Workshop on Causal Discovery CaDis 2025

This book contains the proceedings of the Third Workshop on Causal Discovery (CaDis 2025). The workshop took place in Xalapa, Veracruz, on December 8th - 9th, 2025. It was made possible thanks to the collaboration of the National Institute of Astrophysics, Optics, and Electronics of Mexico (INAOE) and the Instituto de Investigaciones en Inteligencia Artificial (IIA) of Universidad Veracruzana, with the support of IBERAMIA and the Mexican Academy of Computing (AMexComp).

Causal discovery offers a way for understanding the causal relations among data as well as the effects of interventions, modeling hypothetical scenarios, and performing counterfactual inferences from data. Controlled experiments are often unfeasible, difficult to replicate, or too expensive to perform. For this reason, the scientific community seeks innovative ways to learn causal models from observational data in order to analyze the hidden causal relations among variables of interest. This challenge has led to advances that combine causality with other areas such as deep learning and reinforcement learning, generating applications across multiple domains

This edition of the CaDis proceedings includes the eight papers presented at the workshop, that cover contributions from fundamentals and algorithms for causal discovery, to practical applications; as well as the abstract of keynote talk by the professor Sisi Ma (University of Minnesota).

Video recordings of these talks are available at the workshop website: https:// cadisworkshop.com.mx/program/.

This volume not only documents recent advances, but also aims to inspire new research and collaborations, especially in the Iberoamerican region, encouraging the growth of a global community interested in advances in causal reasoning and discovery.

INAOE
iberamia
Sociedad Iberoamericana de Inteligencia Artificial
amexcomp

Workshop on Causal Discovery CaDis 2025 | IBERAMIA

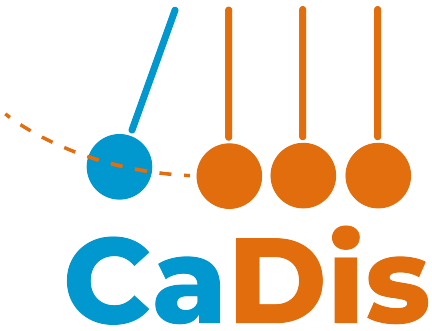
Workshop on Causal Discovery

CaDis 2025

Luis Enrique Sucar Succar
Julio César Muñoz Benítez
Marcela Quiroz Castellanos
Editors

CaDis

Workshop on Causal Discovery
CaDis 2025



Ibero-American Society of Artificial Intelligence

Proceedings of the 3rd Workshop on Causal Discovery CaDis 2025
Editors: Luis Enrique Sucar, Julio César Muñoz–Benítez and Marcela Quiroz Castellanos

In collaboration with Academia Mexicana de Computación and IBERAMIA.

First Edition 2025.
Sociedad Iberoamericana de Inteliencia Artificial

ISBN: 978-84-09-85175-1

Style correction: Luis Enrique Sucar.
Cover design: Instituto Nacional de Astrofíscia, Óptica y Electrónica.
Editing: Luis Enrique Sucar.

Printed in Mexico.

Proceedings of the 3rd Workshop on Causal Discovery
CaDis 2025

Workshop Chairs:

Luis Enrique Sucar, INAOE, Mexico
Julio César Muñoz-Benítez, INAOE, Mexico
Marcela Quiroz Castellanos, UV, Mexico

Program Committee:

Adnan Darwiche, University of California, Los Angeles, USA
Nicandro Cruz, UV, Mexico Hugo Jair Escalante, INAOE, Mexico
Mauricio González, ITESM, Mexico
Samuel Montero, University of Birmingham, UK
Eduardo Morales, INAOE, Mexico
Julio César Muñoz-Benítez, INAOE, Mexico
Billy Peralta, CENIA, Chile
Rubén Sánchez-Romero, Rutgers University-Newark, USA
Luis Enrique Sucar, INAOE, Mexico

Foreword

This volume contains the proceedings of the 3rd Workshop on Causal Discovery (CaDis 2025). The workshop was held at the Instituto de Investigaciones en Inteligencia Artificial (IIA) de la Universidad Veracruzana, in Xalapa, Veracruz, Mexico. This edition was made possible through the collaboration of the National Institute of Astrophysics, Optics, and Electronics of Mexico (INAOE) and the Instituto de Investigaciones en Inteligencia Artificial. We acknowledge the support the Mexican Academy of Computing (AMexComp) and IBERAMIA.

Causal models have many advantages, including the ability to reason about the effects of interventions, as well as the results of different scenarios or counterfactuals. The traditional approach for building causal models is by conducting experiments, however these are often infeasible, unethical or too expensive. Recently there has been a lot of interest in the scientific community to learn causal models from observational data, but this is a great challenge, as just from observations is not possible, in general, to define a unique causal model.

The objective of this workshop was to present recent advances in causal discovery, including different approaches that consider observational and/or interventional data, and also building models with the help of human experts. It is also of interest the combination of causal discovery with other areas of machine learning, such as reinforcement learning and deep learning; as well as real-world applications.

The CaDis 2025 program included an invited talk by Prof. Sisi Ma (Assistant Professor of Medicine, Division of General Internal Medicine, University of Minnesota) and a discussion panel on the current challenges of causal discovery. Video recordings of these talks are available at the workshop website: `https://cadisworkshop.com.mx/program/`.

After a review by at least three members of the program committee, eight papers were accepted for publication in these proceedings. In an analogous way as the workshop, this proceedings are divided in three parts: (i) Invited Talks abstracts, (ii) Fundamentals and Algorithms for Causal Discover, and (iii) Applications.

We hope that this workshop will help to increase the interest of the academic community in causal reasoning and discovery, in particular in the Iberoamerican region; and help to foster new collaborations between different research groups.

Luis Enrique Sucar, Julio César Muñoz-Benitez and Marcela Quiroz Castellanos.
Workshop Chairs

Contents

Section 1

Invited Talk

Multiple Markov Boundaries: the Good, the Bad, and the Ugly

Author: Prof. Sisi Ma, University of Minnesota

Abstract: The Markov boundary of a response variable T is the minimal set of variables that renders all other variables in the dataset statistically independent of T. While some distributions admit a unique Markov boundary, others contain multiple distinct Markov boundaries for the same response. In this talk, I will introduce the theory of multiple Markov boundaries and explore potential mechanisms underlying their emergence in data, with a particular emphasis on biomedical data. I will also discuss the implications of multiple Markov boundaries for predictive modeling, causal modeling, and model translation into real-world decision support tools.

Video recording of the invited talk is available at the workshop website: https://cadisworkshop.com.mx/program/

Section 2

Fundamentals and Algorithms for Causal Discovery

Causal Interpretation of DBSCAN: Dynamic Modeling for Epsilon Estimation

Kay Garcia-Sanchez[0009−0009−2169−1991], Jorge-Luis Perez-Ramos[0000−0002−0444−9230], Selene Ramirez-Rosales[0000−0001−6635−5427], Luis-Antonio Diaz-Jimenez[0000−0003−1519−105X], Ana-Marcela Herrera-Navarro[0000−0001−7711−9585], Hugo Jimenez-Hernandez[0000−0003−0827−6645], and Daniel Canton-Enriquez[0000−0002−6543−5078]

Centro de Investigación e Innovación en Ciencias de la Computación y Tecnología Educativa, Facultad de Informática, Universidad Autónoma de Querétaro, Avenida de las Ciencias S/N, 76230, México
kgarcia85@alumnos.uaq.mx, daniel.canton@uaq.mx

Abstract. One of the most widely used tools for detecting structures in unsupervised data is the DBSCAN algorithm. However, its performance critically depends on the proper selection of the neighborhood parameter ε. This work offers a physical-mathematical perspective that treats this parameter as a controllable variable within a dynamic system composed of a spring, a damper, and a mass. The variation of ε is modeled through ordinary differential equations, in which jerk and accelerations represent structural changes in the density space. The proposal includes a numerical simulation using the fourth-order Runge-Kutta method and a smoothing procedure via the Savitzky-Golay filter, which enables the determination of the critical point that optimizes cluster stability. The structural causal model, compared to DBSCAN Kneedle, optimizes clustering metrics across two synthetic datasets and the Covtype dataset while maintaining an overall complexity of $O(nlog(n))$. The suggested method not only improves the estimation of ε but also provides a causal, mechanical explanation of the clustering process, opening new perspectives for interpretable unsupervised learning.

Keywords: Density-based clustering · Interpretable machine learning · Causal DBSCAN · Causal Model · Parameter optimization

1 Introduction

In the field of unsupervised learning, clustering algorithms seek to identify hidden structures in data; in this way, regions with similar characteristics and attributes are identified without labeled data [1]. Among the most influential techniques, the Density-Based Spatial Clustering of Applications with Noise (DBScan) algorithm stands out for its ability to detect dense regions and distinguish noisy or isolated elements. This helps overcome the classic limitations of methods such as

K-means [2] and hierarchical clustering [3], which assume structures determined by the data's inherent geometry.

Despite its advantages, DBSCAN presents a constant challenge: its sensitivity to the hyperparameters ε (neighborhood radius) and minPts (minimum number of points in the neighborhood). While minPts tends to be stable, the ε parameter is a critical variable that determines the data's structural connectivity. In this context, the variation of the ε parameter determines when clusters emerge, merge, or disappear, leading to significantly different cluster formations [4, 5]. This has motivated numerous studies to develop automated methods for ε selection, ranging from heuristic techniques such as the k-distance plot [6] to hierarchical or adaptive approaches [7]. However, most of these studies remain in the empirical domain, without formally explaining why a local modification in ε causes a global change in the topology of the density space [6, 8].

Consequently, there are researchers seeking to combine unsupervised learning with causal discovery. Historically, causal discovery methods, such as the PC algorithm [9] or models based on conditional response [10, 11] are used to analyze experimental data in order to identify directed acyclic graphs. However, recent works are beginning to apply these principles to the clustering domain. These approaches enhance the ability of unsupervised learning to detect hidden mechanisms of causal variability, offering a more robust theoretical foundation for interpreting structures in complex data [12]

In this work, an alternative approach is proposed where ε is interpreted not as an adjustment parameter but as a causal variable within a dynamic system that governs the topology of the clusters. From this point of view, clustering is understood as an interaction between unobservable structural forces, in which small local perturbations (variations in ε) induce changes in the dataset's topology.

To consolidate this concept, a physical-mathematical model using a mass-spring-damper system is presented. Here, the k-distance curve is interpreted as a dynamic signal, with its acceleration and underlying energy dynamics representing the structural change in density. Simulating the system's behavior under manipulation in ε allows identification of the critical transition point where the curve's slope changes abruptly. This was carried out by solving ordinary differential equations (ODEs) using the fourth-order Runge-Kutta (RK4) method. This point is understood as the optimal value of ε, which represents the causal boundary between dense and sparse areas.

Thus, the current work goes beyond a traditional heuristic, proposing a Structural Causal Model (SCM) to calculate parameters in density algorithms, combining the physical essence of dynamic systems with the causal semantics of interventions. The result is an interpretive framework that not only enables a robust choice of the ε parameter but also explains how the algorithm works, thereby favoring its application in high-dimensional data scenarios.

The experimental validation of the proposal is performed across three scenarios: two synthetic datasets with 10^6 samples and one real-world multivariate dataset. For each scenario, the experimental process was conducted by applying the techniques of DBSCAN as an SCM; and the *Kneedle* heuristic procedure,

comparing the obtained results. Furthermore, clustering quality metrics, such as Davies–Bouldin, Silhouette, and Calinski–Harabasz, were employed.
The remainder of the article is structured as follows: Section 2 presents the theoretical foundations supporting the proposal; Section 3 covers computational complexity; Section 4 presents the experimental results and discussion; finally, Section 5 summarizes the conclusions and proposes future work.

2 Materials and Methods

2.1 Conceptual Foundations

The proposal stems from the idea that the ε parameter should be reinterpreted as a causal control variable in a dynamic system. It is assumed that ε functions as an intervention mechanism that changes the local connectivity of the points and, consequently, the global structure of the clusters, rather than being seen as a mere adjustable hyperparameter.
From this perspective, the data density behaves like a physical system in which modifications to ε generate a dynamic response that can be measured and is similar to the accelerations and energy changes seen in mass-spring-damper models. Formally, the model aims to locate the critical transition point (the elbow point) in the k-distance curve, where a slight local variation of ε generates a global change in the system's density. This point corresponds to the optimal value of ε that maximizes the clustering's structural stability.

2.2 *K*-distances curve extraction

The K-distance curve characterizes the local density distribution of the data points by quantifying the distance from each data point to its k-th nearest neighbor [8]. In this context, the K-distance plot is obtained by calculating the k nearest neighbors of each point using Euclidean distance and the Nearest Neighbors algorithm.
First, Euclidean distance is defined as a metric that estimates the distance between two points within a system in R^2 [13], represented in (1):

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

Secondly, the Nearest Neighbors algorithm can be defined: given a finite dataset S in a Euclidean space E^d and a query q, Nearest Neighbors obtains the k nearest neighbors R of q by evaluating $\delta(x, q)$ where $x \in S$ [14]. R is described according to(2).

$$R = \underset{R \subset S, |R| = k_{\min}}{\arg\min} \sum_{x \in R} \delta(x, q) \tag{2}$$

Analogously, in clustering, neighborhoods define which points causally influence the local density and the subsequent formation of clusters.

2.3 Curve Smoothing

One of the most important prerequisites for working with signals is a clean input signal. When working with a k-distance curve, the signal is implicitly noisy, which makes numerical analysis difficult. In this case, a Savitzky-Golay filter was used; it is a smoothing method applied to signals to preserve local features, such as peaks and curvatures, while reducing the effect of noise [15]. This technique is represented by (3):

$$y_i^{(m)} = \frac{m!}{\Delta t^m} \sum_{j=-n}^{n} c_j^{(m)} x_{i+j} \tag{3}$$

where $y_i^{(m)}$ is the m-th smoothed derivative at point i, x_{i+j} are the signal values, $c_j^{(m)}$ are the Savitzky-Golay filter coefficients obtained by fitting a polynomial of degree p over a window of $2n+1$ points, Δt is the time step of the samples, and m is the order of the derivative to be calculated.

2.4 Mass-Spring-Damper System

A mass-spring-damper system is a dynamic system consisting of a mass m attached to a spring with a stiffness constant k and a damper with coefficient b, subjected to a force F [16]. First, the stiffness constant k is a magnitude that quantifies how rigid a spring is; it relates the force applied to the spring with the displacement produced by it, following Hooke's Law [17]. This law is detailed in (4)

$$F = kx \tag{4}$$

where F is the force exerted, k is the stiffness constant whose unit is N/m, and x is the displacement generated by the spring. Secondly, the damping coefficient b is a parameter that quantifies a system's ability to dissipate kinetic energy through frictional forces [18]; its value depends on the restrictive medium used. Figure 1 depicts a classic model of the mass-spring-damper system.

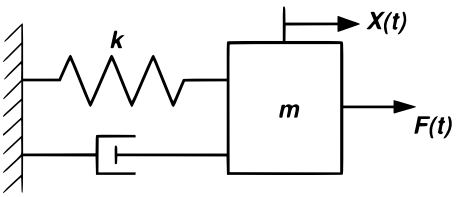


Fig. 1: Representation of a mass-spring-damper dynamic system. The interaction between the mass (m), the spring stiffness constant (k), the damping coefficient (X(t)), and the applied force (F) is illustrated. [19]

Mathematically, the dynamics of the system are described by the second-order differential equation represented in (5):

$$m\ddot{x}(t) + b\dot{x}(t) + kx(t) = F(t) \tag{5}$$

In this context, $F(t)$ is considered a causal intervention; that is, a change in ε leads to a perceptible alteration in the system's acceleration and, consequently, in the cluster configuration. To solve the differential equation, the fourth-order Runge-Kutta method was used. It is a numerical method for solving differential equations. It attempts to find a numerical solution by approximating values x_j at a finite set of points $t_j = t_0 + jh$, with $j = \{1, 2, ..., m\}$ [20].The discrete numerical integration process is expressed in (6):

$$\begin{aligned} \boldsymbol{k}_1 &= h\,\mathbf{f}(t_n, \mathbf{x}_n), \\ \boldsymbol{k}_2 &= h\,\mathbf{f}\left(t_n + \tfrac{h}{2},\, \mathbf{x}_n + \tfrac{\boldsymbol{k}_1}{2}\right), \\ \boldsymbol{k}_3 &= h\,\mathbf{f}\left(t_n + \tfrac{h}{2},\, \mathbf{x}_n + \tfrac{\boldsymbol{k}_2}{2}\right), \\ \boldsymbol{k}_4 &= h\,\mathbf{f}\left(t_n + h,\, \mathbf{x}_n + \boldsymbol{k}_3\right), \\ \mathbf{x}_{n+1} &= \mathbf{x}_n + \tfrac{1}{6}\left(\boldsymbol{k}_1 + 2\boldsymbol{k}_2 + 2\boldsymbol{k}_3 + \boldsymbol{k}_4\right) \end{aligned} \tag{6}$$

This system is an example of a causal dynamic model in which each of the variables F, x, $\dot{x}$, and $\ddot{x}$ has a direct, predictable effect on the others. Thus, simulating the system allows exploring how changes in the force (or in the system parameters) cause measurable changes in the mass's trajectory; in terms of causality, manipulating one variable causes an observable effect in other variables of interest. Therefore, the mass's displacements depend on external forces and system properties, so the cause-and-effect relationship between the variables is explicit and quantifiable.

2.5 System Parameter Identification

The proposed algorithm should be understood within the field of system identification, a field of control engineering that seeks to obtain mathematical models of dynamic systems from measurements taken in the process, such as inputs or control variables, outputs or controlled variables, and disturbances [21]. In this project, the identification process follows a series of structured steps to estimate the physical parameters that make up the system, which arise directly from the data.

The k-distance curve must be treated as a discrete signal $x(t)$ in a pseudo-time domain t, where t corresponds to the indices i for each sorted point. From this position signal $x(t)$, we calculate the velocity $\dot{x}(t)$ (first derivative) and the acceleration $\ddot{x}(t)$. These time series $x_{obs}(t)$, $\dot{x}_{obs}(t)$, and $\ddot{x}_{obs}(t)$ represent the observed dynamics of the data density.

On the other hand, the objective of system identification is to find the optimal values of m (mass), b (damping), and k (stiffness) described in Equation 5. To

achieve this, a least-squares optimization is performed; by rearranging (5), the acceleration (the term to be predicted) can be isolated (see (7)).

$$\ddot{x}_{pred} = \frac{1}{m}(F(t) - b\dot{x} - kx(t)) \tag{7}$$

The objective function to be minimized is the mean squared error between the observed acceleration ($\ddot{x}_{obs}(t)$) and the predicted acceleration $\ddot{x}_{pred}(t)$. The process is summarized in (8):

$$f(\theta) = \sum(\ddot{x}_{obs}(t) - \ddot{x}_{obs}(t))^2 \tag{8}$$

For the force $F(t)$, we assume a simple Newtonian force $F(t) = C$ that acts as an initial excitation to set the system in motion.
Finally, a parameter estimation is performed to find the parameter vector $\theta = [m, b, k]$. This vector is detailed in (9):

$$\boldsymbol{\theta} = \arg\min_{\theta} f(\theta) \tag{9}$$

More specifically, the cost function $f(\theta)$ is understood as a function $\mathbf{R} \to \mathbf{R}$ that measures how costly a set of parameters is. To perform this type of optimization, a quasi-Newtonian method called Broyden-Fletcher-Goldfarb-Shanno (BFGS) was used. This method attempts to update an approximation to the inverse Hessian matrix. This update follows (10):

$$H_{k+1} = H_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{H_k s_k s_k^T H_k}{s_k^T H_k s_k} \tag{10}$$

where s_k is $\theta_{k+1} - \theta_k$, and y_k is $\nabla f(\theta_{k+1}) - \nabla f(\theta_k)$. On the other hand, the update step is described in (11):

$$\theta_{k+1} = \theta_k - \alpha_k H_k^{-1} \nabla f(\theta_k) \tag{11}$$

In this way, we understand that the parameters m, b, and k are estimated, and that the model itself selects them based on the k-distance curve; this assertion is corroborated in Section 4.

2.6 Elbow Point Estimation

The elbow point of a function $f(x)$ is a random and spontaneous event on the curve that marks the end of one state and the beginning of another [6]. In DBSCAN, the elbow point represents the change from a dense region to a sparse region (noise); the value of this point on the y-axis represents the value of the neighborhood radius ε.

