	1
Varianza del ruido	0.01 – 0.1
Observaciones	1,000

Si se aumenta el retraso mínimo con el que los enlaces causales pueden surgir, se debe aumentar el tiempo de observación para el análisis causal. De esta forma, el analizar estos escenarios tiene como objetivo verificar si el aumento del tiempo para el análisis modifica el descubrimiento de los enlaces causales. La comparación de la gráfica causal original y la gráfica reconstruida mediante el análisis se muestra en la Figura 13.

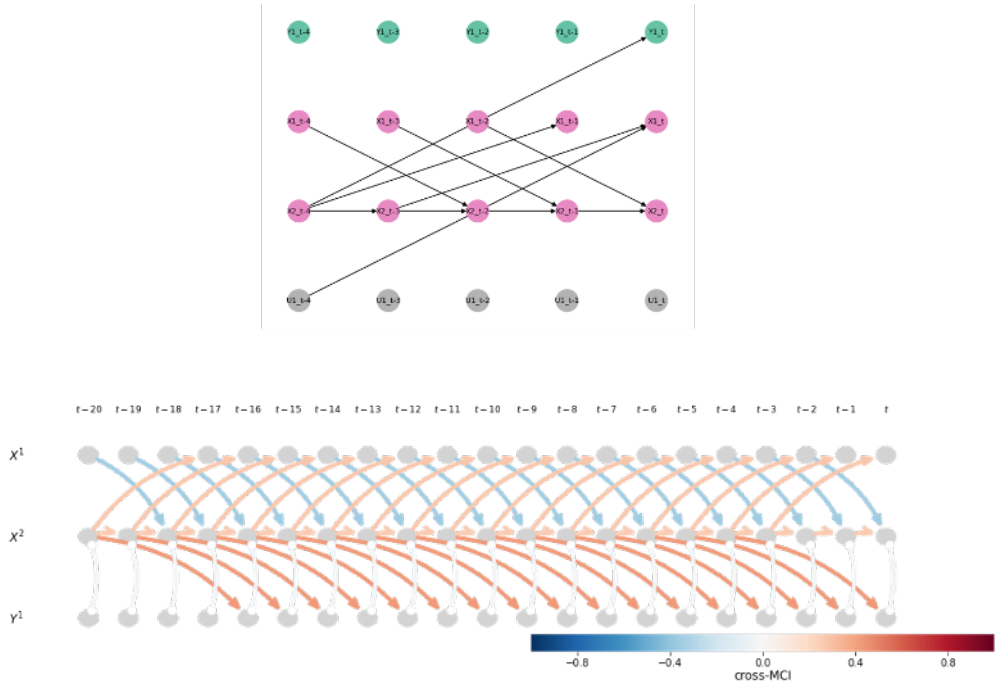

Figura 13. Análisis causal aumentando el retraso en el que pueden aparecer los enlaces causales.

Como se puede apreciar, el análisis de la serie de tiempo para el descubrimiento causal recupera los enlaces originales, especificando el retraso en el que estos aparecen. Aunque es de notar que existe un enlace falso positivo de X2 y Y1, sin embargo, la probabilidad de que este enlace sea un enlace real es bastante baja. La definición de la intensidad de estos enlaces es de utilidad para determinar si estos enlaces se tratan de enlaces reales o de falsos positivos que deben ser eliminados del análisis contextual.

En el mismo sentido, se aumentó nuevamente el retraso mínimo con el que pueden aparecer los enlaces causales (Tabla 7).

Tabla 7. Configuración de la generación de datos sintéticos aumentando el retraso.

Atributo	Valor
Lag mínimo	3
Lag máximo	6
Número de objetivos	1
Número de características	2
Número de variables latentes	1
Varianza del ruido	0.01 – 0.1
Observaciones	1,000

La Figura 14 muestra el resultado del análisis del descubrimiento causal con el aumento en el retraso mínimo y el retraso máximo. Como se observa, al aumentar el retraso mínimo se afecta la reconstrucción de los enlaces causales

de las variables en la serie de tiempo. Esto es, la probabilidad o intensidad con la que se muestran los enlaces es poco. Sin embargo, aun así se puede obtener información sobre las relaciones causales de las variables. Esto puede servir para realizar un análisis focalizado en alguna de las variables o llevar a cabo un análisis profundo configurando los algoritmos de correlación para definir los enlaces causales de la serie de tiempo.