In the proposed system, assuming a particle moves along the k-distance curve under the action of stiffness and damping forces, the calculation of the jerk of the particle in each iteration is of vital importance, whose formula is represented in (12):

$$J(t) = \frac{d\ddot{x}(t)}{dt} \tag{12}$$

The location where $J(t)$ reaches its local maximum indicates a sharp change in the system's acceleration, which in turn indicates the transition between dense and sparse areas. The ε value related to this point is therefore considered optimal.

2.7 Structural Causal Model

This section presents the SCM, which integrates the direction of dependencies, the functional mechanisms, and the interventions required to satisfy the criteria established in Pearl's causal theory [10].

The total process unfolds through a set of internal variables that operate in a hierarchy:

- U: exogenous variables associated with the intrinsic variability and measurement noise of the dataset.
- X_{knn}: k-NN distances and local density features after standardization.
- S_{sg}: smoothed signal achieved through the Savitzky-Golay filter.
- D_{rk4}: dynamic trajectory generated from the equation of motion, which has been integrated using the fourth-order Runge-Kutta method.
- J: the jerk is the result of the third numerical derivative of the path.
- ε^*: the final value taken by the ε parameter used in DBSCAN.

These variables constitute a unidirectional causal flow, which can be represented by a Directed Acyclic Graph (DAG): the dataset alters the smoothing, which generates the dynamics, and this, in turn, establishes the internal indicator that determines ε. The causal DAG summarizing these dependencies and formalizing the mechanisms that structure the SCM is shown in Figure 2.

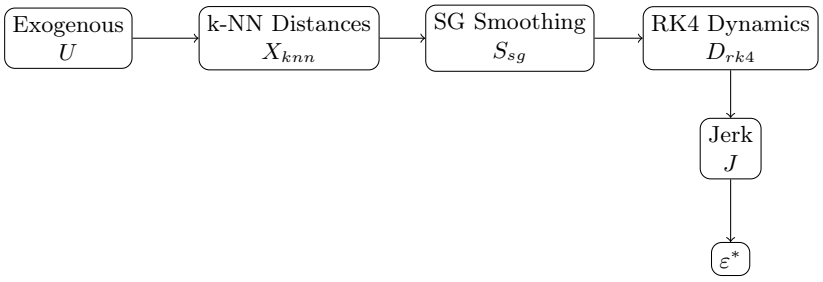


Fig. 2: Internal causal structure of the SCM for the estimation of ε.

The causal relationships specified in the DAG are formalized through the relevant structural equations in (see (13)):

$$\begin{aligned} X_{knn} &= f_1(data, U_1), \\ S_{sg} &= f_2(X_{knn}; h, p), \\ D_{rk4} &= f_3(S_{sg}; \theta), \\ J &= f_4(D_{rk4}), \\ \varepsilon^* &= f_5(J) \end{aligned}$$

where h and p are the smoothing parameters, while θ groups the parameters of the induced dynamic system (external force, spring constant, and damping). The SCM formulation maintains that each node depends only on its causal parents, upholding the factorization imposed by the DAG.

After defining the SCM's structural equations, the manner in which internal system alterations are expressed under controlled interventions must be determined. For this, the $do(\cdot)$ notation suggested by Pearl is used, where an intervention directly replaces the structural equation of a node and enables the study of how this change influences the distribution of the output variable. In this proposal, the parameter selection is expressed as $\varepsilon^* = f_5(J)$, such that interventions on any of the internal nodes generate observable changes in the final value of ε^*. Formally, these interventions take the form: $do(J = j_0)$, $do(S_{sg} = s_0)$, $do(\theta = \theta_0)$, and allow obtaining distributions of the type: $P(\varepsilon^*|do(J = j_0))$.

Finally, the causal validity of the model is examined through internal interventions that allow verifying the consistency of the causal sequence. The manipulations taken into account are the following: (i) adjusting the window size via $do(h = h_0)$; (ii) modifying the system stiffness by means of $do(\theta = \theta_0)$; and (iii) establishing artificial trajectories via $do(D_{rk4} = d_0)$. In each case, the sequence was noted (see (13)):

$$\Delta S_{sg} \Rightarrow \Delta D_{rk4} \Rightarrow \Delta(J) \Rightarrow \Delta\varepsilon^* \tag{13}$$

which confirms that the parameter calculation is causally linked with the transformations generated in the SCM's internal nodes. This determines that ε^* is not the result of local correlations, but rather the consequence of an explicit causal mechanism that organizes the SCM's dynamics.

2.8 Evaluation Metrics

To provide an objective description of the quality of the generated clusters, three metric categories were employed.
In the first place, it is evaluated if the data tends to cluster. For this, the Hopkins statistic is applied as a quantitative tool to determine whether a dataset possesses a natural clusterable structure [22]. In this context, the hypothesis shown in (14) is applied:

$$\begin{cases} H_0 : H = 0.5 & \text{(Random Distributed Data)} \\ H_1 : H < 0.5 & \text{(Tend to be clustered Data)} \end{cases} \tag{14}$$

In this context, the value of the Hopkins statistic is obtained from (15):

$$H = \frac{\sum_{i=1}^{n} \omega_i}{\sum_{i=1}^{n} u_i + \sum_{i=1}^{n} \omega_i} \tag{15}$$

where ω_i are the n sampled real points and u_i are the n random points.

Next, cohesion metrics are evaluated; these assess the internal compactness of the formed clusters. In this context, we speak of inter-cluster variance ($var(C_k)$), cluster diameter ($Diam(C_k)$), and mean cluster distance ($Dist(C_k)$). These metrics are described in section 2.8 respectively:

$$Var(C_k) = \sum_{p \in C_k} ||p - \mu_k||^2 \tag{16}$$

$$Diam(C_k) = \max_{p_i, p_j \in C_k} (||p_i - p_j)|| \tag{17}$$

$$Dist(C_k) = \frac{1}{|C_k|(|C_k| - 1)} \sum_{p_i \neq p_j \in C_k} ||p_i - p_j|| \tag{18}$$

$$\tag{19}$$

Note that when more than one cluster is found, these metrics are averaged across all K clusters.

Finally, separation metrics or internal validation indices were evaluated, which assess the quality of the cluster partition by measuring how well separated the clusters are from one another. In this context, we find the Silhouette coefficient (measures how similar a point "i" is to its own cluster; this metric is known to favor convex clusters), the Davies-Bouldin index (measures the ratio of the worst similarity between clusters), and the Calinski-Harabasz index or variance ratio (measures the relationship between the variance between clusters and the variance within the clusters). Their calculations are described in section 2.8 respectively:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{20}$$

$$DB = \frac{1}{K} \sum_{k=1}^{K} max_{l \neq k} (\frac{s_k + s_l}{M_{k,l}}) \tag{21}$$

$$CH = \frac{B_K/(K-1)}{W_K/(N-K)} \tag{22}$$

$$\tag{23}$$

2.9 Computational Complexity

Suppose X is a dataset, where N represents the number of observations and d the number of variables. Let k be the number of minimum points (minPts), representing the number of neighbors DBSCAN uses to create the k-distance curve. The proposed SCM process is divided into four stages: (i) preparation and

calculation of k-distances, (*ii*) smoothing using Savitzky-Golay, (*iii*) simulation of the dynamic system (RK4) with the smoothed k-distance signal, and (iv) determination of the optimal point via the maximum jerk to calculate ε.

The step-by-step complexity of the process involves the following analysis:

1. Standardization and k-NN neighborhood: Standardization, by mean and standard deviation per feature, has a complexity of $O(Nd)$. However, in practice, d is usually considered constant, so it is approximated as $O(N)$. Subsequently, the creation of the neighborhood structure uses the Kernel Density Tree (KD-Tree) algorithm, which has an intermediate complexity of $O(N\log(N))$ for fitting the model and $O(Nk\log(N))$ for the neighbor search, where k is the minimum number of points defined by DBSCAN. Finally, sorting the k-distances uses the QuickSort algorithm with a complexity of $O(N\log(N))$.
2. Smoothing using the Savitzky-Golay filter: A polynomial sliding window filter is used to process the ordered distance signal. Since the polynomial degree p and the window dimension w are fixed parameters in the implementation, the convolution procedure has a linear behavior in relation to the number of measurements, so its computational complexity is $O(Nw)$. Nonetheless, in practice, the complexity is $O(N)$.
3. Identified dynamics and numerical solution (RK4): In this step, each measurement of the smoothed signal is understood as a target position value within a mass-spring-damper system. Where its temporal evolution is determined in integration intervals of size Δt. The limited-memory BFGS method is used to optimize the dynamic system parameters, so its computational complexity is $O(N)$. Likewise, the RK4 algorithm performs four evaluations of the acceleration function per iteration, which entails a constant cost per step; therefore, the complexity is $O(N)$.
4. Calculation of jerk, energy, and selection of ε: The jerk is determined using finite differences, and the argmax is obtained through a linear scan. Therefore, the computational cost is $O(N)$.

Therefore, taking into account all steps of the explicit-causal process, the computational cost is determined by the search for the k-nearest neighbors, which leads to an approximate overall complexity of $O(N\log(N))$.

The development of the proposed explicit-causal model faces three main challenges. The first challenge was the sensitivity of the ε parameter to local density variability. For example, in high-dimensional datasets, this can cause spurious peaks in the k-distance curve. To solve this, a Savitzky-Golay filter was applied, which preserves local curvature while reducing noise without altering the signal's overall trend. The second challenge was ensuring numerical stability during simulation of the mass-spring-damper system, which was achieved by using the RK4 method to maintain accuracy at each integration interval. Subsequently, determining the optimal causal transition point is complex due to the intersection between local maxima in the system's derivative. To ensure robust estimation of the optimal ε value, the jerk is discretized and the global maximum identified. Finally, Algorithm 1 refers to the explicit-causal process. This algorithm

receives as input parameters: X, a dataset of dimension $N \times d$, k, the neighborhood used by DBSCAN, θ_0, the proposed physical parameters of the dynamic system (mass, damping, and stiffness), and h, the temporal step for RK4. On the other hand, the algorithm returns the optimal values of the ε parameter and C, as well as the cluster labels generated by DBSCAN, using ε_{opt}.

Algorithm 1 Parameter ε Estimation using Dynamic-Causal Modeling

procedure CausalEpsilonEstimation(X, k, θ, h)
  $X_{std} \leftarrow$ Standardize(X) ▷ Normalize the dataset
  $D \leftarrow$ ComputeKNN_Distances(X_{std}, k) ▷ Compute k-th distances
  $D_{sorted} \leftarrow$ QuickSort(D) ▷ Sort distances in ascending order
  $D_{smooth} \leftarrow$ SavitzkyGolay($D_{sorted}, window, poly$) ▷ Smooth k-distance curve
  $\theta \leftarrow$ BFGS(θ_0) ▷ Identification of dynamic system parameters
  for each $t \in D_{smooth}$ **do**
    Solve ODE RK4
    Calculate acceleration $a(t)$ and jerk $J(t)$
  end for
  $t^* \leftarrow \arg\max_t(J(t))$ ▷ Locate global maximum of the jerk
  $\varepsilon_{opt} \leftarrow D_{smooth}[t^*]$ ▷ Determine optimal ε value
  Labels $\leftarrow$ DBSCAN($X_{std}, \varepsilon_{opt}, k$) ▷ Apply DBSCAN with optimal ε
  return ε_{opt}, Labels
end procedure

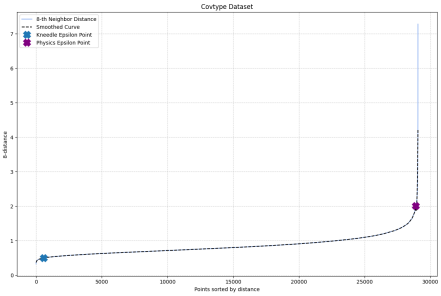
3 Results and Discussion

3.1 Experimental Design and Data Configuration

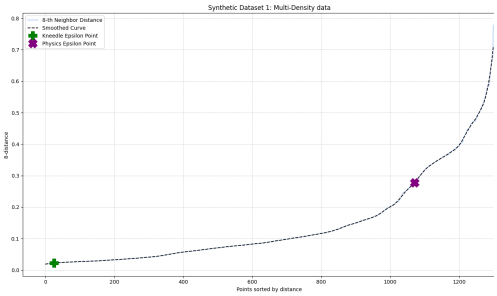
In evaluating the performance of the explicit-causal model, three experimental scenarios were considered:

1. A high-dimensional dataset called *covtype* [23], which consists of 57 variables and 581,013 records. It should be noted that only the first 10 variables were used as they are continuous numerical variables.
2. A synthetic dataset generated through a combination of the `sklearn.datasets.make_blobs` and `numpy.random.uniform` functions, using global seed 23. The set consists of 1,300 records and 2 variables and was designed to simulate a scenario with multiple densities and noise.
3. A synthetic dataset generated entirely using the `sklearn.datasets.make_blobs` function, using a global seed of 23. The set consists of 5,000 two-dimensional records and was specifically designed to create a smooth "S"-shaped k-distance curve, which presents a challenge for elbow detection methods. Its structure is composed of two stacked and concentric Gaussian distributions (both centered at [0, 0]).

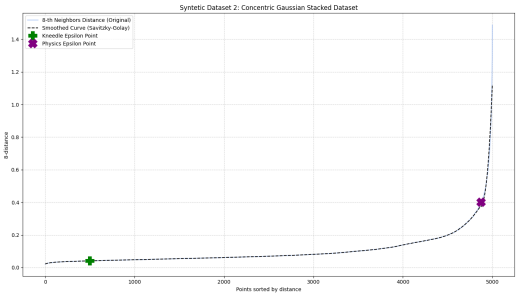
First, the k-distance graph was calculated for all datasets using Nearest Neighbors, demonstrating that the dynamics-based method (mass-spring-damper system) allowed for the detection of the threshold. Next, Figure 3 shows these graphs for each dataset. Demonstrating that the system's acceleration responds causally to the curve's variation, revealing a critical point that determines the optimal density in DBSCAN.



(a) Covtype dataset



(b) Synthetic dataset 1



(c) Synthetic dataset 2

Fig. 3: k-distances graphs for the evaluated datasets.

To reinforce the algorithm's analysis, estimates of the mass, damping, and stiffness parameters were obtained, demonstrating the model's adaptability across the datasets (see Table 1).

Data	**Mass**	**Stiffness**	**Damping**	**Calculated** ε
Covtype	2561.3921	1.0503	0.0000	2.0082
Synthetic Dataset 1	1538.9829	3.3571	0.0000	0.3764
Synthetic Dataset 2	2235.5416	4.1494	0.0000	0.4003

Table 1: Estimation of each Parameter from the model.

An interesting point is the absence of the damping parameter. In these cases, it is attenuated to zero due to the latent structure in the data where acceleration is caused solely by position (via k) and inertia (via m), and the velocity term is an irrelevant confounder whose causal edge was eliminated upon being identified with a zero weight. In this way, the model is simplified to a simple harmonic oscillator, only in these cases.

3.2 Efficiency Metrics

To objectively evaluate the quality, robustness, and significance of the clusterings generated by the ε value, two sets of metrics were applied to the two algorithms: i) cohesion metrics and ii) inter-cluster separation metrics. For cohesion, we measure the Intracluster Variance (IV), the Cluster's Mean Diameter (CMD), and the Cluster's Mean Distance (DMD). In tables 2, 3, and 4, we can find the results of these cohesion metrics. On the other hand, Tables 5, 6, and 7 present the results of the inter-cluster separation metrics.

Algorithm	**IV**	**CMD**	**DMD**
Physics Based Algorithm	143696.7723	8.4074	2.6890
Kneedle	13.3167	1.0395	0.5902

Table 2: Cohesion metric results for the Covtype dataset

Algorithm	**IV**	**CMD**	**DMD**
Physics Based Algorithm	254.9060	1.3471	0.5086
Kneedle	0.0054	0.0528	0.0250

Table 3: Cohesion metric results for the Synthetic dataset 1

Algorithm	**IV**	**CMD**	**DMD**
Physics Based Algorithm	9133.5109	7.0413	1.6767
Kneedle	0.0816	0.1199	0.0534

Table 4: Cohesion metric results for the Synthetic dataset 2

Before discussing the quality of the formed clusters, it is necessary to define the data's clustering tendency. For this, the Hopkins statistic was used; from this context, values of 0.0773 for the Covtype dataset, 0.2646 for Synthetic 1, and 0.1027 for Synthetic 2 were obtained. The low magnitude of the statistic

in all three cases indicates a tendency to cluster, consistent with the inverse formulation, where $H \to 0$.

In this context, Kneedle reports extremely low diameters, indicating that this method is fragmenting natural structures into very tiny components (87 clusters in Covtype, 7 clusters in Synthetic 1, and 64 in Synthetic 2), which suggests oversegmentation. On the other hand, the physics-based algorithm reports clusters of greater dimension and variance; given that Hopkins ensures the data is clustered, the method's greater variance indicates a more faithful representation of the clusters' real extent (core and periphery).

Algorithm	**Silhouette**	**Davies-Bouldin**	**Calinski-Harabasz**
Physics Based Algorithm	0.4124	1.7303	68.6748
Kneedle	-0.5104	1.4870	11.3456

Table 5: Inter-cluster separation metric results for the Covtype dataset

Algorithm	**Silhouette**	**Davies-Bouldin**	**Calinski-Harabasz**
Physics Based Algorithm	0.0717	2.4286	29.8945
Kneedle	-0.1932	1.1303	8.7197

Table 6: Inter-cluster separation metric results for the Synthetic dataset 1

Algorithm	**Silhouette**	**Davies-Bouldin**	**Calinski-Harabasz**
Kneedle	-0.4655	2.3219	3.2747

Table 7: Inter-cluster separation metric results for the Synthetic dataset 2

These three metrics allow evaluating the separation quality of the detected groupings; on the one hand, the Silhouette coefficient indicates that, across the three datasets presented, Kneedle consistently obtains negative values, mathematically demonstrating that the partitions are artificial. In contrast, the proposed method maintains positive values (0.4124 for the Covtype dataset and 0.0717 for Synthetic dataset 1), indicating that the data's natural structure is preserved.

An interesting point is the absence of separation metrics for the Physics-Based Algorithm (in Table 7). This fact refers to the separation metrics, which require at least two groups, which makes sense since the Synthetic 2 dataset was generated as a concentric structure with a dense core surrounded by a diffuse nebula, both centered at the origin, meaning there are no spatially separated

clusters. This fact justifies the appearance of 64 clusters in Kneedle and the negative silhouette coefficient.

3.3 Analysis and Causal Discussion

The dynamic analysis of the mass-spring-damper system showed that the jerk peaks coincide with the structural transitions detected in the density space. This correspondence validates the assumption that the ε parameter functions as a causal intervention variable, capable of inducing quantifiable reorganizations in the data's connectivity. The suggested model establishes transitions based on the system's physical response, which imparts a mechanistic character to the clustering process. This contrasts with the heuristic criterion of Kneedle, which is based on geometrically detecting the inflection point.

The findings indicate, from an interpretative perspective, that the proposed model not only optimizes clustering quality but also provides an explanatory framework for understanding how local density evolves in response to changes in the ε parameter. Identifying jerk as a transition descriptor provides a new tool for defining critical points of structural transformation in complex systems.

4 Conclusion

This article presents the view that the tuning of the ε hyperparameter in the DBSCAN algorithm should be understood as an explicit-causal process governed by the dynamics of the distance distribution. By modeling the curve using differential equations, we establish a link between local variations in the slope and structural changes in data density. This reinterpretation converts the estimation of the neighborhood radius into the identification of a causal inflection point: a slight modification in the neighborhood triggers a significant change in the cluster configuration.

The experimental results challenged the method with known complex scenarios where DBSCAN fails by definition in the literature: high-dimensional data (Covtype) and data with smooth density transitions (Synthetic 2). Both scenarios drive the emergence of variable-density algorithms such as OPTICS and HDBSCAN.

Despite this limitation, quantitative validation demonstrates how the ε radius optimized by the physics-based method produces superior quality clusterings. First, Kneedle's Silhouette coefficient yields negative scores, suggesting invalid partitions. Therefore, the physics-based method represents an absolute increase of 0.923 points in Covtype and 0.265 points in Synthetic 1. This structural superiority is also quantified by the Calinski-Harabasz index, where the physics-based method outperforms Kneedle by 505% on the Covtype dataset and 243% on Synthetic 1. Finally, the Physics-based method correctly identified the unitary topology of the Synthetic 2 dataset, demonstrating its robustness against the oversegmentation that Kneedle failed to avoid.