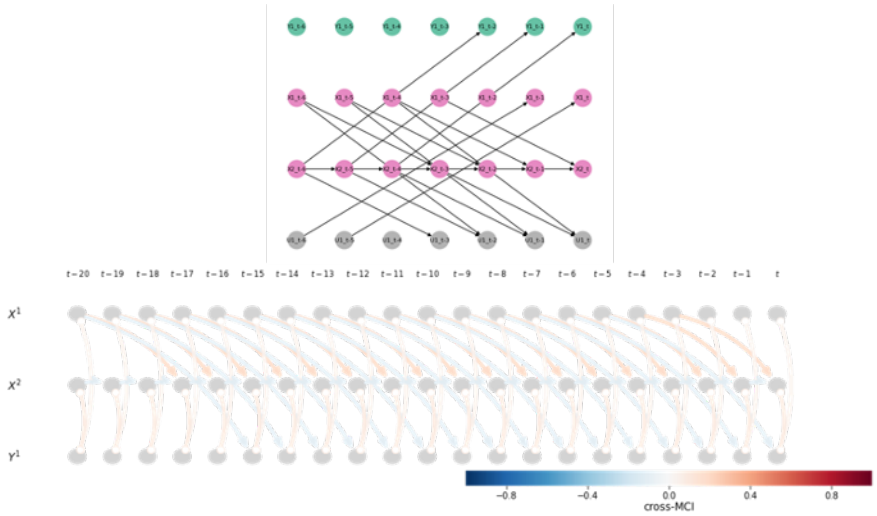

Figura 14. Análisis causal aumentando el retraso mínimo y máximo.

Sin embargo, el análisis se hace más complejo al si se incrementa la distancia entre el retraso mínimo y el máximo. Para esto se cambiaron las configuraciones de la generación de los datos sintéticos. La Tabla 8 muestra esta configuración.

Tabla 8. Configuración para el aumento de la distancia entre el retraso mínimo y el retraso máximo.

Atributo	Valor
Lag mínimo	2
Lag máximo	6
Número de objetivos	1
Número de características	2
Número de variables latentes	1
Varianza del ruido	0.01 – 0.1
Observaciones	1,000

La Figura 15 muestra el resultado del análisis causal de la serie de tiempo. Como se aprecia, al aumentar la diferencia que existe entre el retraso mínimo y el retraso máximo, la definición de los enlaces causales entre las variables se vuelve más complejo. Esto es, algunos de los enlaces reales no son visibles en el análisis del descubrimiento causal, aunque parte de la información original, el enlace de X1 a X2, es recuperada.

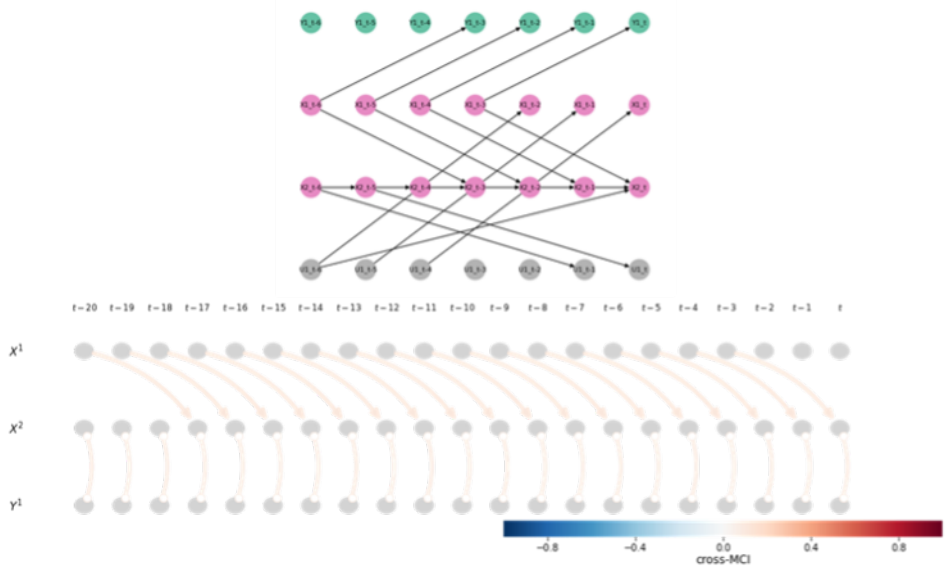

Figura 15. Análisis causal con el aumento de la diferencia entre el retraso mínimo y el retraso máximo.

Por último, se realizó un último escenario, en esta etapa preliminar, para analizar el resultado aumentando aún más el retraso. De esta forma, la Tabla 9 muestra el cambio realizo al retraso mínimo y máximo, siendo estos de 8 y 10 respectivamente.