As future work, the following lines of research are proposed:

1. Developing an adaptive multi-scale analysis that allows for the local calculation of ε in areas with heterogeneous densities.
2. The SCM via DBSCAN can be extended to other clustering algorithms, such as graph-based and feature-learning methods, to broaden its use to non-Euclidean spaces.
3. It is recommended to analyze the combination of SCM with structural inference techniques, such as PC, to identify relationships between clustering parameters and unobservable variables in complex datasets.

Acknowledgments. The authors thank the Center for Research and Innovation in Computer Science and Educational Technology (CIICCTE) of the Faculty of Informatics at UAQ for providing the space to carry out this work.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences, 622, 178–210.
2. Torres Aguirre, V., et al. (2023). El algoritmo KMeans y el problema de los centroides iniciales. Boletín de la Sociedad Mexicana de Computación Científica y sus Aplicaciones.
3. Ran, X., Xi, Y., Lu, Y., Wang, X., & Lu, Z. (2023). Comprehensive survey on hierarchical clustering algorithms and the recent developments. *Artificial Intelligence Review,* 56(8), 8219–8264.
4. Deng, D. (2020, September). DBSCAN clustering algorithm based on density. In *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)* (pp. 949-953). IEEE.
5. Hanafi, N., & Saadatfar, H. (2022). A fast DBSCAN algorithm for big data based on efficient density calculation. *Expert Systems with Applications*, *203*, 117501.
6. Baptista, M. L., Henriques, E. M., & Goebel, K. (2021). More effective prognostics with elbow point detection and deep learning. Mechanical systems and signal processing, 146, 106987.
7. Stewart, G., & Al-Khassaweneh, M. (2022). An implementation of the HDBSCAN* clustering algorithm. *Applied Sciences*, *12*(5), 2405.
8. Li, M., Su, M., Zhang, B., Yue, Y., Wang, J., & Deng, Y. (2025). Research on a DBSCAN-IForest Optimisation-Based Anomaly Detection Algorithm for Underwater Terrain Data. *Water*, *17*(5), 626.
9. Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. Social science computer review, 9(1), 62-72.
10. Pearl, J. (2000). Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress, 19(2), 3.
11. Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of causal inference: foundations and learning algorithms (p. 288). The MIT press.
12. Kim, K., Kim, J., & Kennedy, E. H. (2024). Causal k-means clustering. arXiv preprint arXiv:2405.03083.

13. Hohmann, M., Devriendt, K., & Coscia, M. (2023). Quantifying ideological polarization on a network using generalized Euclidean distance. *Science Advances*, *9*(9), eabq2044.
14. Wang, M., Xu, X., Yue, Q., & Wang, Y. (2021). A comprehensive survey and experimental comparison of graph-based approximate nearest neighbor search. *arXiv preprint arXiv:2101.12631*.
15. dos Santos, J. V. A., & Lopes, H. (2024). Savitzky-Golay smoothing and differentiation filters for damage identification in plates. *Procedia Structural Integrity*, *54*, 575–584.
16. Бовнегра, Л. В., & Стрельбіцький, В. В. (2022). Modeling mass-spring-damper system using scilab. *Праці Одеського політехнічного університету*, (2 (66)), 96–99.
17. Wulandari, S., Iswanto, B. H., & Sugihartono, I. (2021). Determination of Springs Constant by Hooke's Law and Simple Harmonic Motion Experiment. In *Journal of Physics: Conference Series* (Vol. 2019, No. 1, p. 012053). IOP Publishing.
18. Gianasso, M. (1971). The damping coefficient in welded bodies: A theoretical study. *Meccanica*, *6*, 241–246.
19. Gómez-Aguilar, J. F., Yépez-Martínez, H., Calderón-Ramón, C., Cruz-Orduña, I., Escobar-Jiménez, R. F., & Olivares-Peregrino, V. H. (2015). *Modeling of a Mass-Spring-Damper System by Fractional Derivatives with and without a Singular Kernel*. Entropy, 17(9), 6289–6303. https://doi.org/10.3390/e17096289
20. Blum, E. K. (1962). A modification of the Runge-Kutta fourth-order method. *Mathematics of Computation*, 16(78), 176–187.
21. Franco, J. S., & Flecha, J. R. V. (2005). *Introducción a la identificación de sistemas*. Técnica Industria, 256(1).
22. A. Banerjee and R. N. Dave, "Validating clusters using the Hopkins statistic," in *2004 IEEE International Conference on Fuzzy Systems*, vol. 1. IEEE, 2004, pp. 149–153.
23. Blackard, J. (1998). *Covertype [Dataset]*. UCI Machine Learning Repository. https://doi.org/10.24432/C50K5N.

CausalMorph: Preconditioning Data for Linear Non-Gaussian Acyclic Models

Mario De Los Santos-Hernández[1][0000−0002−5283−271X], Felipe Orihuela-Espina[2][2222−3333−4444−5555], and L. Enrique Sucar[1][1111−2222−3333−4444]

[1] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, México
madlsh3517@gmail.com
[2] School of Computer Science, University of Birmingham, United Kingdom

Abstract. Linear non-Gaussian acyclic models (LiNGAMs) are a powerful class of algorithms for causal discovery, but their strict assumptions of linearity and non-Gaussian noise are often violated by real-world data, compromising identifiability. We propose CausalMorph, a three-stage preconditioning algorithm that transforms a dataset to better align with these assumptions. The method linearizes causal mechanisms using an MDL-guided polynomial fit, replaces the original residuals with synthetic non-Gaussian noise, and orthogonalizes the new noise against its causes. On a large-scale benchmark of 17,640 synthetic datasets, applying CausalMorph reduces the mean Structural Hamming Distance (SHD) of a downstream learner by a statistically significant 38.4%. Crucially, CausalMorph also improves performance in ideal settings, revealing its role as a regularizer against finite-sample artifacts. Our work establishes data preconditioning as a practical and effective strategy to broaden the applicability and enhance the robustness of the LiNGAM framework for real-world causal inference.

Keywords: Causal Discovery · Data Preconditioning · LiNGAM, Assumption Violation, CausalMorph.

1 Introduction

Within the field of causal discovery, algorithms are typically classified according to the assumptions they impose on the data-generating process, most notably the functional form of causal relations and the statistical properties of noise. Among these, the non-Gaussian Linear family of models, epitomized by ICALiNGAM and its widely adopted successor DirectLiNGAM, stands out for its theoretical strength [15, 16, 6]. These methods can, in principle, recover the full causal graph from purely observational data. However, this identifiability is strictly based on the dual assumptions of linearity and non-Gaussianity [15]. In practice, data from domains such as neuroimaging or economics seldom conform to both, severely constraining the real-world utility of this otherwise powerful algorithmic family.

To address these limitations, a broad spectrum of non-linear causal discovery models has been proposed [3, 17, 5, 9]. Although such approaches capture richer

forms of dependence, they typically incur substantial computational costs, require large sample sizes, or introduce their own restrictive assumptions. As a result, practitioners often face a trade-off: employ a sophisticated but potentially unstable non-linear model, or rely on a simpler, well-understood linear approach that risks misspecification. In this work a third path is explored. Rather than developing yet another discovery algorithm, the central question posed is whether observational data can be systematically transformed to satisfy the stringent assumptions of the non-Gaussian linear framework [15, 16].

To this end, CausalMorph is introduced as a targeted data preconditioning algorithm for LiNGAM-based learners. The procedure follows a principled three-stage pipeline: (i) linearization of potentially non-linear relationships, (ii) synthesis of non-Gaussian residuals, and (iii) enforcement of independence between causes and residuals. An extensive evaluation across 17,640 synthetic datasets demonstrates several contributions. First, the CausalMorph algorithm constitutes a novel preconditioning method that significantly improves the accuracy of downstream learners, yielding a statistically significant 38. 4% reduction in the mean structural Hamming distance (SHD). Second, a pronounced regularization effect is observed: CausalMorph improves discovery accuracy even under ideal LiNGAM conditions, indicating mitigation of finite-sample artifacts. Finally, these gains broaden the practical applicability of the LiNGAM family, increasing robustness and reliability across a wider range of real-world settings.

2 Theoretical Framework and Problem Formulation

2.1 Structural Equation Models and Causal Graphs

The relationships among a set of p variables $V = \{X_1, \ldots, X_p\}$ can be represented by a *Structural Equation Model (SEM)*. The causal structure of such a model is encoded as a *Directed Acyclic Graph (DAG)* $G = (V, E)$, where each directed edge $X_j \to X_i$ in the set of edges E indicates that X_j is a direct cause of X_i. The set of direct causes of X_i is denoted by $\text{pa}(X_i)$. Acyclicity implies the existence of a valid *causal order*, denoted by $\prec$, such that for every edge $X_j \to X_i$, it holds $j \prec i$. The observational data are given as a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, consisting of n independent samples of the variables in V.

2.2 The Linear Non-Gaussian Acyclic Model (LiNGAM)

The preconditioning target of this work is the data-generating process assumed by the *Linear Non-Gaussian Acyclic Model (LiNGAM)* framework. Algorithms in this family, including ICALiNGAM [15] and DirectLiNGAM [16], are notable for their ability to identify the complete causal structure from purely observational data. This identifiability is possible only under a set of strict assumptions. The SEM for a LiNGAM process is given by:

$$X_i = \sum_{X_j \in \text{pa}(X_i)} \beta_{ij} X_j + E_i, \tag{1}$$

where β_{ij} represents the linear causal effect of X_j on X_i, and E_i is an unobserved noise term. The four assumptions required for identifiability are:

1. **Linearity**: The functional relationship between each variable and its direct causes is linear and additive, as specified in Equation 1.
2. **Acyclicity**: The causal graph G contains no directed cycles, ensuring a well-defined causal ordering.
3. **Non-Gaussianity**: The noise terms E_i are drawn from non-Gaussian distributions. This condition enables the use of techniques such as Independent Component Analysis (ICA) or iterative regression to resolve causal directions, which are otherwise unidentifiable under Gaussian noise using only second-order statistics.
4. **Independence**: Each noise term E_i is statistically independent of the direct causes $\text{pa}(X_i)$ and also independent of all other noise terms E_j for $j \neq i$.

The synthetic data generator used in this study was designed to precisely control these conditions. By tuning parameters such as the degree of non-linearity and the noise distribution ('dist'), datasets can be created that either adhere to or systematically violate the LiNGAM assumptions. This provides a controlled environment for assessing the effectiveness of the proposed approach.

2.3 Problem Formulation

The problem considered here is the mismatch between real-world observational data and the idealized assumptions of the LiNGAM framework. In particular, it is assumed that the observed dataset $\mathbf{X}$ is generated by a process that can violate both linearity and non-Gaussianity, thus limiting the identifiability guarantees of LiNGAM-based discovery algorithms.

The objective is to define a projection operator P such that the projected dataset $\mathbf{X}^* = P(\mathbf{X})$ lies closer to the space of data distributions that satisfy the LiNGAM assumptions. When applying a LiNGAM-based discovery algorithm to $\mathbf{X}^*$, the estimated causal graph $\hat{G}^*$ should provide a more accurate approximation of the ground truth graph G than the graph $\hat{G}$ obtained directly from $\mathbf{X}$.

The proposed solution to this problem is *CausalMorph*, an algorithm that realizes P through a principled preconditioning pipeline. Its purpose is to project data distributions into a LiNGAM-compatible space in a manner that enhances the robustness and reliability of causal discovery while preserving underlying causal relationships.

3 Background and Related Work

The field of causal discovery has produced a diverse ecosystem of algorithms, each defined by a distinct set of assumptions and trade-offs. To properly situate the contribution of CausalMorph, this section reviews the LiNGAM family, alternative non-linear approaches, and the role of data transformations in causal

discovery. Particular emphasis is placed on the advantages of LiNGAM that motivate its pairing with CausalMorph.

3.1 The Linear Non-Gaussian Acyclic Model (LiNGAM) Family

LiNGAM represents a cornerstone of modern causal discovery. Its foundational principle, first established in **ICALiNGAM** [15], is that a linear causal model with non-Gaussian noise is equivalent to the independent component analysis model (ICA). This property enables unique identification of the full causal structure from observational data. Subsequent refinements, such as **DirectLiNGAM** [16], further improved efficiency by directly identifying a causal order without requiring iterative ICA.

The key distinguishing feature of LiNGAM-based methods is that they return a fully directed causal graph, rather than a Markov equivalence class. This is a decisive advantage for applications where edge orientation is critical, such as neuroimaging or econometrics. Moreover, the relative computational efficiency and interpretability of LiNGAM make it a practical choice for high-dimensional data. The limitation, however, lies in its reliance on strict assumptions of linearity and non-Gaussianity.

3.2 Beyond Linearity: Non-Linear Causal Discovery

To relax the LiNGAM assumptions, a broad class of methods has been developed. Of particular interest are *direct modifications and extensions of the algorithm*, which preserve its linear–non-Gaussian identifiability lever while broadening scope:

- **GroupLiNGAM** [7]: Extends LiNGAM to sets of variables. By detecting and ordering groups whose internal structure may be unresolved, GroupLiNGAM returns a partially ordered block DAG that remains compatible with the non-Gaussian identifiability of LiNGAM.
- **VAR-LiNGAM / SVAR via non-Gaussianity** [4]: Brings LiNGAM ideas to time series by exploiting non-Gaussianity in structural vector autoregressive models. Temporal structure plus non-Gaussian shocks enable identification of contemporaneous effects along with lagged dynamics.
- **LiNGAM with latent confounders** [2]: Addresses hidden variables in non-Gaussian linear models. Under appropriate constraints, parts of the causal graph remain identifiable despite latent confounding, extending the applicability of LiNGAM beyond strictly observed DAGs.
- **LiM (Linear Mixed data)** [18]: Target variables mixed-type (e.g. continuous and discrete) within a LiNGAM-style setting, adapting estimation to heterogeneous data while retaining non-Gaussian orientation cues.

For context, broader nonlinear lines remain relevant:

- **Additive Noise Models (ANMs)**: Relax linearity while assuming additive, independent noise [3].

- **Neural and gradient-based methods**: e.g., NOTEARS [17], which casts structure learning as smooth optimization; flexible but computationally heavier and hyperparameter sensitive.
- **Score-based and representation-learning approaches**: e.g., TCL/nonlinear ICA [5], and NoGAM [9], which leverage auxiliary structure (non-stationarity) or scores to move beyond Gaussian noise. Some return only MECs; others provide full orientation under additional assumptions.

This spectrum highlights a persistent trade-off: flexibility and generality often increase computational and sample complexity or weaken identifiability. In contrast, LiNGAM-family methods retain exact orientation under their assumptions, motivating the CausalMorph projection toward that regime.

3.3 Data Transformations in Causal Discovery

The idea of transforming data prior to analysis has a long history in statistics and machine learning. Generic preprocessing, such as whitening, is commonly applied, but these transformations can be counterproductive in causal discovery [6, 11, 10]. For example, whitening destroys the distributional structure that ICA-based methods require [6].

What is needed is not a generic preprocessing step, but a causally aware projection that aligns the data with the assumptions of the chosen discovery algorithm. This is precisely the role of the pressented algorithm. By projecting data into a LiNGAM compatible space, CausalMorph retains the unique advantages of the LiNGAM family: efficient computation, interpretability, and recovery of a fully directed causal structure, while mitigating the vulnerability to real-world deviations from linearity and non-Gaussianity [15, 16, 8].

4 Method: CausalMorph

CausalMorph is a data preconditioning pipeline that projects observational data toward a regime more closely aligned with LiNGAM assumptions. The procedure is applied sequentially to each variable in a preliminary causal order. For a node Y with tentative parent set X_p, three stages are performed: (i) MDL-guided local linearization of the conditional mechanism, (ii) synthesis of non-Gaussian residuals and (iii) orthogonalization of residuals with variance matching. Methodological details are given below.

- **Stage I: MDL-guided local linearization.** Parent data X_p are standardized to zero mean and unit variance. The conditional relationship between Y and X_p is approximated by fitting a polynomial $p(X_p)$, with the degree chosen by minimizing the Minimum Description Length (MDL) criterion:

$$\text{MDL} \;=\; n \log(\text{MSE} + \epsilon_{\log}) + \lambda k, \tag{2}$$

  where n is the sample size, MSE the mean squared error of the fit, k the number of polynomial terms, λ a penalty parameter and $\epsilon_{\log}$ a small constant

for numerical stability. A first-order Taylor expansion around a robust anchor $\mathbf{x}_0$ (coordinate-wise median of X_p) yields

$$Y_{\text{lin}} \;=\; p(\mathbf{x}_0) + \big(\nabla p(\mathbf{x}_0)\big)^{\top}(X_p - \mathbf{x}_0), \tag{3}$$

with $\nabla p(\mathbf{x}_0)$ estimated by finite differences. The residual $E_{\text{orig}} = Y - Y_{\text{lin}}$ captures stochastic noise and the remaining nonlinear effects.

– **Stage II: synthetic non-Gaussian residuals**. Residuals E_{orig} are whitened to obtain a decorrelated representation and an associated coloring matrix. Candidate non-Gaussian distributions (Laplace, Uniform, Exponential, and Student's t) are sampled, recolored using the same matrix, and evaluated with the Shapiro–Wilk test. The candidate exhibiting the strongest evidence against normality is selected, producing the synthetic residual vector E_{synth}.

– **Stage III: orthogonalization and variance matching.** An orthonormal basis Q for $\text{span}(X_p)$ is calculated by QR decomposition. The projection of E_{synth} onto this subspace is removed,

$$E_{\text{ortho}} \;=\; E_{\text{synth}} - QQ^{\top}E_{\text{synth}}, \tag{4}$$

ensuring linear uncorrelatedness with the parent set. To preserve the original signal-to-noise ratio, E_{ortho} is rescaled to match $\sigma(E_{\text{orig}})$, and the variable is reconstructed as:

$$Y_{\text{new}} \;=\; Y_{\text{lin}} + E_{\text{ortho}} \cdot \frac{\sigma(E_{\text{orig}})}{\sigma(E_{\text{ortho}})}. \tag{5}$$

Repeating the three stages for all variables in the specified order produces the projected dataset $\mathbf{X}^*$. The resulting data better align with the LiNGAM assumptions while retaining the original causal dependencies. Numerical safeguards (e.g., skipping variables with near-zero variance in parents or residuals) enhance stability. The complete specification appears in Algorithm 1.

Algorithm 1 CausalMorph Algorithm

Require: Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, preliminary causal order $\mathcal{O}$
Ensure: Projected data matrix $\mathbf{X}^* \in \mathbb{R}^{n \times p}$

$\mathbf{X}^* \leftarrow \mathbf{X}$
for each variable index i in $\mathcal{O}$ **do**
  $Y \leftarrow \mathbf{X}^*_{:,i}$
  Tentative parent set $X_p \leftarrow \{\mathbf{X}^*_{:,j} \mid j \prec i \text{ in } \mathcal{O}\}$
  if $X_p \neq \emptyset$ **and** $\min_j \mathrm{Var}(X_{p,j}) > \epsilon$ **then**
    Stage I: MDL-Guided Local Linearization
    $X_p^{\text{scaled}} \leftarrow \mathrm{Standardize}(X_p)$
    Choose $d^* \in \{1, \ldots, D\}$ minimizing $\mathrm{MDL} = n \log(\mathrm{MSE} + \epsilon_{\log}) + \lambda k$
    Fit polynomial $p(\cdot)$ of degree d^*; compute Y_{lin} via first-order Taylor expansion around $\mathrm{median}(X_p^{\text{scaled}})$
    $E_{\text{orig}} \leftarrow Y - Y_{\text{lin}}$
    Stage II: Synthetic Non-Gaussian Residuals
    $(Z_{\text{white}}, C) \leftarrow \mathrm{Whiten}(E_{\text{orig}})$ *(C is the coloring matrix)*
    $\mathcal{D}^* \leftarrow \underset{\mathcal{D} \in \{\text{Laplace}, \text{Uniform}, \text{Exponential}, t\}}{\arg\min} \mathrm{SW\_pvalue}\big(\mathrm{Color}(Z \sim \mathcal{D}, C)\big)$
    $E_{\text{synth}} \leftarrow \mathrm{Color}(Z \sim \mathcal{D}^*, C)$
    Stage III: Orthogonalization and Variance Matching
    $E_{\text{ortho}} \leftarrow E_{\text{synth}} - \mathrm{Proj}_{X_p}(E_{\text{synth}})$
    $E_{\text{final}} \leftarrow E_{\text{ortho}} \cdot \dfrac{\sigma(E_{\text{orig}})}{\sigma(E_{\text{ortho}})}$
    $Y_{\text{new}} \leftarrow Y_{\text{lin}} + E_{\text{final}}$
    $\mathbf{X}^*_{:,i} \leftarrow Y_{\text{new}}$
  end if
end for
return $\mathbf{X}^*$

Remark (Preliminary causal order). The input order $\mathcal{O}$ is a *preliminary* or reference order and does not necessarily coincide with the true causal order. It may be obtained from prior knowledge, domain heuristics, or a rough initial estimator. CausalMorph does not attempt to learn $\mathcal{O}$; it uses $\mathcal{O}$ only to define tentative sets of parents during projection. Inaccuracies in $\mathcal{O}$ do not invalidate the procedure but can reduce the efficacy of the projection. In practice, $\mathcal{O}$ can be iteratively refined (e.g. using a downstream LiNGAM estimate on $\mathbf{X}^*$) and the projection can be reapplied.

5 Experimental Setup

5.1 Synthetic Data Generation

A large-scale benchmark of synthetic datasets was constructed to investigate performance under controlled violations of LiNGAM assumptions. For each dataset, a ground-truth DAG on the p nodes was sampled by adding a directed edge $i \to j$ for $i < j$ with probability p_{conn}, ensuring acyclicity. Given parents $\mathrm{pa}(X_i)$, the node values were generated by the SEM

$$X_i \;=\; f_i\big(\mathrm{pa}(X_i)\big) \;+\; E_i, \tag{6}$$

where parent effects are aggregated linearly with random weights and then optionally passed through a nonlinearity. Specifically, letting $u_i = \boldsymbol{\beta}_i^\top \mathbf{x}_{\mathrm{pa}(i)}$, the mechanism is:

$$f_i(\mathbf{x}_{\mathrm{pa}(i)}) \;=\; (1-\alpha)\, u_i \;+\; \alpha\, g(u_i), \tag{7}$$

with g and $\alpha \in [0,1]$ controlling the degree of nonlinearity. Additive noise E_i is drawn node-wise from one of normal, uniform, Laplace, and exponential with scale set by $\sigma_{\mathrm{noise}} = \texttt{deviation}$.

5.2 Experimental Parameters and Evaluation Protocol

Synthetic data sets are generated on a factorial grid summarized in Table 1. For each dataset, **DirectLiNGAM** is executed (i) on the raw data and (ii) on the projected data $\mathbf{X}^*$ produced by CausalMorph, using reference hyperparameter defaults. The resulting estimates $\hat{G}$ and $\hat{G}^*$ are compared with the ground truth DAG.

Table 1. Factorial grid for synthetic data generation. Parameter names mirror the implementation.

Parameter	**Values**
Graph structure	
Number of variables (p)	{5, 25, 40}
Connection probability (p_{conn})	{0.25, 0.5, 0.75}
Data properties	
Sample size (n)	{50, 500, 5000}
Noise scale (σ_{noise})	{0, 0.25, 0.5, 0.75}
Proportion of Gaussian noise sources	{0%, 25%, 75%, 100%}
Mechanism nonlinearity	
Mode (`mode`)	{`linear`, `nonlinear`}
Nonlinearity weight (α)	{0.25, 0.5, 0.75} (only if `mode`=`nonlinear`)

Primary metrics.

- **Structural Hamming Distance (SHD).** Number of edge additions, deletions, or reversals required to transform the estimate into the ground truth (lower is better).
- **F1 (adjacency).** Harmonic mean of precision and recall of the presence of the edge, regardless of direction.

Statistical testing.

For each configuration, the SHD is analyzed in a paired design by computing the differences per replica $\Delta = \mathrm{SHD}_{\mathrm{raw}} - \mathrm{SHD}_{\mathrm{CM}}$. The normality of Δ is assessed with the Shapiro–Wilk test [14]. If normality is not rejected, a paired test t is used and Cohen's d_z is reported; otherwise, the Wilcoxon signed rank test is applied and the size of the rank-biserial effect is reported. For all effect sizes, 95% bias corrected and accelerated bootstrap confidence intervals (BCa) are provided. Family-wise error across configurations is controlled through the Holm procedure. An overall summary across the full grid is also reported by aggregating the paired differences and applying the same testing protocol.

6 Results

CausalMorph was evaluated using the synthetic benchmark described in Section 5. Across conditions, preconditioning with CausalMorph yields substantial and statistically significant gains in the accuracy of causal discovery. Across the entire grid, applying CausalMorph before **DirectLiNGAM** reduces mean Structural Hamming Distance (SHD) from 0.396 (raw) to 0.244, a relative improvement of **38.4%**. Improvements occur in more than 93% configurations and are highly significant (paired t-test, $t(17279) = 141.7$, $p < 0.001$; Cohen's $d = 1.07$). Benefits persist in LiNGAM-specific regimes (purely linear mechanisms with fully non-Gaussian errors), where SHD still decreases significantly ($p < 0.001$; Cohen $d = 1.32$), indicating a data-level regularization effect beyond correcting assumption violations.

Gains are highest with stronger violations of the LiNGAM assumptions. As the proportion of *Gaussian* noise sources increases, a known challenge to identifiability, the SHD gap between CausalMorph and the baseline widens (Fig. 1, left). Under nonlinear mechanisms, the SHD is consistently lower with CausalMorph, with a larger margin than in the linear case (Fig. 1, right).

The advantages are robust to the characteristics of the system. SHD decreases with sample size for both methods, while a consistent CausalMorph advantage is maintained across all (n, p) combinations (Fig. 2). The absolute SHD increases with the density of the graph p_{conn} for both methods, yet the improvement from CausalMorph persists at all densities, indicating its applicability in sparse and dense graphs (Fig. 3).

7 Discussion

Empirical results indicate that CausalMorph functions as an effective preconditioning step for LiNGAM-based discovery. Gains increase as the data-generating process departs from LiNGAM assumptions (greater Gaussian noise, stronger nonlinearity) and remain present in regimes favorable to LiNGAM, consistent with a regularization effect. The three-stage design targets key drivers of identifiability: local linearization reduces model misspecification; residual synthesis

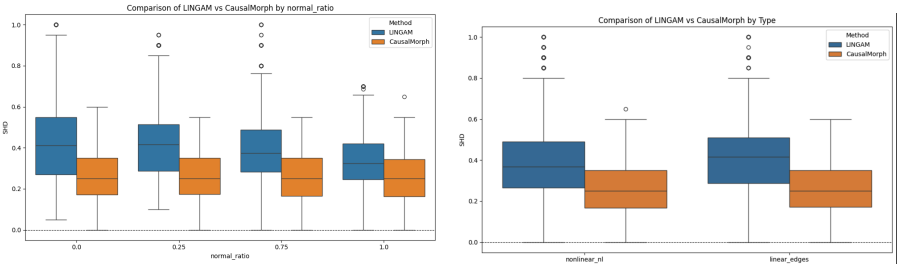


Fig. 1. Assumption stress tests. Left: SHD vs. proportion of Gaussian noise sources; improvements increase as errors become more Gaussian. **Right:** SHD under linear vs. nonlinear mechanisms; larger gains under nonlinearity.

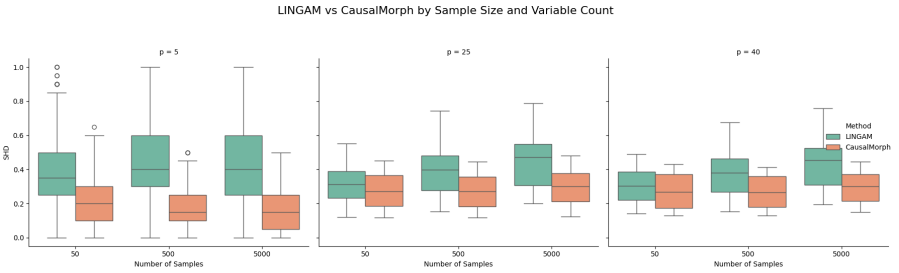


Fig. 2. Robustness over sample size and dimensionality. SHD for DirectLiNGAM with and without CausalMorph across $n \in \{50, 500, 5000\}$ and $p \in \{5, 25, 40\}$. Lower error with CausalMorph in every panel.

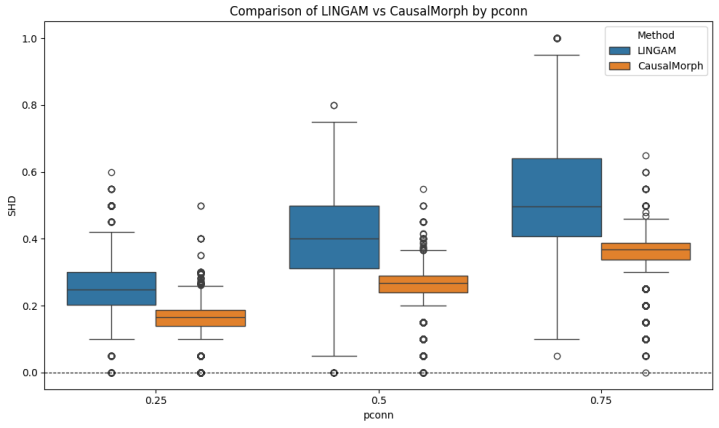


Fig. 3. Robustness over graph density. SHD across $p_{\text{conn}} \in \{0.25, 0.5, 0.75\}$; improvements persist at all densities.

strengthens the non-Gaussian signal; orthogonalization removes first-order dependence between causes and errors.

The trends in Figs. 1–3 are consistent with these mechanisms. Improvements in ideal regimes suggest mitigation of finite-sample artifacts (e.g., sampling-induced correlations, attenuated higher-order signal), effectively regularizing the empirical distribution toward the target SEM. Current limitations include reliance on a preliminary causal order and the absence of finite-sample guarantees.

7.1 Future Work

- **Comparative evaluation within the LiNGAM family.** Perform head-to-head studies against direct LiNGAM modifications and extensions only (e.g., ICA-/DirectLiNGAM variants, GroupLiNGAM, LiNGAM with latent confounders, VAR-/SVAR-LiNGAM, mixed data adaptations). Use matched protocols and metrics to isolate where a LiNGAM-targeted projection is beneficial relative to algorithmic variants that retain the linear–non-Gaussian identifiability lever.
- **Real-world application on fNIRS data.** Apply the pipeline to functional near-infrared spectroscopy datasets, evaluating robustness under physiological noise and hemodynamic confounding, and compare specifically against LiNGAM family baselines using domain-relevant endpoints (e.g., task-evoked directed connectivity).
- **Mathematical validation.** Establish formal properties of the projection independent of finite-sample analysis: (i) conditions under which the projection preserves (or improves) LiNGAM identifiability; (ii) invariance results (e.g., equivariance under admissible reparameterizations and scaling); (iii) monotonicity of a suitable risk or contrast under the projection; (iv) existence/uniqueness and stability of the three-stage mapping; and (v) convergence of iterated projection–estimation schemes to fixed points consistent with LiNGAM assumptions.

8 Conclusion

CausalMorph projects observational data toward a LiNGAM-compatible regime via local linearization, synthetic non-Gaussian residuals, and orthogonalization. In large-scale simulations, the mean SHD is reduced by **38.4%** on average, with robust gains across sample sizes, dimensionalities, and graph densities, and with regularization effects evident even in ideal regimes. These results broaden the practical applicability of LiNGAM-style methods and motivate further theoretical analysis and real-world validation.

References

1. Grünwald, P.D.: *The Minimum Description Length Principle.* MIT Press, Cambridge, MA (2007)

2. Hoyer, P.O., Shimizu, S., Kerminen, A., Palviainen, M.: Estimation of causal effects using linear non-Gaussian causal models with hidden variables. International Journal of Approximate Reasoning **49**(2), 362–378 (2008)
3. Hoyer, P.O., Janzing, D., Mooij, J.M., Peters, J., Schölkopf, B.: Nonlinear Causal Discovery with Additive Noise Models. In: Advances in Neural Information Processing Systems **21**, pp. 689–696 (2009)
4. Hyvärinen, A., Shimizu, S., Hoyer, P.O.: Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity. Journal of Machine Learning Research **11**, 1709–1731 (2010)
5. Hyvärinen, A., Morioka, H.: Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. In: Advances in Neural Information Processing Systems **29**, 3772–3780 (2016)
6. Hyvärinen, A., Oja, E.: Independent Component Analysis: Algorithms and Applications. Neural Networks **13**(4–5), 411–430 (2000). https://doi.org/10.1016/S0893-6080(00)00026-5
7. Kawahara, Y., Shimizu, S., Washio, T.: GroupLiNGAM: Linear non-Gaussian acyclic models for sets of variables. arXiv:1006.5041 (2010)
8. Liu, H., Lafferty, J., Wasserman, L.: The Nonparanormal: Semiparametric Estimation of High-Dimensional Undirected Graphs. In: Advances in Neural Information Processing Systems **22**, 2295–2303 (2009)
9. Montagna, F., Noceti, N., Rosasco, L., Zhang, K., Locatello, F.: Causal Discovery with Score Matching on Additive Models with Arbitrary Noise. In: Proceedings of the 2nd Conference on Causal Learning and Reasoning (CLeaR 2023), PMLR, vol. 213, pp. 726–751 (2023)
10. Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J., Schölkopf, B.: Distinguishing cause from effect using observational data: methods and benchmarks. Journal of Machine Learning Research **17**(32), 1–102 (2016)
11. Peters, J., Janzing, D., Schölkopf, B.: *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA (2017)
12. Rissanen, J.: Modeling by shortest data description. Automatica **14**(5), 465–471 (1978). https://doi.org/10.1016/0005-1098(78)90005-5
13. Rolland, P., Cevher, V., Kleindessner, M., Russell, C., Janzing, D., Schölkopf, B., Locatello, F.: Score Matching Enables Causal Discovery of Nonlinear Additive Noise Models. In: ICML 2022, PMLR, vol. 162, pp. 18741–18753 (2022)
14. Shapiro, S.S., Wilk, M.B.: An Analysis of Variance Test for Normality (Complete Samples). Biometrika **52**(3–4), 591–611 (1965). https://doi.org/10.1093/biomet/52.3-4.591
15. Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A.: A Linear Non-Gaussian Acyclic Model for Causal Discovery. Journal of Machine Learning Research **7**, 2003–2030 (2006)
16. Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P.O., Bollen, K.: DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. Journal of Machine Learning Research **12**, 1225–1248 (2011)
17. Zheng, X., Aragam, B., Ravikumar, P., Xing, E.P.: DAGs with NO TEARS: Continuous Optimization for Structure Learning. In: Advances in Neural Information Processing Systems **31**, 9472–9483 (2018)
18. Zeng, Y., Shi, C., Zheng, S., Zhang, K.: Causal Discovery for Linear Mixed Data. In: Proceedings of the 1st Conference on Causal Learning and Reasoning (CLeaR 2022), PMLR, vol. 177, pp. 1698–1722 (2022)

Time Series Prediction Based on Causal Discovery

Julio Muñoz-Benítez[1] and L. Enrique Sucar[1]

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro No.1, Sta. María Tonantzintla, Puebla, México.
{jcmunoz,esucar}@inaoep.mx

Abstract. Prediction based on causal models has the advantage that the relevant variables that affect the variable of interest can be identified, providing, in principle, better results than using traditional statistical (associative) models. In this work, we propose an approach to time series prediction based on causal models. Historical data from a time series is fed to a causal discovery algorithm; and the resulting structure is used to perform predictions on the time series, based on causal relations between variables. We also consider the case of subsampled time series, for which the causal structure is recovered based on imputing missing data using adversarial neural networks, and this is used for prediction. Experiments on synthetic time series show that the predictions based on the causal relations for both, the model learned with the original data and the one with imputed data, obtain a significant reduction in the mean square error compared to predictions based only on the variable of interest, and on predictions based on all the variables. This reflects the importance of knowing the causal interactions to make predictions based on the most relevant variables. Then we demonstrate the application of the proposed method on a real-world scenario on energy markets, to predict the daily energy demand and electricity price.

Keywords: Causal Discovery · Time Series · Subsampling · Time Series Prediction

1 Introduction

Time series analysis and forecasting constitute a well established and continually evolving area of research and interest, motivated by their critical role in numerous domains, including economics, business, public policy, social sciences, environmental monitoring, medicine, and finance. The ability to accurately anticipate future observations is fundamental to a wide range of strategic planning and decision-making processes. Although traditional forecasting models are often based on statistical patterns and correlations, these approaches may not capture the true generative mechanisms underlying the data. In particular, when temporal resolution is limited due to subsampling, standard models may infer spurious associations or miss critical dependencies altogether.

In this context, causal knowledge offers a more robust mechanism for predictive analysis by enabling not only the capture of observed correlations, but also the inference of generative mechanisms that explain how variables interact and influence each other. Incorporating causal information into the modeling process can enhance interpretability, improve robustness under changes in the data distribution, and guide the selection of relevant variables for forecasting. This perspective is particularly valuable when dealing with high-dimensional and temporally complex data (Peters et al., 2017). Incorporating causal knowledge in variable selection has gained increasing attention as a means to improve predictive modeling. Traditional feature selection methods often rely on statistical associations, which may include spurious correlations that do not generalize across different settings. In contrast, causal variables, those that have a direct or indirect effect on the target variable, are more likely to maintain their predictive relation under distributional changes or interventions.

Inferring causal relations from time series data has served as the basis for causal discovery in various fields such as climate systems, ecological networks, effective connectivity in the brain, and finance (Hyttinen et al., 2017). However, one of the main challenges of causal discovery from time series is that causal interactions may occur on a time scale faster than the frequency of measurement (Hyttinen et al., 2017; Lawrence et al., 2020), this phenomena is known as *subsampling*. This can lead to a loss of valuable information to determine the true causal relationships between events, resulting in significant errors in the obtained causal structure, as shown in previous work (Danks and Plis, 2014). Although causal discovery in subsampled time series is relatively under explored, it is a challenge that must be addressed as it is common in practical applications.

In this work, we explore the use of causal discovery techniques in time series prediction, with a focus on data affected by subsampling. We propose a two-stage methodology: first, we reconstruct the missing data using adversarial neural networks to approximate original time series; second, we apply a causal discovery algorithm to recover the causal structure, which is then used to select input variables that are not merely correlated, but causally related to the target variable, resulting in more robust and accurate predictions.

We validate the proposed approach using both, synthetic data and a real world energy consumption dataset. Our experiments show that the predictions based on the causal relations obtain a significant reduction in the mean square error compared to predictions based only on the variable of interest, and on predictions based on all the variables. Additionally, incorporating causal knowledge not only improves prediction accuracy but also enhances model interpretability and robustness under challenging data conditions.

The paper is organized as follows. Background is provided in Section 2. Section 3 reviews related work. In Section 4, we describe time series prediction using causal knowledge. Section 5 presents the experimental results obtained from a synthetic dataset and a real world scenario. Section 6 gives the conclusions and directions for future work.

2 Background

Time series are a sequence of data that can be denoted as $X_t = (X1_t, X2_t, ..., Xk_t)$ where k is the number of variables measured at discrete time steps, $t \in [1, T]$. Over the past several years, several approaches have been developed for time-series prediction. In the following subsections, we briefly describe Bayesian and neural network methods; and then we introduce the problem of subsampling.

2.1 Dynamic Bayesian Networks

Dynamic Bayesian Networks (DBNs) are an extension of Bayesian Networks (BNs) to model temporal processes by capturing dependencies across time steps. In this framework, the system is represented as a two-slice temporal Bayes net, where each time slice includes a set of random variables, and directed edges represent conditional dependencies both within and across slices. This allows DBNs to represent how a system evolves over time. DBNs are particularly useful when the time series presents temporal and probabilistic dependencies, and they support both inference and learning from time series data. For prediction tasks, once the model has been learned from observed sequences, future values of the variables can be inferred by propagating beliefs through the network using algorithms such as particle filtering. These models have been applied in areas such as speech recognition, finance, and health monitoring, offering a probabilistic approach to temporal prediction that incorporates uncertainty and domain knowledge (Murphy, 2002).

2.2 Neural Networks

Neural networks have gained popularity in time series prediction due to their ability to capture nonlinear and complex interactions among variables. Artificial Neural Networks (ANNs) have been applied to both, univariate and multivariate prediction tasks, and architectures like Convolutional Neural Networks (CNNs) have been adapted to capture local temporal patterns within fixed-size windows (Lim and Zohren, 2021).