Tabla 9. Aumento de los atributos en el retraso máximo y mínimo

Atributo	Valor
Lag mínimo	8
Lag máximo	10
Número de objetivos	1
Número de características	2
Número de variables latentes	1
Varianza del ruido	0.01 – 0.1
Observaciones	1,000

Como resultado, Figura 16, la reconstrucción de la gráfica causal de la serie de tiempo muestra los enlaces causales de las variables, los cuales corresponden con los enlaces originales.

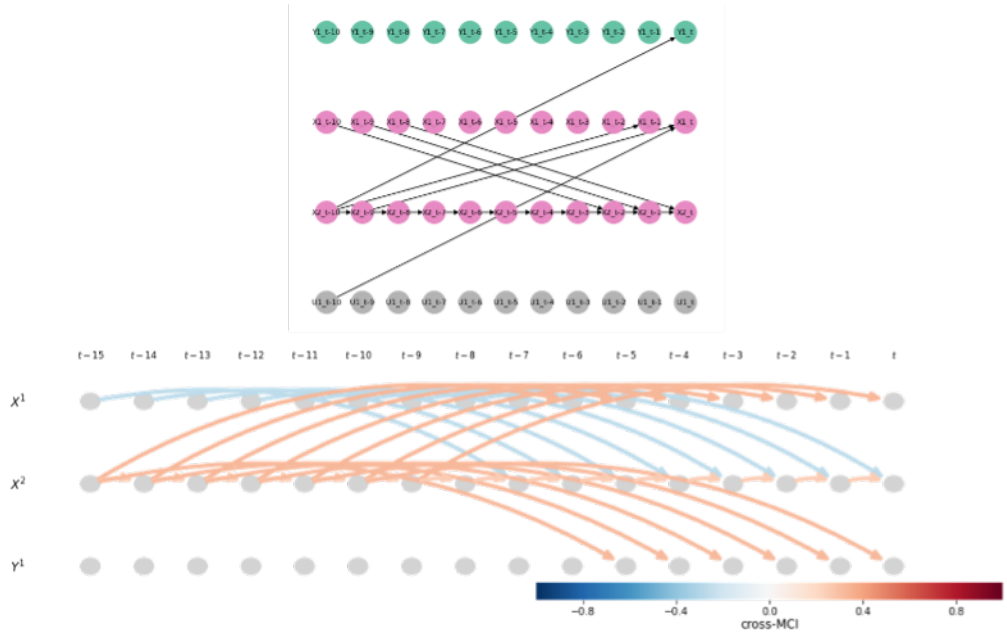

Figura 16. Resultado del análisis causal aumentando el retraso mínimo y máximo.

Debe hacerse notar que, aunque se aumentó el retraso máximo y mínimo, la diferencia es de dos retrasos de tiempo. Como se puede apreciar, la gráfica del resultado del análisis causal se ve mayormente afectada cuando la diferencia entre el retraso mínimo y máximo es mayor (Figura 15). Esto puede ser un campo de oportunidad, en la presente investigación, para generar experimentos y verificar el grado de afectación del descubrimiento causal bajo estas circunstancias.

4. Conclusiones

Se llevó a cabo un análisis preliminar del estado del arte sobre el descubrimiento causal en series de tiempo, detectando los principales desafíos que presenta el análisis causal de variables con respecto al tiempo. En este sentido, se realizaron las primeras pruebas y configuraciones de dos módulos programados en Python para el análisis causal en series de tiempo. Siendo el primero un módulo para la generación de datos sintéticos y el segundo un módulo para analizar los enlaces causales de variables para la reconstrucción de la gráfica causal.

Como resultados preliminares se realizaron las configuraciones para generar datos sintéticos bajo diferentes escenarios, teniendo como principales aspectos el submuestreo y el aumento del retraso en que pueden surgir los enlaces causales. Esto significó una gran ventaja ya que se contó con una forma de comparar las relaciones causales originales de las variables y aquellas detectadas mediante el análisis causal. De esta forma, se pudo entender como el aumento del retraso y el submuestreo de la serie de tiempo afectan la detección de causalidad de los datos observados. Siendo este último uno de los desafíos que deben de tomarse en consideración ya que si no se cuenta con información sobre el grado de submuestreo con el que cuenta la serie de tiempo se puede perder información causal relevante y concluir en análisis erróneos sobre las relaciones entre variables.

Asimismo, contar con un módulo para el análisis de la causalidad y la intensidad de los enlaces sirve como punto de partida para generar modificaciones para un mejor análisis de las relaciones causales en las series de tiempo de acuerdo con el contexto del área de investigación. En este sentido, se propone continuar con el análisis de estas aportaciones para detectar un área de oportunidad de la presente investigación.