In recent years, deep learning approaches have shown an outstanding performance in time series prediction, especially in scenarios involving long-term dependencies and nonlinear dynamics. Recurrent Neural Networks (RNNs), and particularly their variants such as long-short-term memory (LSTM) networks, have been widely adopted to capture temporal correlations by maintaining memory of relevant past inputs. More recently, Transformer models have emerged as state-of-the-art approaches for modeling time series to learn long-range dependencies without recurrence (Wen et al., 2023). These models typically outperform classical methods when trained on large datasets, but come with increased computational demands and model complexity.

Traditional forecasting methods rely on the identification of statistical relations such as correlations or temporal patterns in the data. Although these

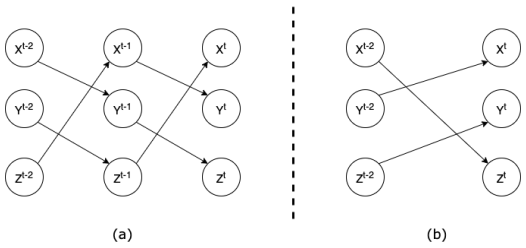


Fig. 1. Causal structures for a time series with variables X, Y, Z. (a) Original structure. (b) Structure obtained from subsampled data (every two time steps).

approaches have achieved notable success in various applications, they do not explicitly model causal relationships between variables. As a result, predictions may be based on spurious associations. Ignoring causality also limits interpretability, making it difficult to understand the factors that influence the variable of interest. In contrast, incorporating causal knowledge provides a way to identify the true relationships among data, offering predictions that are more robust to interventions and distributional changes. Recent work suggests that using causal information can enhance model generalization and lead to more reliable forecasting, particularly in complex or non-stationary environments (Runge, 2018).

2.3 Subsampling in Time Series

One of the main challenges of using data collected in time series is that causal interactions may occur at a different resolution than the one at which observations are made (Hyttinen et al., 2017; Lawrence et al., 2020; Solovyeva et al., 2023). When the sampling frequency is too low relative to the causal relations, crucial information about cause-effect relationships can be lost. This is illustrated in Figure 1, where the original causal structure is depicted in Figure 1(a), and the modified structure that results from observing the system every two time steps is shown in Figure 1(b). If it is assumed that the structure of Figure 1(b) is correct, valuable information about the true causal relationships between the variables is lost. This may lead to believe that variable Z can be intervened to control Y, but the true influence of Z on Y is mediated by X. Thus, an intervention in X would be more effective. Similarly, if the structure of Figure 1(b) is used, the predictions of the behavior of the variables can be completely different from those obtained if the true causal structure of the time series is used (Hyttinen et al., 2017). This loss of information due to subsampling can significantly impact the analyzes and the accuracy of predictions derived from observed time series data. Specifically, when the sampling rate is insufficient to capture the underlying causal mechanisms this not only affects causal inference but can also affect predictive performance, as models may miss key temporal dependencies or infer spurious relationships, leading to biased or less accurate forecasts (Runge, 2018).

3 Related Work

We present a summary of related work, including approaches that incorporate causal discovery in time series prediction; and causal discovery for the case of subsampling.

3.1 Causal Discovery in Time Series

Recent studies have explored the integration of causal knowledge in time series prediction, showing that incorporating causal relationships can improve the precision and robustness of the forecast. Kristjanpoller et al. (2025) proposed a framework that integrates causal relations to predict financial time series, outperforming traditional approaches. Tierney et al. (2023) developed a multivariate Bayesian dynamic model for causal prediction, enabling sequential learning and the analyzes of intervention effects in time series data. (Li et al., 2021) addressed the challenge of domain adaptation in time series forecasting by proposing a method that leverages stable causal structures across domains to improve generalization under causal conditional changes. In (Faruque et al., 2024), a deep learning approach is proposed to learn temporal causal relations from nonlinear, non-stationary time series data. More recently, Tang (2024) conducted an empirical study applying time series causal discovery algorithms to equity markets, showing that strategies informed by causal structures can lead to profitable investment outcomes. However, while these approaches incorporate causal information to improve prediction, they generally assume that the observed data reflect causal relations in the system and do not take into account the effects of subsampling, an issue that can affect the causal relationships in the model and degrade predictive performance.

3.2 Causal Discovery in Subsampled Time Series

As mentioned previously, if the time series data is undersampled, the true causal relationships can not be obtained, in general. For this, Danks and Plis (2014) developed an algorithm that allows learning a set of causal structures even if the level of subsampling is unknown. This is performed through a graphical representation of the causal structure of the time series. Subsequently, all the possible causal structures are obtained, comparing them with the initial causal structure, which may be affected by some degree of subsampling. In this way, if the new structures are consistent with the original structure they are considered as a possible causal structures, obtaining an *equivalence class* of causal structures.

Solovyeva et al. (2023) extend the previous approach by proposing a constraint satisfaction procedure which is computationally more efficient, and can also recover from conflicts due to statistical errors.

The previous developments can find the set of possible causal structures that are consistent with the under-sampled data, but can not select among these the *correct* one. Muñoz-Benítez and Sucar (2024) proposes an approach to impute the missing data due to subsampling in order to recover a *unique causal structure*, which is incorporated in this work for prediction.

4 Time Series Prediction Using Causal Knowledge

The approach proposed in this work aims to minimize the impact of subsampling on time series data and to incorporate causal knowledge to optimize prediction. The method is structured into two stages: i) imputing missing data in the subsampled time series to recover the true causal relationships among the variables; and ii) include causal knowledge to identify the most relevant variables for predicting the target variables.

4.1 Data Imputation in Time Series Data

Imputation methods estimate missing values by leveraging the available data, resulting in a more complete dataset. Muñoz-Benítez and Sucar (2024) proposed a method to impute the missing data in a subsampled time series, based on an adversarial neural network to generate the missing values. This work demonstrates promising results in the context of causal discovery; in this paper we apply this technique for time series prediction.

4.2 Prediction Based on Causal Knowledge

We propose a feedforward neural network architecture designed to incorporate causal knowledge for predicting variables in multivariate time series data. Based on a prior causal analysis, the model identifies which variables directly affect the target variable. This information is used in a more focused, efficient, and interpretable predictive model.

The model architecture consists of a dense neural network with an input layer followed by three hidden layers, see Figure 2. The input layer includes only those variables that have been identified as causal parents of the target variable, as determined by the causal graph. The output layer is a dense layer with a single unit and linear activation, producing the predicted value of the target variable. The model learns to map combinations of causally relevant input variables to the target variable, which reduces dimensionality and simplifies the training process.

Causal knowledge is incorporated through the selection of input variables. Instead of using all available variables, the model receives only those that have been identified as direct causal parents of the target variable. This selective input reduces noise and enables more accurate predictions. Formally, the prediction of a variable X_i is modeled as:

$$X_i = f_\theta\left(\mathrm{Pa}(X_i)\right) \tag{1}$$

where $\mathrm{Pa}(X_i)$ denotes the set of causal parents of X_i, and f_θ is the function approximated by the neural network with parameters θ. This enables a more robust architecture, guided by the causal structure of the system. Figure 2 shows a time series with several observable variables $X1, ..., X10$ but only those relevant to the target variable are selected $X5, X6, X8$.

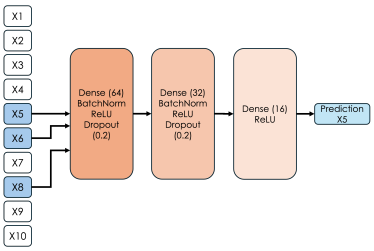


Fig. 2. Architecture of the prediction model. The relevant variable are selected and used as input for the model. The output is the target variable which is affected by its parents in the causal graph.

5 Experimental Results

5.1 Synthetic Data

As an initial experiment, we generated a synthetic dataset consisting of 10,000 samples and 10 variables. Causal relationships between variables were allowed to emerge with a minimum time lag of 2 and a maximum of 4. The probability of a causal edge between any pair of variables was set to 10%. All variable values in the dataset were sampled from a Gaussian distribution. The PCMCI algorithm (Runge, 2018) was used as causal discovery algorithm in order to obtain the causal structure used as ground truth for our approach. Figure 3(a) shows the causal structure obtained from the time series data.

To simulate the effect of subsampling in the time series, the dataset was modified to simulate observations every 2, 3, and 4 time lags. As can be seen in Figure 3, the causal structure is degraded, modifying the times at which the causal relationships are present, and even eliminating some.

The imputation approach was used to generate the missing data for the time series. These time series with imputed data were used to learn the causal structures for each subsampling scenario. The recovered causal structures are shown in Figure3(e)-(g). The recovered causal structures maintain most of the original causal relationships, although in some cases the times at which the relationships are reflected are not the same.

Incorporating causal knowledge into time-series forecasting can help to improve prediction accuracy, as only relevant variables that affect the target variable are used. Furthermore, the use of time series with imputed data minimizes the effect of subsampling on the predictions. In all experiments, a prediction horizon of one time step was used, meaning the model predicts the value X_{t+1} based on observations up to time t. To validate the accuracy of the prediction, 200 data points not observed by the model were used (Figure 4), which serve to compare the predictions of variable $X5$ using only historical data for the same variable (red), the predictions using all the variables (blue), and the predictions using only the relevant variables (green). The results show a significant improvement on the Mean Square Error (MSE) when using the variables of the causal

model learned with the imputed data against the one based on the subsampled data; and even against using all the variables. Table 1 summarizes the results for the different levels of subsampling. Increased levels of subsampling severely affect the accuracy of time series predictions; however, the predictions using the models discovered with the imputed data still have a relatively low error, significantly lower than the other models.

Level of subsampling	**Data affected by subsampling**	**Data using all variables**	**Data using relevant variables**
2	1.459	0.044	0.009
3	1.55	0.10	0.04
4	3.18	0.92	0.049

Table 1. Prediction results for a horizon of one time step, in terms of MSE, of the data affected by different levels of sub-sampling.

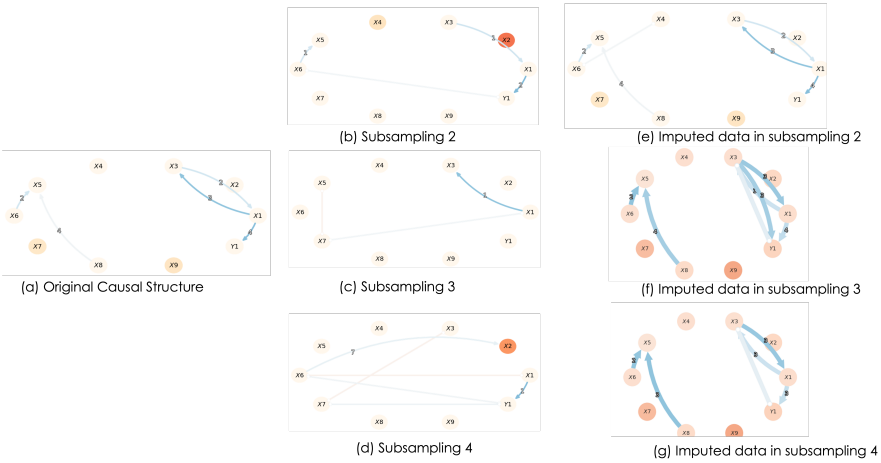


Fig. 3. Causal structures discovered for the synthetic dataset. (a) Original causal structure. (b), (c), (d) Causal structures of the dataset affected by subsampling every 2, 3 and 4 time steps, respectively. (e), (f), (g) Causal structures learned with the imputed data, corresponding to the same subsampling rates, 2, 3 and 4 time steps, respectively. The causal graphs are shown using a compact representation known as *rolled graph*, that indicates the causal links with a number corresponding to the time lags, and the width the indicates the strength of the relation.

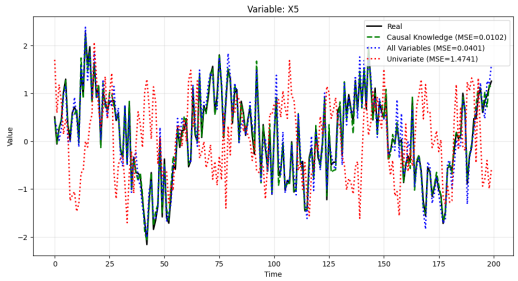


Fig. 4. Prediction of variable $X5$ for a time series affected by a subsampling rate of 2. The predictions are compared using 200 data points not seen by the model (black). If the data set affected by subsampling is used, the predictions are inaccurate and present the highest error (red). However, if causal information from the model learned from the imputed data is used, the predictions resemble the original data (green), providing the lowest MSE, even lower that using all the variables (blue). (Best seen in color.)

5.2 Real Scenario: Predicting Daily Energy Price

The proposed approach was evaluated on a real-world dataset of an energy market. It includes data from a total of 2,016 days, covering the period from January 1, 2015 to October 6, 2020 (Kozlov, 2020), from residential users in Victoria, Australia. The data collected reflected the total daily demand in MWh (demand); the recommended retail price in AUD$ MWh (RRP); daily demand at positive RRP in MWh (demand_pos_RRP) and their respectively price (RRP_positive)[1]; daily demand at negative RRP (deman_neg-RRP) and their respectively price (RRP_negative)[2]; and the dataset included weather data from temperature (min_temperature and max_temperature); solar exposure (solar_exposure) and rainfall in mm (rainfall).

The causal discovery algorithm was applied to infer the causal structure among the variables. The resulting graph, shown in Figure 5(a), highlights how the variables influence one another within the system. From the model we can observe some interesting relations. When the fraction of time with negative prices (frac_at_neg_RRP) increases, the overall average daily electricity price (RRP) tends to decrease. This is, when licensed users produce more energy than required, the price tends to decrease. Electricity prices, particularly RRP and RRP_positive, emerge as key drivers of demand patterns. This suggests that sustained periods of oversupply lead to both lower peak prices and more frequent or prolonged instances of negative pricing. The causal graph also indicates that solar exposure plays a role in shaping energy demand and electricity consumption behavior. This reflects the power of causal analysis, where causal relationships help to explain the behavior of the observed phenomenon beyond the correlation between variables.

[1] This represents the energy sold by the state.
[2] This represents the energy sold to the state by licensed users.

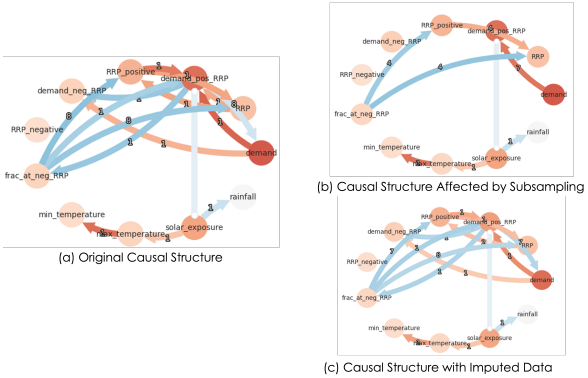


Fig. 5. Causal structure of the dataset on daily energy demand. (a) Original causal structure. (b) Causal structure affected by a subsampling rate of 2. (c) Causal structure with imputed data.

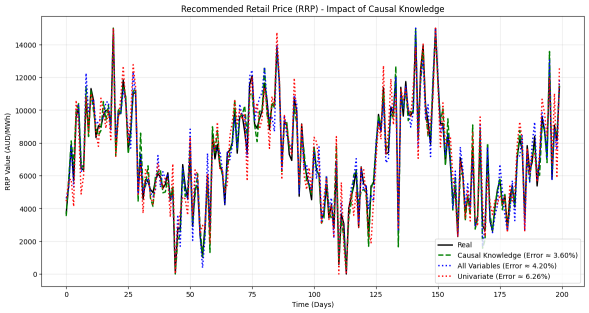


Fig. 6. Prediction of the recommended retail price (RRP) based on 200 data points not seen in training (black). The use of historical data (red) is compared against using all variables (blue) and only the related variables that affect the target variable (green). (Best seen in color.)

Based on the causal model, the relationships obtained in the causal graph were used for prediction, with the recommended retail price (RRP) as the target variable. In this way, by using only the relevant variables that affect the target variable, it was possible to obtain a prediction much closer to the real values. As can be seen in Figure 6, using historical data for the same variable generates an accuracy with a 6.26% MSE (red). If all the variables are used, this error is reduced to 4.20% (blue). When using only the direct causal influences, the error is even lower, 3.6% (green).

The original dataset was collected daily, so to evaluate the impact of subsampling, we consider a scenario in which the collected data is affected by a subsampling rate of 2. As shown in Figure 5(b), most of the causal relations are lost and the causal relations that remain appear at incorrect time lags compared

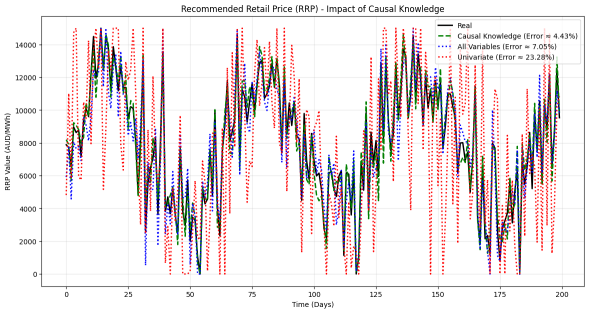


Fig. 7. Prediction of the recommended retail price (RRP) using the dataset affected by subsampling. The results from the model learned from the data affected by subsampling (red) are very different to the original values (black). Based on the causal model obtained with the imputed data, taking into account all variables (blue) and only the related variables that affect the target variable (green), a significant reduction in the MSE is obtained. (Best seen in color.)

to the original structure. However, the causal structure obtained from the imputed data, Figure 5(c), is very close to the original structure. This structure was used for selecting the relevant variables for predicting RRP. Figure 7 shows the results using the data affected by subsampling, increasing the MSE to 23.7% (uni-variate). However, incorporating the imputed data the error decreases to 7% using all the variables and 5% using only the causally-relevant information. This demonstrates the advantage of the proposed approach in minimizing the effect of subsampling, and of using causal knowledge to improve predictions in complex, real-world scenarios.

6 Conclusions

This work proposes a novel methodology for time series prediction based on causal discovery, particularly in contexts where the data is affected by subsampling. By using adversarial neural networks for data imputation, the proposed approach reconstructs a more complete time series, enabling accurate causal discovery. The resulting causal graph is then used to focus only on the causally relevant variables for prediction. The approach was validated on synthetic data and a real-world energy market, showing that predictions based on causal knowledge consistently outperform those based on historical data of the predicted variable, and even those based on all the variables. The proposed method is particularly effective in mitigating the negative impact of subsampling on causal discovery and prediction. As future work, we will test our approach on longer term predictions, and in other application domains.

Bibliography

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

Faruque, A. et al. (2024). Ts-causalnn: A deep learning approach for causal discovery in non-stationary time series. *arXiv preprint arXiv:2404.01466*.

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

Kozlov, A. (2020). Daily electricity price and demand data.

Kristjanpoller, W. et al. (2025). A framework for integrating causal knowledge in financial time series forecasting. *Journal of Financial Econometrics*. To appear.

Lawrence, A., Kaiser, M., Sampaio, R., and 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*.

Li, H. et al. (2021). Domain adaptation in time series forecasting with stable causal structures. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9):3245–3255.

Lim, B. and Zohren, S. (2021). Time series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209.

Muñoz-Benítez, J. and Sucar, L. E. (2024). Data imputation with adversarial neural networks for causal discovery from subsampled time series. In Calvo, H., Martínez-Villaseñor, L., and Ponce, H., editors, *Advances in Soft Computing. Proceedings of 22nd Mexican International Conference on Artificial Intelligence*, pages 39–51, Cham. Springer Nature Switzerland.

Murphy, K. P. (2002). Dynamic bayesian networks: representation, inference and learning.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Runge, J. (2018). Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310.

Solovyeva, K., Danks, D., Abavisani, M., and Plis, S. (2023). Causal learning through deliberate undersampling. In *2nd Conference on Causal Learning and Reasoning*.

Tang, X. (2024). Trading with time series causal discovery: An empirical study. *arXiv preprint arXiv:2408.15846*.

Tierney, L. et al. (2023). Multivariate bayesian dynamic models for causal prediction in time series. *Journal of Causal Inference*, 11:135–152.

Wen, Q., Wu, Y., Sun, L., Xu, J., Wang, C., and Long, G. (2023). Transformers in time series: A survey. *ACM Computing Surveys (CSUR)*, 55(9):1–36.

Section 3

Applications

Clustering-based Causality Analysis of GDP and Financing levels nexus

Roberto Flores-Nava[1] and Edgar Roman-Rangel[1][0000−0002−0590−1698]

[1] Instituto Tecnológico Autónomo de México
[2] {rflore53,edgar.roman}@itam.mx

Abstract. Causal analysis and discovery from observational data is highly relevant for understanding how economic drivers like productivity ones can have an effect on key indicators like financing, or viceversa. While a single causality approach can provide important insights, there are cases when it might be insufficient to capture the complexity of the real effects in the economic sector. For this reason, it becomes relevant to complement traditional models with other techniques such as Machine Learning ones, which allow a richer and more flexible exploration of patterns. Combining these two elements is an innovative approach to explore, understand, contrast and challenge different existing theories, and to provide a different perspective and interpretation of them. This paper then integrates both perspectives by first implementing an unsupervised analysis with a clustering methodology to segment similar countries into groups based on their macroeconomic structures and trajectories, followed by applying Granger causality tests to evaluate the dynamic-causal relationship between economic growth and private sector financing within each cluster. The results show that the dynamic-causal effect is not universal, enriching the way that economic theory is understood and applied.

Keywords: Economic growth · Private sector credit · Unsupervised learning · Clustering · Granger causality

1 Introduction

The relationship between financial development and economic growth has been widely studied and has been a subject of debate across different theorists [1,2]. Two perspectives are the most representative: while some authors argue that the economic growth follows the levels on financing, driven by the intuition that with more available credits, the producers of goods and services could produce more, then the Gross Domestic Product (GDP) would be increased too; some others defend a position which states that favorable economic conditions, including growth, must be seen previous to see financial companies motivated to increase their loans.

From the first approach, one of the most influential contributions come from Schumpeter [1], who proposed that financial institutions play a crucial role by

channeling resources into innovative projects, which in turn drive long-term economic growth.

In contrast, Robinson [2] claimed that it is the dynamic economic environment what drives the conditions for financial development, suggesting the causal direction runs the other way around. Within his analysis, he emphasize that in an evironment of sustained economic growth and certainty on future perspectives, the creation of new companies and projects is encouraged, which increases both the demand for and supply of credit.

Later on, empirical evidence reinforced both sides of the debate. King and Levine [6] showed that countries with more developed financial systems tend to experience faster growth, while Beck, Levine, and Loayza [8] found that private credit is positively linked to productivity improvements. Still, these studies often treat countries as a homogeneous group, which limits the ability to capture structural differences across economies.

This paper builds on the idea that the financing and economic growth relationship is not universal but context-dependent, which means that one variable is caused by the other depending on the type of economy that it is observed. To test this hypothesis, we apply unsupervised learning techniques to segment countries into groups with similar macroeconomic trajectories and structural conditions. Within each cluster, we then analyze correlations and conduct Granger causality tests to identify the dynamic direction of the relationship between private credit and GDP. By combining clustering with causal testing, the study provides evidence that the financing levels and economic growth nexus varies across contexts and cannot be explained by a single theoretical framework; instead, this should be studied differently depending on the economic context. Concretely, the contribution of this paper is to enrich the economic theory analysis by providing a novel approach that combines, for the first time, both macroeconomic-structural conditions with clustering and causal analyses.

2 Related Work

The debate, on whether finance leads or follows economic growth, has been open for more than a century. Schumpeter [1] saw financial systems as engines of innovation, while Robinson [2] argued that it is growth itself that creates the demand for finance.

Later, some early empirical works tried to measure this link more systematically. Goldsmith [3] showed that countries with deeper financial systems also tended to grow faster. McKinnon [4] and Shaw [5] added to this view with the "financial repression" hypothesis, suggesting that restrictive policies slow down investment and that liberalization could instead boost economic activity.

In the 1990s, the debate gained momentum with broader cross-country evidence. King and Levine [6] showed that the financial depth is strongly associated with long-run growth. Levine [7] reviewed the literature and outlined the main channels linking finance and growth, helping to consolidate the previous results. Then, Beck, Levine, and Loayza [8] also showed that private credit matters for

productivity and GDP, while recognizing at the same time that causality could go both ways.

More recently, the perspective has moved to heterogeneity in the approaches. Mhadhbi et al. [9] used a bootstrap panel Granger approach and found that the direction of causality differs across countries. Abbas et al. [10] added a time–frequency perspective, showing that results can change in the short versus the long run. Kchikeche and Mafamane [11], looking at Morocco, confirmed that causality can even shift depending on the horizon: short-run effects go from credit to growth, but in the long run feedback prevails.

Although our approach shares conceptual similarities with Mhadhbi et al. [9], our approach differs in an important way. Instead of testing heterogeneity directly on a pooled panel, we first segment countries into groups with similar macroeconomic structures and then apply Granger causality within each group. This allows us to control explicitly for cross-country structural differences before evaluating dynamic relationships. As a result, our findings complement and extend it by showing that heterogeneity is not only present, but also that it follows clear macroeconomic patterns derived from the clustering step.

Overall, these contributions reveal two key points: first, that the finance–economic growth relationship is not universal; and second, that both the empirical method and the context are important factors in the results. However, none of these consider an approach that combines a segmentation of countries with causal testing inside each group. This is precisely what our study contributes: by clustering economies with similar macroeconomic conditions and then applying Granger causality, we show that the direction of the finance-economic growth link depends on the type of economy, not on a single global pattern.

3 Data and Exploratory Analysis

3.1 Data Sources

This study takes into account three widely recognized international databases. Quarterly GDP was obtained from the IMF's Quarterly National Accounts [12]. Quarterly Private sector credit (as a percentage of GDP) was taken from the Bank for International Settlements (BIS) [13]. Finally, annual indicators such as credit, GDP variation, per capita GDP, inflation and unemployment were sourced from the World Development Indicators (World Bank) [14]. These sources ensure consistency and comparability across countries. The last source of information with annual data was leveraged to perform the clustering analysis, while the other two sources with quarterly data were leveraged for the causal analysis because of their more frequent representation. As a result, the time series leveraged for the cluster analysis contain the annual data from 2005 to 2022, while the data for the causality tests are represented from 1Q2005 (i.e., first quarter of 2025) to 3Q2024.

3.2 Descriptive Statistics

Taking into account the availability of data across the three sources of information, and driven by three main criteria: i) keep the bigger amount of countries with consistent data; ii)capture representative periods with different economic and financial conditions globally; and iii) maintain a sufficient number of observations to apply statistical tests, this analysis focuses on 30 countries: Australia, Austria, Belgium, Brazil, Chile, China, Colombia, Denmark, Finland, France, Germany, Greece, Hong Kong, Hungary, Ireland, Israel, Italy, Japan, Luxembourg, Malaysia, Mexico, Netherlands, Norway, Poland, Portugal, South Africa, Spain, Sweden, United Kingdom, and United States.

3.3 Descriptive Statistics

The dataset for clustering analysis covers 30 countries between 2005 and 2022. Table 1 reports descriptive statistics for the main variables. In addition, in Figure 1, it is possible to note that the variables show a high heterogeneity, for example, in the variables of private credit or inflation, where there are outliers and large dispersions.

Table 1: Descriptive statistics of main variables

Variable	Min	Mean	Max	Std. Dev.
GDP growth (%)	-10.94	2.27	24.62	3.64
Per capita GDP growth (%)	-11.37	1.56	23.44	3.58
Inflation (%)	-4.45	2.52	14.61	2.27
Unemployment (%)	2.35	7.89	34.01	5.25
Private lending (% of GDP)	15.31	106.36	264.44	47.12

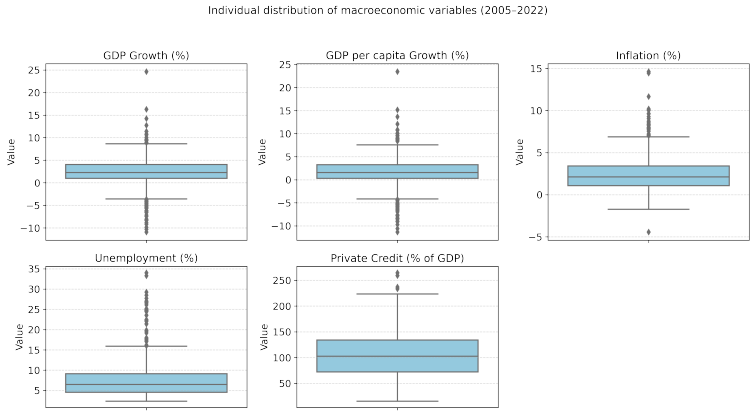


Fig. 1: Distribution of macro economic variables for cluster analysis (2005–2022).

3.4 Preprocessing

Annual data from the World Bank used in the clustering exercise were employed as published, without additional transformations. For the causal analysis, however, quarterly GDP and private sector credit were adjusted to obtain comparable measures in terms of year-on-year growth rates. This required first converting credit, originally expressed as a percentage of GDP, into local currency values using IMF GDP series as the base. Both GDP and credit were then transformed into annual growth rates by comparing each quarter with the same quarter of the previous year.

After this step, additional datasets were created by including up to four quarterly lags for each variable, ensuring that every observation contained both the current value and its lagged terms.

3.5 Target Variables and Data Organization

For the clustering analysis, the objective was to group countries into homogeneous clusters based on their structural macroeconomic profiles. Five annual indicators from the World Bank were selected for this purpose: GDP growth, GDP per capita growth, inflation, unemployment, and private sector credit, as defined above.

In contrast, the causal analysis required explicit target variables, which for this study are represented as Y. Quarterly GDP and private credit were transformed into year-on-year growth rates, allowing comparability across countries of different sizes and reducing the effect of trending levels. Two specifications were tested:

- Annual percentage change (Δ %) of Credit (C) at time t as a function of the lagged annual percental change of GDP at time (quarter) *t-1*, *t-2*, *t-3*, *t-4*:

$$\Delta\% \; C_t = f\left(\Delta\% \; PIB_{t-1},\; \Delta\% \; PIB_{t-2},\; \Delta\% \; PIB_{t-3},\; \Delta\% \; PIB_{t-4}\right) \quad (1)$$

- Annual percentage change (Δ %) of GDP at time t as a function of the lagged annual percentage change of Credit (C) at time (quarter) *t-1*, *t-2*, *t-3*, *t-4*:

$$\Delta\% \; PIB_t = g\left(\Delta\% \; C_{t-1},\; \Delta\% \; C_{t-2},\; \Delta\% \; C_{t-3},\; \Delta\% \; C_{t-4}\right) \quad (2)$$

where $f(\cdot)$ and $g(\cdot)$ are distinct functions representing the causal relationship in each direction. In the first case, the objective is to evaluate whether past values of GDP growth help explain credit growth (C_t); in the second, whether past values of credit growth help explain GDP growth (GDP_t). This separation emphasizes that the functional form and significance of lagged effects may differ depending on the causal direction being tested.

The datasets were structured according to the type of analysis. For clustering, a wide panel was created, with each row representing a country and columns capturing annual values of the five macroeconomic indicators from 2005 to 2022. For causal testing, longitudinal datasets were built for each cluster, where each row represented a country–quarter observation with the annual growth rate of the

target variable and its corresponding lags. In total, six datasets were generated (three clusters, two directions of causality), each providing the necessary structure for the subsequent econometric models.

4 Methodology

The methodological pipeline consists of three main steps: (i) unsupervised learning by clustering analysis of countries into homogeneous groups, (ii) correlation analysis between GDP and credit growth with quarterly lags; and (iii) Granger causality tests to evaluate dynamic interactions. Figure 2 illustrates the overall process.

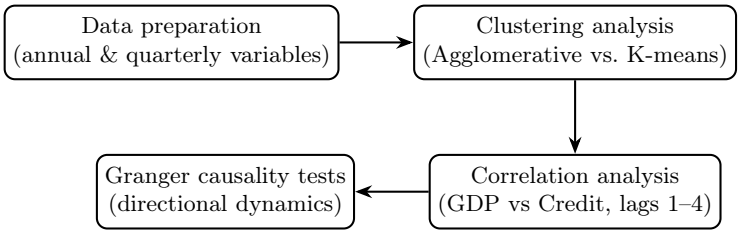


Fig. 2: Methodological pipeline.

4.1 Clustering analysis

Clustering algorithms were applied to group countries according to their structural macroeconomic characteristics. Two algorithms were tested: K-means and Agglomerative Clustering [17,18]. Performance was assessed using the Silhouette Score [19], complemented by dimensionality reduction visualizations [15,16]. The final grouping was obtained with Agglomerative Clustering, as it provided both statistical consistency and economic interpretability.

It's worth to notice that countries were not clustered based on raw time-series trajectories but on their annual macroeconomic indicators arranged as a wide panel, where each country is represented by a vector containing its values from 2005 to 2022. Thus, each country is represented by a multi-dimensional feature vector rather than by its full time-series sequence. For this reason, no DTW was required; Euclidean distance on the standardized panel was used.

Then, boxplots of GDP growth, GDP per capita growth, inflation, unemployment, and private credit were used to characterize each cluster, highlighting distinctive macroeconomic patterns.

4.2 Correlation analysis

After clustering, pairwise correlations were computed between GDP and credit growth, considering up to four quarterly lags. This step served as an initial

exploration to detect potential dynamic associations. Heatmaps were used to visualize correlation patterns by cluster.

4.3 Granger Causality Tests

Correlation alone cannot establish directional relationships. Therefore, Granger causality tests [20] were performed to evaluate whether past values of one variable improve the prediction of the other. Two competing models are estimated:

$$\text{Restricted model (no } X\text{):} \quad Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + u_t, \tag{3}$$

$$\text{Unrestricted model (with } X\text{):} \quad Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{i=1}^{p} \gamma_i X_{t-i} + \varepsilon_t. \tag{4}$$

where:

- Y_t: value of the dependent series at time t
- X_{t-i}: value of the explanatory series X at lag i (period $t-i$)
- Y_{t-i}: lagged value of the dependent series Y at lag i (period $t-i$)
- α_0, β_0: intercept terms of each model
- α_i, β_i: coefficients associated with the lagged values of Y
- γ_i: coefficients capturing the effect of lagged values of X on Y
- u_t, ε_t: stochastic error terms of each model
- p: number of lags considered in the specification
- t: time index of the series, with $t = 1, 2, \ldots, T$

The null hypothesis of no Granger causality from X to Y is:

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_p = 0$$

against the alternative that at least one $\gamma_i \neq 0$. The test is conducted using an F-statistic comparing restricted and unrestricted models.

The procedure was applied in both directions (GDP $\rightarrow$ credit and credit $\rightarrow$ GDP) within each cluster, allowing the analysis to capture whether the dynamic relationship depends on the type of economy.

It is important to note that Granger causality evaluates whether past values of X improve the prediction of Y relative to a restricted model that includes only lagged values of Y. In our context, this corresponds to comparing the unrestricted specification in Equation (4) with the restricted model in Equation (3). Thus, the test captures predictive direction based on temporal precedence, but not structural or mechanistic causality.

5 Results

5.1 Clustering analysis

We evaluated two algorithms: K-means and Agglomerative (both with $k = 2$ to $k = 5$), on the five annual macro indicators (GDP growth, GDP per capita growth, inflation, unemployment, private credit). Both techniques achieved close Silhouette scores for each k, as shown in Table 2.

Table 2: Silhouette scores by clustering technique ($k = 2$–5).

k	Agglomerative	KMeans
2	0.4506	0.4237
3	0.4081	0.4048
4	0.3274	0.3622
5	0.3522	0.3667

Although the highest Silhouette score values were obtained for $k = 2$, having only two clusters was considered too restrictive to capture the heterogeneity, and difficult to describe intuitively. In comparison, for $k = 3$, the Silhouette scores remained reasonably high while the country distribution across clusters reflected clearer and more interpretable macroeconomic profiles. Therefore, $k = 3$ was selected as the most appropriate parameter.

Regarding the choice of algorithm, K-means provided acceptable results but showed less well-defined boundaries between groups. In contrast, the Agglomerative produced clusters that were more coherent and aligned with expected macroeconomic behavior, as confirmed both numerically and visually. The hierarchical structure highlighted in the Ward-linkage dendrogram (Figure 3) supported the robustness of the solution with three clusters; in addition, the PCA projection showed in Figure 4 confirms a good segmentation. Consequently, Agglomerative clustering was chosen as the preferred technique.

Figure 5 highlights the economic interpretation of each group: Private credit represents the most discriminative variable with more dispersion between clusters: Cluster 2 shows very high levels (aprox. 180% of GDP, on average), Cluster 3 the lowest, and Cluster 1 an intermediate between the other two. Inflation and unemployment are lowest in Cluster 2 and highest in Cluster 3, while economic growth indicators (total and per capita) remain less dispersed across groups.

Clustering-based Causality Analysis of GDP and Financing levels nexus

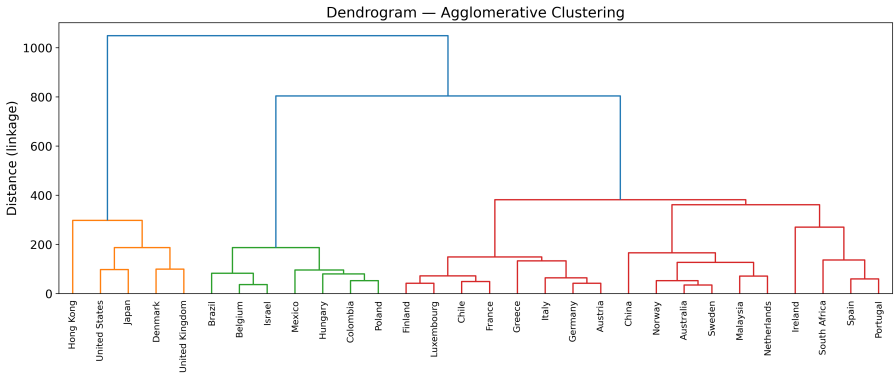


Fig. 3: Ward-linkage dendrogram for the Agglomerative solution

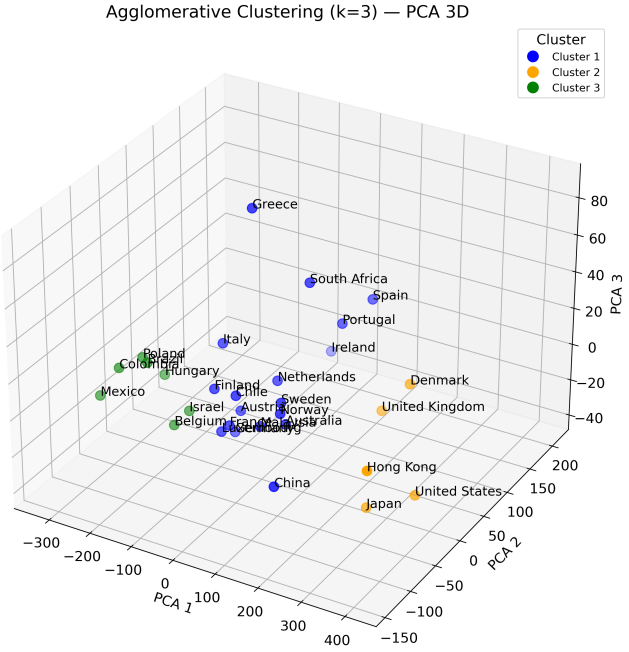


Fig. 4: Agglomerative clustering visualized with 3D PCA

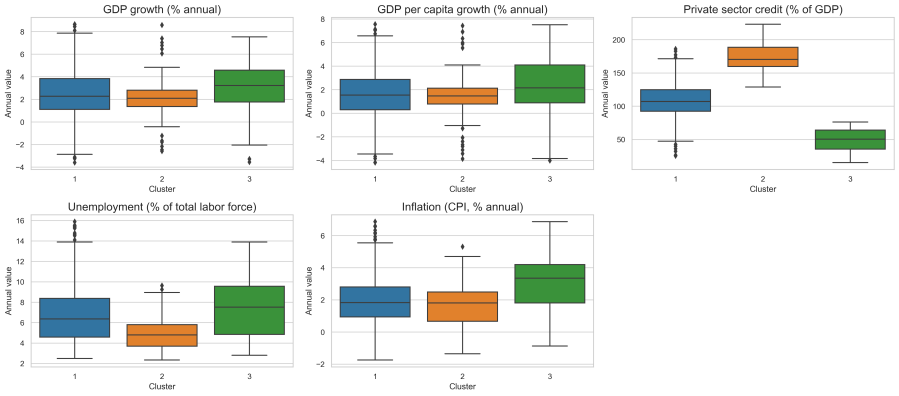


Fig. 5: Distribution of macroeconomic indicators by cluster (outliers smoothed)

The clustering allowed to then identify three groups of coutries with similar patterns inside each group but with differences between each other. The cluster 1, that we named "Developed and in transition economically stable" gathers economies mostly developed with others which are in advanced transition, which are characterized by a moderate growth, contained levels of inflation and unemployment, and an intermediate financial deepening. The cluster 2, "Highly developed, competitive and stable", groups the most consolidated economies, with a low inflation and unemployment, and high credit access. Finally, the cluster 3, "Emerging and dynamic intermediate markets with macroeconomic risk", represents economies with a major growth dynamism, but with high levels of inflation and unemployment at the same time with lower levels of financing. This segmentation provides the structural backdrop for the dynamic analysis.