5. Trabajo Futuro

Derivado de estos primeros experimentos se propone seguir evaluando el desempeño de estas propuestas para el análisis de series de tiempo en el ámbito del descubrimiento causal, teniendo especial atención al desafío presentado por el submuestreo de los datos. Cabe señalar que uno de los principales objetivos de la presente investigación es realizar una propuesta para la aportación al estado del arte, por lo que se sugiere continuar con la revisión de trabajos relacionados e incluso realizar experimentos con otras metodologías y algoritmos de descubrimiento causal.

Como trabajo inmediato se propone extender y profundizar la revisión del estado del arte y los trabajos relacionados para trabajar en una primera propuesta de contenido para un artículo científico.

Asimismo, se continuarán realizando pruebas de configuraciones para mejorar el desempeño del análisis y reconstrucción de la serie de tiempo y comparar los resultados obtenidos.

Referencias

Danks, D., y Plis, S. (2014). Learning causal structure from undersampled time series. *JMLR: Workshop and Conference Proceedings*.

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

Hyttinen, A., Plis, S., Järvisalo, M., Eberhardt, F., y Danks, D. (2017). A constraint optimization approach to causal discovery from subsampled time series data. *International Journal of Approximate Reasoning*, *90*, 208–225.

Lawrence, A., Kaiser, M., Sampaio, R., y Sipos, M. (2020). Data generating process to evaluate causal discovery techniques for time series data. *Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems*.

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

Anexo I

```
import matplotlib.pyplot as plt

from data_generation_configs import (
    CausalGraphConfig, DataGenerationConfig, FunctionConfig, NoiseConfig, RuntimeConfig
)
from time_series_generator import TimeSeriesGenerator


if __name__ == '__main__':

    # Set general attributes.
    complexity = 20

    # Set attributes for causal graph. More are set directly in the configuration below.
    min_lag = 1
    max_lag = 2
    num_targets = 1
    num_features = 2
    num_latent = 1

    # Set attributes for data generation.
    num_samples = 1000

    # complexity is only used to initialise any unprovided configs. Here they are all initialised so it is ignored.
    config = DataGenerationConfig(random_seed=1, complexity=complexity, percent_missing=0.0,
                                  causal_graph_config=CausalGraphConfig(
                                      graph_complexity=complexity,
                                      include_noise=True,
                                      max_lag=max_lag,
                                      min_lag=min_lag,
                                      num_targets=num_targets,
                                      num_features=num_features,
                                      num_latent=num_latent,
                                      prob_edge=0.3,
                                      max_parents_per_variable=1,
                                      max_target_parents=2, max_target_children=0,
                                      max_feature_parents=3, max_feature_children=2,
                                      max_latent_parents=2, max_latent_children=2,
                                      allow_latent_direct_target_cause=False,
                                      allow_target_direct_target_cause=False,
                                      prob_target_autoregressive=0.1,
                                      prob_feature_autoregressive=0.5,
                                      prob_latent_autoregressive=0.2,
                                      prob_noise_autoregressive=0.0,
                                  ),
                                  function_config=FunctionConfig(
                                      function_complexity=complexity
                                  ),
                                  noise_config=NoiseConfig(
                                      noise_complexity=complexity,
                                      noise_variance=[0.01, 0.1]
                                  ),
                                  runtime_config=RuntimeConfig(
```

```
                num_samples=num_samples, data_generating_seed=42
            )
            )

# Instantiate a time series generator.
ts_generator = TimeSeriesGenerator(config=config)

# Query for the completed config now that causal graph and SCM have been created.
full_config_dict = ts_generator.get_full_config()

# Generate data sets from this configuration.
data = ts_generator.generate_datasets()
datasets, causal_graph = data
df = datasets[0]  # We only generated one data set so just look at the first (and only) DataFrame.

# View data.
for node in df.columns:
    plt.figure()
    plt.plot(df.index, df[node])
    plt.xlabel('Time')
    plt.ylabel(node)
    plt.title(node)

# Compare target against its parents.
for parent in causal_graph.get_parents('Y1_t'):
    if causal_graph.is_feature_node(parent):
        plt.figure()
        var, lag = causal_graph.get_var_and_lag(parent)
        if lag > 0:
            plt.scatter(df[var][:-lag], df['Y1'][lag:])
            plt.title(f'Y1 vs. {var} lag {lag}')
        else:
            plt.scatter(df[var], df['Y1'])
            plt.title(f'Y1 vs. {var}')
        plt.xlabel(var)
        plt.ylabel('Y1')

# View causal graph.
causal_graph.display_graph(include_noise=False)  # Set to True to graph noise nodes.

# Show plots.
plt.show()
```