5.2 Correlation analysis (GDP vs. Credit)

Pearson correlations between year-on-year GDP growth and credit growth with lags 1–4 in both directions (lagged GDP with current credit; lagged credit with current GDP) were computed. Figure 6 shows heatmaps by cluster.

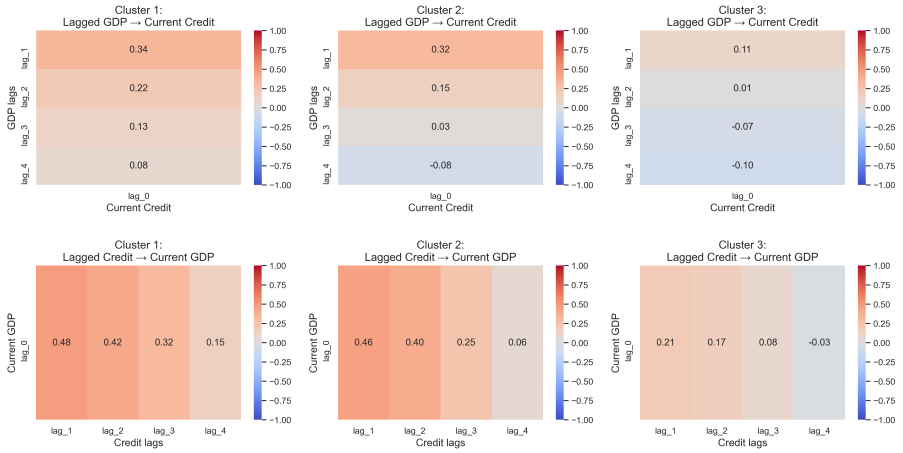


Fig. 6: Correlation heatmaps by cluster. Top: lagged GDP vs. current credit. Bottom: lagged credit vs. current GDP.

The previous evidence shows that correlation magnitudes differ across clusters and tend to be stronger at shorter lags (1–2 quarters). This motivates formal directional tests. However, simple correlations suggest dynamic association but cannot establish direction nor control for own-lag persistence, hence the move to Granger tests.

5.3 Granger Causality Tests

For each country, we performed Granger tests in both directions using quarterly lags $p \in \{1, 2, 3, 4\}$, based on the models defined in equations (3) and (4), and tested $H_0 : \gamma_1 = \cdots = \gamma_p = 0$ with a F-test at the 5% significance level [20].

The results were summarized by cluster using the following rules: (i) *Unidirectional* if H_0 is rejected in $\alpha = 0.05$ in one direction only (at least one significant lag in one direction by the F test and none in the opposite direction); (ii) *Bidirectional* if H_0 is rejected at $\alpha = 0.05$ in both directions (at least one significant lag in both directions); and (iii) *No evidence* if H_0 is not rejected at $\alpha = 0.05$ in either direction.

Table 3: Granger causality outcomes by cluster (number of countries).

Cluster	**Bidirectional**	**Credit → GDP**	**GDP → Credit**	**No evidence**
Developed and in transition economically stable	3	11	2	1
Highly developed, competitive and stable	4	1	0	0
Emerging and dynamic intermediate markets with macroeconomic risk	3	2	0	2

According to the Table 3, it is clear that cluster of "Developed and in transition economically stable" concentrates bidirectional causality, with this both variables can cause the other. The cluster of "Highly developed, competitive and stable" countries is mostly *Credit* → *GDP*. While, finally, cluster "Emerging and dynamic intermediate markets with macroeconomic risk" exhibits mixed patterns with frequent bidirectionality, for which it is more difficult to reach a sigle direction or rule of causality.

6 Conclusions

The paper addresses the complex and context-dependent relationship between economic growth and private credit. While much of the literature assumes a universal or unidirectional causality, this study examines whether this causality depends on country heterogeneity by grouping economies with similar macroeconomic characteristics.

Our contribution lies in combining unsupervised clustering with Granger causality tests. Countries were first segmented by five macroeconomic indicators, and then, causal dynamics between GDP and credit growth were analyzed using quarterly data. At the dynamic level, we find that causality patterns vary by cluster: in advanced economies, credit frequently precedes growth; in intermediate or emerging economies, the relationship is more often bidirectional, reflecting feedback loops; while in some cases in the less advanced economies, no causal evidence was found, which might imply that economic growth and financing are complex variables, difficult to determine one with the other.

These findings suggest that the finance–growth nexus cannot be generalized but must be interpreted in each structural context. For policymakers, financial deepening should be tailored to local conditions, while researchers can benefit from combining machine learning and causal inference, to generate new insights into long-standing debates in economics. Future research could expand the dataset, explore alternative clustering methods or structural causality methods, or incorporate additional macro-financial variables.

References

1. Schumpeter, J.A.: *The Theory of Economic Development*. Cambridge, MA: Harvard University Press (1911)
2. Robinson, J.: *Essays in the Theory of Economic Growth*. London: Macmillan (1962)
3. Goldsmith, R.: *Financial Structure and Development*. New Haven: Yale University Press (1969)
4. McKinnon, R.I.: *Money and Capital in Economic Development*. Washington, D.C.: Brookings Institution (1973)
5. Shaw, E.S.: *Financial Deepening in Economic Development*. New York: Oxford University Press (1973)
6. King, R.G., Levine, R.: Finance and growth: Schumpeter might be right. *Quarterly Journal of Economics* **108**(3), 717–737 (1993)

7. Levine, R.: Financial development and economic growth: Views and agenda. *Journal of Economic Literature* **35**(2), 688–726 (1997)
8. Beck, T., Levine, R., Loayza, N.: Finance and the sources of growth. *Journal of Financial Economics* **58**(1–2), 261–300 (2000)
9. Mhadhbi, K., Terzi, C., Bouchrika, A.: Banking sector development and economic growth in developing countries: A bootstrap panel Granger causality analysis. *Journal of Economics and Development* **22**(3), 271–288 (2020)
10. Abbas, E., et al.: Frequency matters: A wavelet-based investigation of the financial development–economic growth nexus. In: *Proceedings of ICBMASS 2025*, NUST, Islamabad (forthcoming, 2025)
11. Kchikeche, A., Mafamane, D.: Examining the interactions of economic growth and bank credit to the private sector in Morocco: A causality analysis. *International Journal of Economics and Finance Studies* **15**(2), 45–60 (2023)
12. International Monetary Fund (IMF): Quarterly National Accounts. https://data.imf.org/, last accessed 2025-09-21
13. Bank for International Settlements (BIS): Credit to the non-financial sector. https://www.bis.org/statistics/totcredit.htm, last accessed 2025-09-21
14. World Bank: World Development Indicators (WDI). https://databank.worldbank.org/source/world-development-indicators, last accessed 2025-09-21
15. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn. New York: Springer (2009)
16. Bishop, C.M.: *Pattern Recognition and Machine Learning*. New York: Springer (2006)
17. Xu, D., Tian, Y.: A comprehensive survey of clustering algorithms. *Annals of Data Science* **2**(2), 165–193 (2015)
18. Everitt, B.S., Landau, S., Leese, M., Stahl, D.: *Cluster Analysis*. 5th edn. Chichester: Wiley (2011)
19. Kaufman, L., Rousseeuw, P.J.: *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: Wiley (2009)
20. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**(3), 424–438 (1969)
21. Pedregosa, F., et al.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)

Causal inference applied to the calculation of insulin bolus in patients with type 1 diabetes using the GRaSP algorithm.

Contreras-Jiménez Rocío[1][0009-0006-7184-3859], Olivares-Rojas Juan Carlos[2][0000-0001-5302-1786], Téllez-Anguiano Adriana del C.[3][0000-0002-0945-2076], Alcaráz-Chávez J. Eduardo[4][0009-0004-6514-4571], Gutiérrez Gnecchi J. Antonio[5][0000-0001-7898-604X] and Reyes-Archundia Enrique[6][0000-0003-3374-0059]

TecNM/Instituto Tecnológico de Morelia, Morelia, Michoacán. 58120, México
[1]rocio.cj@morelia.tecnm.mx

Abstract. This work presents the development of an intelligent system based on causal inference to optimize the insulin bolus calculation for patients with type 1 diabetes. The GRaSP (Greedy Relaxations of the Sparsest Permutation) algorithm was applied to the HUPA-UCM 2024 dataset to identify causal relationships between physiological and behavioral variables. Different data transformations were tested, and medical knowledge constraints were integrated to improve the consistency of the inferred graphs. Preliminary results show structural improvements with the applied constraints, although inconsistencies due to unobserved confounders and non-normality in the data persist. This study demonstrates the potential of causal models to support personalized insulin calculation and glycemic control.

Keywords: Artificial Intelligence, Causal Inference, GRaSP, Health Technology, Insulin Bolus Calculation, Type 1 Diabetes.

1 Introduction

Type 1 diabetes mellitus (T1DM) is a chronic, non-preventable, and incurable disease that requires strict control of blood glucose levels [1]. Insulin bolus calculation is one of the pillars of treatment; however, current methods rely on empirical or predictive approximations that do not account for causal relationships among the variables involved. Accurate insulin bolus calculation in patients with type 1 diabetes mellitus (T1DM) is one of the most complex challenges in glycemic control. Applying a higher-than-necessary bolus can lead to hypoglycemia with fatal consequences, while using an insufficient bolus can lead to hyperglycemia, which, if persistent, will cause long-term complications such as diabetic retinopathy, kidney failure, heart disease or stroke, diabetic neuropathy, limb loss, and others [2]. This implies high costs for patients and their families, the public health system, and the patient's economically active life [3]. Traditionally, dose adjustment has been based on empirical equations that consider the

carbohydrate-insulin ratio, patient sensitivity, and pre-meal glucose levels. However, these approaches lack adaptability to daily physiological variability and the multiple external factors that influence glycemic response.

Over the past decade, several studies have incorporated artificial intelligence (AI) techniques to automate and personalize bolus calculation. Among them, Noaro, Capon et al used quadratic LASSO regression (LASSO\(_Q\)) to improve dose estimation, reducing the frequency of hypoglycemia in simulations [4]; Zhu, Taiyu et al, implemented deep reinforcement learning (DRL) to optimize prandial dosing, increasing time in range (TIR) in adults and adolescents [5]; and Kalita et al. [6] developed InsNET, a deep neural network capable of predicting basal and bolus doses with high accuracy. Although these systems have shown promising results, they share a fundamental limitation: they are black boxes. These correlational models prioritize predictive accuracy over interpretability, understanding of underlying physiological mechanisms, and identification of the causes of post-pandemic glucose, which depend on the patient's context and the insulin bolus administered.

In this context, the GRaSP (Greedy Relaxations of the Sparsest Permutation) algorithm [7] emerges as an alternative based on causal inference capable of identifying structural relationships between clinical, behavioral, and metabolic variables. Unlike traditional AI methods, GRaSP not only predicts but also uncovers direct causal dependencies between factors such as preprandial glucose, carbohydrate intake, exercise, basal and bolus insulin doses, providing a solid theoretical basis for clinical decision-making.

The GRaSP algorithm is a causal discovery method that seeks to identify the most parsimonious directed acyclic graph (DAG) structure from observational data. Unlike other conditional-independence-based algorithms, GRaSP uses a permutation search strategy and a tuck operation, thereby relaxing faithfulness assumptions [7].

In healthcare settings, this algorithm facilitates the identification of underlying mechanisms between physiological and behavioral variables. Causal models can outperform correlational approaches in interpreting complex clinical processes [8], and glucose behavior in a patient with type 1 diabetes is particularly complex, involving many variables. Unfortunately, not all of them can be monitored, but the more variables that are considered, the more accurate the calculation of the insulin bolus required by the patient will be.

In biomedical contexts, where data are high-dimensional and dependencies are complex, GRaSP is an interesting candidate for causal network inference (omics, EHRs, physiological signals). However, its use must be accompanied by robust validation and caution regarding unobserved confounders and sample limitations. Available implementations facilitate their empirical evaluation on clinical datasets [9].

The application of the GRaSP algorithm to insulin bolus calculation offers several contributions:

- Clinical interpretability, because the generated directed acyclic graphs (DAGs) allow visualization of how each variable influences postprandial glucose levels, making it easier to understand the model in a clinical setting.

- Physiological personalization: By modeling individual relationships, the system can adjust the dose according to the patient's specific sensitivity or insulin ratio, and daily context, as in the dataset used with the variables: date, time, exercise (steps), carbohydrates, preprandial glucose, insulin bolus, postprandial glucose, heart rate, and other tracking variables.
- Hybrid integration: The causal structure can be combined with predictive models like XGBoost or neural networks to improve robustness and reduce errors due to spurious correlations.
- Causal effect estimation: In this work, we are using libraries such as DoWhy or EconML, with average treatment effects (ATE) and conditional effects (CATE) for the bolus, then the future glucose can be calculated, contributing to more rational and safer optimization.
- Clinical potential: The causal-interpretive approach could be integrated into automated (closed-loop) insulin infusion systems to strengthen the transparency, traceability, and medical oversight of the algorithmic decision-making process.

This work seeks to apply the GRaSP algorithm within a causal artificial intelligence framework to generate personalized, clinically consistent insulin dosing recommendations.

2 Theoretical framework

GRaSP is a family of causal discovery algorithms that explore the space of variable permutations to construct parsimonious DAGs and use local operations: tuck to iterate toward more parsimonious solutions. GRaSP introduces a hierarchy of relaxations: GRaSP 0, 1, 2. This offer point-consistency guarantees under progressively weaker assumptions than faithfulness, and benchmarks demonstrate good accuracy and scalability compared to other AI methods [7].

GRaSP explores the permutation space of variables; given a permutation, it constructs a "sparse" DAG consistent with that ordering and uses local operations —tucks and swaps — to traverse the permutation space, searching for permutations that yield more parsimonious DAGs. This combines the advantages of ordering-based and score-based methods, GRaSP comes with point consistency guarantees under the assumptions established by the authors; furthermore, in simulations, the algorithm demonstrated competitive performance against contemporary methods in terms of accuracy and scalability for more than 100 variables, GRaSP is designed to be computationally efficient compared to exact graph searches; practical implementations and utilities have also been published (official repository and documentation in benchmarking tools and causal discovery libraries), [7] [9].

2.1 Comparison with other causal discovery approaches

Against methods based on independence tests (e.g., Peter Clark algorithm, Fast Causal Inference): GRaSP is score- or ordering-based, so it is generally more robust to Causal Inference test errors in high-dimensional or complex dependency scenarios. However, this depends on the choice of score and the conditional family model [7]

Compared with other score-based methods like Fast Greedy Search (FGES), Generalized Sequential Pattern (GSP), or NOTEARS algorithms. In benchmarks, the more relaxed version of GRaSP was competitive or superior in several scenarios, especially when the graph was dense; however, relative performance can vary depending on sample size, noise, and the presence of deterministic or nonlinear correlations. Additional evaluations show that methods such as Greedy Fast Causal Inference (GFCI), FGES, or GRaSP can produce dense graphs in specific temporal or autocorrelated scenarios. [7]

2.2 Healthcare Applications — Where is GRaSP useful?

Although GRaSP was proposed in the general context of causal discovery, its applicability to healthcare is promising in several domains:

- Biomedical/omics network inference (genes, metabolites, proteins): The need to identify causal relationships in high-dimensional data makes permutation-based search attractive, and variants of the GRaSP/GRASP appear in biological network inference work. Furthermore, GRaSP may be part of the set of algorithms evaluated for constructing causal/functional graphs in omics [9] [10].
- EHR and clinical observational studies: GRaSP is a candidate for uncovering causal dependencies between treatments, biomarkers, and clinical outcomes in large electronic databases (due to its scalability compared to dense graphs); its use has been reported or included in causal discovery pipelines in recent studies applying algorithms to complex data (e.g., studies from 2024–2025 that integrate GRaSP into comparative analyses or analysis workflows) [11].
- Analysis of biomedical time series and physiological signals: Although GRaSP was designed for contemporaneous variables, some extensions and comparisons include it to estimate contemporaneous relationships (e.g., in temporal causal discovery / contemporary edges studies), but there are caveats — specialized temporal methods can outperform GRaSP if the variables or scoring are not tailored [12].
- Published cases and preprints: There are preprints and recent works (2024–2025) that already employ GRaSP as part of their causal discovery pipeline for applied problems (e.g., papers from 2025 that use it in interaction analysis contexts and in works combining causal discovery and counterfactual reasoning). These examples show that the method is entering practical applications, including domains close to medicine and digital health [11].

Causal inference applied to the calculation of insulin bolus in patients with type 1 diabetes using the GRaSP algorithm.

2.3 Practical recommendations for applying GRaSP to health problems

- Preprocessing: imputation and treatment of categorical and continuous variables, standardization and transformation of non-Gaussian variables, if the score assumes normality. Control and labeling of times and lagged variables.
- Selection of GRaSP variants: Test the variants of the algorithm GraSP 0,1,2 and compare them with FGES and GSP; the more relaxed variant is usually more robust if you suspect faithfulness violations [7].
- Evaluation and Validation: Use simulations with plausible structures to assess sensitivity/specificity. Contrast findings with expert knowledge and with interventions in this case; we expected to be able to conduct a trial with real patients and validation by endocrinology doctors.
- Software and reproducibility: Implementations and materials from the paper are available in the repository and code, and GRaSP integrates into benchmark frameworks, facilitating reproducible testing [9].

3 Methodology

In this work, we apply the Tetrad library, developed by Carnegie Mellon University for causal analysis, to clinical data from patients with type 1 diabetes. Tetrad supports multiple algorithms, integrates with Python, and provides a graph-based analysis interface. Figure 1 shows the project's progress to date.

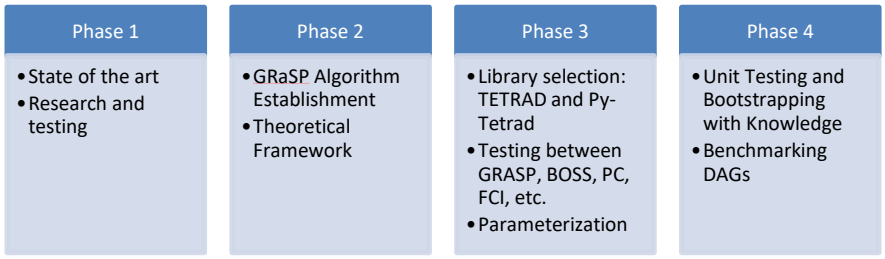


Fig. 1. Phases of the methodology that were realized to date.

The HUPA-UCM Diabetes 2024 dataset was selected, which includes the following information: date, time, preprandial glucose, bolus insulin delivered, carbohydrates, calories, exercise or steps, and heart rate. We calculated postprandial glucose and other relevant factors [13]. After cleaning and normalization, three versions of the dataset were generated: original, log-transformed, and Yeo-Johnson-transformed. The GRaSP algorithm was implemented with the SEM-BIC score and the Fisher Z test of independence. To improve causal validity, knowledge constraints were applied to prohibit illogical relationships between variables. The generated structures were visualized with Graphviz and validated using bootstrapping methods, and causal effect estimates were obtained with DoWhy and EconML.

4 Results and discussion

Preliminary results show that incorporating knowledge-based constraints significantly improves the structural consistency of the generated causal graphs. However, violations of the Markov principle and the non-Gaussian distribution of some variables limit the stability of the inferences. Graphs with medical constraints were found to reflect better the dependencies among glucose, insulin, and exercise, although the parameterization still needs to be fine-tuned to improve consistency. A comparative analysis with traditional predictive models (such as XGBoost) is underway and aims to evaluate the complementarity between statistical and causal prediction. Figure 2 shows DAG output by standard GRaSP run (22 total relationships):

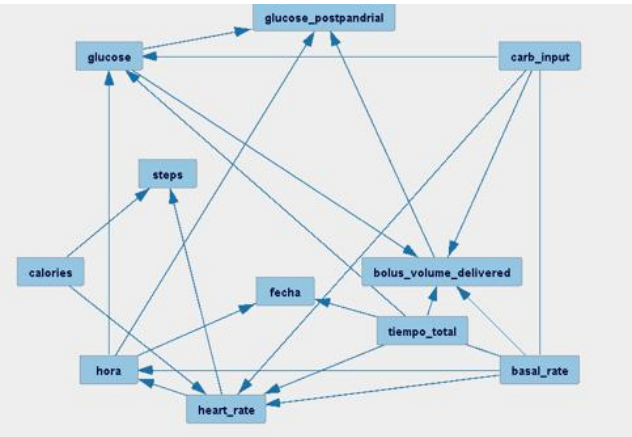


Fig. 2. DAG output by standard GRaSP

Figure 3 shows the DAG output by a Bootstrap Run of 1000-folds with knowledge (13 total relationships):

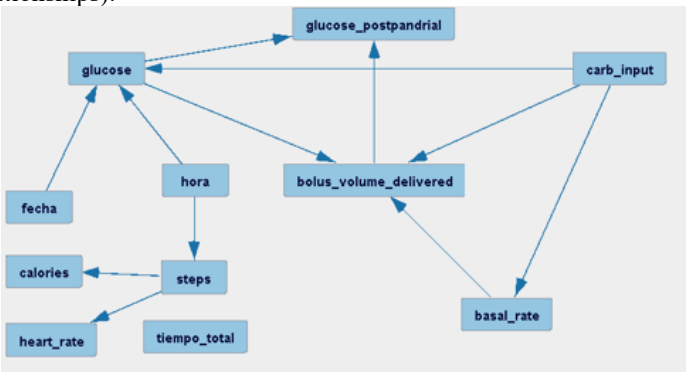


Fig. 3. DAG output by a Bootstrap Run of 1000 folds with knowledge

Causal inference applied to the calculation of insulin bolus in patients with type 1 diabetes using the GRaSP algorithm.

4.1 Future work

It is still pending:

- The justified definition of the optimal parameterization for the GRaSP algorithm.
- Validation of the final causal DAG.
- Specify the weights for each variable and their noise (error) coefficients.
- Obtain the causal structure and the linear equation that represent the probability distribution of all variables based on the target variable (postprandial glucose), given the treatment variable (postprandial insulin bolus), in the context of data from patients with diabetes. This equation models how the postprandial insulin bolus, along with other relevant variables (such as carbohydrate consumption, pre-meal glucose, etc.), causally influences postprandial glucose.

Final Part: Causal Inference:

- Use of the expression of causal relationships for counterfactual (intervention) testing.
- Create a causal predictive model for the postprandial insulin bolus.

5 Conclusions

GRaSP constitutes a modern, theoretically grounded alternative to the arsenal of causal discovery methods. Its features (permutation search, tuck operation, relaxation hierarchy) make it particularly well-suited to high-dimensional problems and dense graphs, which are common in biomedicine and digital health. However, its practical application in healthcare must be framed within pipelines that include simulation validation, methodological comparison, and triangulation with external evidence to mitigate risks associated with unobserved confounding and model assumptions.

Acknowledgments. We thank the Tecnológico Nacional de México for funding this project with grant 21636.25-P, as well as the student Maria Fernanda Abarca Colín of the Dolphin Program, who actively participated in the development of this project.

Disclosure of Interests. The authors declare that they have no conflicts of interest in this article.

References

[1] World Health Organization, "Diabetes," 14 11 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed 16 07 2025].

[2] NHS, "Complications of type 1 diabetes," NHS, [Online]. Available: https://www.nhs.uk/conditions/type-1-diabetes/complications/. [Accessed 11 10 2025].

[3] Metlife, "¿Cuánto cuesta vivir con diabetes en México?," 01 11 2024. [Online]. Available: https://www.metlife.com.mx/blog/bienestarfinanciero/gastos-de-vivir-con-diabetes-cuanto-necesitas/. [Accessed 16 07 2025].

[4] G. C. M. V. G. S. S. D. F. a. A. F. G. Noaro, "Machine-Learning Based Model to Improve Insulin Bolus Calculation in Type 1 Diabetes Therapy," *IEEE Transactions on Biomedical Engineering,* vol. 68, no. 1, pp. 247-255, 2021.

[5] T. L. K. K. L. H. P. &. G. P. Zhu, "An insulin bolus advisor for type 1 diabetes using deep reinforcement learning," *Sensors,* vol. 20, no. 18, p. 5058, 2020.

[6] D. K. a. K. B. Mirza, "InsNET: Accurate Basal and Bolus Insulin Dose Prediction for Closed Loop Diabetes Management," *45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC),* pp. 1-4, 2023.

[7] W. A. B. &. Lam and J. Ramsey, "Greedy relaxations of the sparsest permutation algorithm," *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence,* 2022.

[8] X. P. S. L. J. e. a. Wu, "Causal inference in the medical domain: a survey," *Appl Intell,* vol. 54, pp. 4911-4934, 2024.

[9] W. Y. A. B. &. R. J. Lam, *GRaSP Project Repository,* 2022.

[10] S. J. Maradei, "Ciencias ómicas: ¿Eso qué es y con qué se come?," Biotecgen , [Online]. Available: https://www.biotecgen.com.co/blog/cienciasomicas. [Accessed 11 10 2025].

[11] D. L. R. S. Y. D. X. I. K. S. P. &. Z. K. Zeng, "Causal discovery and counterfactual reasoning to optimize persuasive dialogue policies.," *Behaviour & Information Technology,* pp. 1-15, 2025.

[12] U. H. M. O. G. Muhammad Hasan Ferdous, "CDANs: Temporal Causal Discovery from Autocorrelated and Non-Stationary Time Series Data," *Proceedings of Machine Learning Research,* 2023.

[13] J. I. Hidalgo, J. Alvarado, M. Botella, A. Aramendi, J. M. Velasco and O. (. Garnica, *"HUPA-UCM Diabetes Dataset",* Madrid, Madrid, 2024.

Probabilistic Logic Twin Networks for Safe Driving Decisions: Edge-Constrained vs. Unconstrained DAG Learning

Héctor Avilés[1 [0000−0001−5310−3474]], Ingridh Gracia[1 [0009−0004−7101−6656]], Rafael Kiesel[2 [0000−0002−8866−3452]], Verónica Rodríguez[3 [0000−0002−5976−9338]], Rubén Machucho[1 [0000−0002−5731−6677]], Alberto Reyes[4 [0000−0002−8509−6974]], Marco Negrete[5 [0000−0002−5468−2807]], Gabriel Ramírez[6 [0000−0001−5226−5615]], Nicolás Luévano[1 [0009−0004−2527−6086]], Myriam Pequeño[1 [0009−0005−3619−3259]], Jesús Medrano[1 [0009−0004−2527−6086]], and Felix Weitkämper[7,8 [0000−0002−3895−8279]]

[1] Polytechnic University of Victoria, 87138 Tamaulipas, Mexico
{havilesa,2330319,rmachuchoc,2330215,2330006,jmedranoa}@upv.edu.mx
[2] Vienna University of Technology, 1040 Vienna, Austria
rafael.kiesel@tuwien.ac.at
[3] Technological University of the Mixteca, 69004 Oaxaca, Mexico
veromix@mixteco.utm.mx
[4] National Institute of Electricity and Clean Energy, 62490 Cuernavaca, Mexico
areyes@ineel.mx
[5] National Autonomous University of Mexico, 04510 Mexico City, Mexico
marco.negrete@ingenieria.unam.edu
[6] Center for Research and Advanced Studies of the National Polytechnic Institute, Tamaulipas Campus, 87138 Tamaulipas, Mexico
grtorres@cinvestav.mx
[7] German University of Digital Science, Potsdam, Germany
[8] Ludwig-Maximilians-Universität München, München, Germany
felix.weitkaemper@german-uds.de

Abstract. In this paper, we compare two types of probabilistic logic twin networks (*PLTNs*) learned *with* and *without* edge constraints for the selection of "safe" counterfactual driving actions in autonomous vehicles. The PLTN models are constructed from causal Bayesian network (*cBN*) hypotheses learned from data: (a) using edge-constrained learning of the directed acyclic graph (*DAG*), where specific causal relations are forbidden via an edge *blacklist*, and (b) learning a DAG with no edge constraints. Our results indicate that PLTNs learned without edge constraints always identify a unique action closely aligned with the training data, but one that is often judged as "unsafe" according to expert knowledge. In contrast, PLTNs learned with edge constraints produce both a broader variety of safe actions and several ties for the safest action, but never recommend unsafe ones. These findings highlight the relevance of edge constraints in safety-critical decision-making for autonomous driving.

Keywords: Counterfactual reasoning · Probabilistic logic · Autonomous driving.

1 Introduction

Collision avoidance through the evaluation of alternative, "safer" courses of action is a critical requirement for achieving fully autonomous driving capability. In [1, 16], it was shown that Pearl's counterfactual reasoning, implemented through probabilistic logic twin networks (*PLTNs*) [2, 11], is a promising framework for preventing collisions in autonomous vehicles. PLTNs are twin networks built on causal Bayesian networks (*cBNs*) [13] encoded as probabilistic logic programs [15]. Probabilistic logic is a paradigm well-suited for modeling causal relationships due to its clear and flexible rule-based representation and the availability of sophisticated probabilistic inference procedures [3, 7]. This framework helps answer questions about prospective, hypothetical situations such as:

"*What is the probability that a potential collision could occur, given that the current action and the system state are known, if a different action were chosen?*".

This type of "*What-if*" question enables self-driving cars to evaluate the convenience of alternative driving actions as safer maneuvers under hazardous situations [17].

In this work, we compare probabilistic logic twin networks (*PLTNs*) built on causal Bayesian network (*cBN*) hypotheses learned *with* and *without* imposing edge constraints for the construction of their directed acyclic graph (*DAG*). The overall goal is to identify safe driving actions in self-driving cars facing potential collisions. In the first case, DAGs are optimized from an empty edge set but subject to specific blacklisted edges. These constraints take into account expert judgment about implausible causal associations in our setting. In the second case, a DAG is fully learned from data without any constraints. Inference is carried out through the Counterfactuals package [9], an efficient and effective tool recently developed to solve probabilistic counterfactual queries in PLTNs. Our preliminary results indicate that PLTNs with fully learned cBNs consistently recommend a single safest action that aligns with the training data. However, this action often corresponds to the same maneuver that currently leads to a potential collision. In contrast, PLTNs with edge constraints offer greater action diversity, with up to five actions tied as safest, but no unsafe actions. Our empirical findings suggest that edge constraints greatly influence the selection of safe driving actions, posing a significant concern for safety-critical autonomous driving.

2 Methodology

2.1 Testbed and datasets

Our testbed consists of a self-driving car simulated in a race-like environment (see Figure 1).[9] The self-driving car travels on a two-lane road with straight seg-

[9] The source code of our self-driving vehicle and some videos of the self-driving system running are available at: https://github.com/hector-aviles/self-driving-car-2025.

Fig. 1. Race-like environment considered in this study for our self-driving car (in bright red).

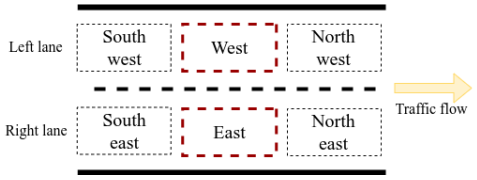


Fig. 2. Predefined locations for other vehicles around the self-driving car (dashed red lines indicate the space the self-driving car occupies on each lane). When the vehicle is on the left (respectively, right) lane, only the locations Northwest, Northeast, East and Southeast (respectively, Northeast, Northwest, West and Southwest) are meaningful.

ments and curves and up to 10 obstacle vehicles distributed over the road, either static or in motion. The maximum tested speed of the self-driving car is near 50 km/h. The architecture of the self-driving car includes modules for *perception*, *motor control*, *decision-making*, and *collision detection* based on radar and laser readings. State variables are `curr_lane`, which identifies the lane in which the self-driving car travels, and 6 occupancy variables called `free_E`, `free_NE`, `free_NW`, `free_SE`, `free_SW` and `free_W` which indicate whether there is a vehicle or not in the location indicated in the name of each variable, relative to the self-driving car. Figure 2 depicts the locations represented by the occupancy variables on each lane. These state variables form a 7-tuple, hereinafter referred to as `state`. In addition, a multi-valued variable `action` identifies one of 6 driving maneuvers for the self-driving car: **change_to_left** and **change_to_right** used for changing to the left lane and changing to the right lane, respectively, **cruise** to reach a steady (maximum) speed, **keep** (distance) to maintain a steady distance to a vehicle ahead, and **swerve_right** and **swerve_left** to veer to the side of the lane while reducing speed. These two latter are reserved for potential collision scenarios, and together with **keep**, they are considered the safest actions.

We have recorded a dataset with data from either autonomous control or human control of the autonomous car using a human-friendly interface[10]. Each possible instance from the complete `state-action` space of size $2^7 \times 6 = 768$ was manually labeled as leading to a collision when: (a) the self-driving car

[10] These datasets are available at: https://www.kaggle.com/autonomousvehicle/

Table 1. Total number of state-action pairs labeled as potential and non-potential collisions in the integrated dataset, categorized by action.

Action	# of non-potential collisions	# of potential collisions	**Total:**
change_to_left	40,308	289,809	330,117
change_to_right	39,814	294,459	334,273
cruise	617,930	175,429	793,359
keep	496,558	0	496,558
swerve_left	2,883	0	2,883
swerve_right	1,681	0	1,681
Total:	1,199,174	759,695	1,958,869

performs action **cruise** and there is a vehicle close ahead traveling in the same or opposite direction (*rear-end* or *head-on* collision, respectively), and (b) when there is a car either next to or ahead in the lane the self-driving car merges into (*sideswipe* and rear-end crash, respectively). This procedure resulted in 288 `state-action` pairs as leading to a potential collision. All "unsafe" pairs involve the actions **change_to_left**, **change_to_right**, and **cruise** only. With the new list of "safe" and unsafe potential instances as reference, we found that the automated and human control databases contain only a small number of potential crash examples (39,695 state-action pairs, all of which are repetitions of 132 unique pairs). To introduce more examples of potential crashes into the original dataset, the 288 unique state-action pairs were replicated 2,500 times to generate a synthetic dataset of 720,000 potential crashes, a number selected arbitrarily to achieve a balance between safe and unsafe examples in the dataset. All examples were labeled either as potential collision or not using the variable `latent_collision`, thus leading to `state-action-latent_collision` triplets.

Table 1 summarizes the total number of `state-action` pairs labeled as potential collisions or non-potential collisions in the integrated dataset, arranged by action.

2.2 Learning causal Bayesian networks

We implement structural and parameter learning of cBNs with the **bnlearn** package [18] in R [14]. Structural DAG learning is performed via hill-climbing greedy search with the decomposable and score-equivalent Bayesian Information Criterion (*BIC*). Hill-climbing is a score-based algorithm that iteratively modifies DAG edges to maximize the BIC score. This score is equal to the log-likelihood of the data given the DAG, penalized by the number of parameters. In our setting, DAG learning starts with an empty structure (i.e., no edges, which is the default in many **bnlearn** settings) and no restarts (i.e., only a single search path). Although this procedure typically reaches local optima only, it

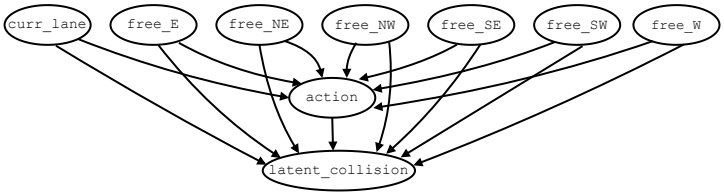


Fig. 3. Admissible edges in causal Bayesian networks *with* edge constraints.

is nevertheless helpful to identify persistent[11] statistical correlations observable from resampling. In addition, we believe this learning setting is also enough to isolate the effects of edge constraints in counterfactual estimations. We estimate parameters using the maximum likelihood estimation criterion (*MLE*) [12] applying uniform smoothing to the child probabilities when parent configurations are unobserved.

The admissible edges in the DAG with edge constraints are shown in Fig. 3. They are allowed since we assume that the observed state of the environment causally influences the driving decisions that the car makes, which, in turn, affect the probability of collisions with other vehicles. We consider the variable `action` as a treatment variable, `latent_collision` as the outcome variable, and the state variables as a set of Boolean confounders. Undesirable edges between variables (e.g., those between state variables or from `action` to state variables) were blacklisted. The hill-climbing algorithm then selects the final subset of edges from the admissible edges based on the training dataset.

To construct the cBN models without structure constraints, we let the hill-climbing algorithm decide the edges between variables. In both cases, as in the vast majority of the unconstrained cBNs, we observed that some associations are consistent. For example, `action` is always the root node of the DAG and is often also the parent of all the other variables. Some relationships are reasonable from a causal perspective. One example is the edge from `action` to `curr_lane`, which could describe how an action affects the lane in which the car is traveling (e.g., during lane changes). These two variables, along with `free_NE` and `free_NW`, consistently had direct edges to `latent_collision`, indicating plausible relevance in rear-end collision risk. Other associations are less justifiable, such as edges between state variables (e.g., it is unclear how the occupancy in the West location could cause the occupancy in the East location). Another common example is the presence of edges from `latent_collision` to state variables.

We interpret the observed structural consistencies as evidence of reliable statistical associations observed in our dataset. However, we emphasize that the DAGs learned are tenable causal hypotheses inferred from data and not validated causal truths.

[11] Persistence is a necessary, although not a sufficient indicator of causality.

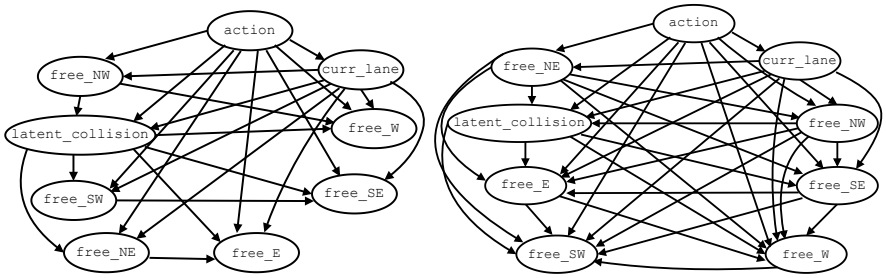


Fig. 4. Examples of DAGs *without* edge constraints learned with 19,514 examples (left) and with 1,756,232 samples (right).

2.3 Modeling and querying probabilistic logic twin networks

The `Counterfactuals` package requires a cBN encoded as a probabilistic logic program in ProbLog[12]. The ProbLog program must be extended with *exogenous*,[13] *unobserved error terms*. These error terms account for unobserved factors that introduce variability into the deterministic cause-effect relationships in the model, as described in the next paragraphs. To syntactically translate cBNs from **bnlearn** into ProbLog, we developed a dedicated R script.

For illustration, consider a simple example of an autonomous vehicle traveling on the right lane only. Thus, its cBN only needs to retain the variables `free_NE`, `action` and `latent_collision`, and the actions **cruise** and **keep**. The simplified cBN, derived from the cBN presented in Figure 3, is shown on the left side of Figure 5. It was learned using a small data sample.

The corresponding ProbLog program is described in Listing 1.1. Lines 3-9 declare error terms. The terms `u2`, `u3` are binary and defined over {`cruise, keep`}. Unlike our previous work, we introduce *annotated disjunctions* (*ADs*) [19] to specify the values and probabilities for multi-valued variables in the causal model. ADs ensure that exactly one value of the variable is true at any time. In this example, we define error terms with ADs in lines 4-5. The other error terms `u1`, `u4`-`u7` are Boolean, and for each, only its probability of being true is required. Lines 15-16 declare deterministic causal-effect rules from `free_NE` to `action`. For example, line 15 states that the probability of executing **cruise** is ≈ 0.33, while performing **keep** is ≈ 0.67 when Northeast location is not empty (i.e., \+ `free_NE`). From line 16, it can be deduced that the preferred action is **cruise**, with a probability of ≈ 0.93 whenever `free_NE` is true (i.e., the space ahead is clear). Lines 18-21 describe the effect of `action` and `free_NE` on `latent_collision`. In 18, according to data, a potential crash is slightly more likely than not (≈ 0.57) whenever **cruise** is executed and there is a car ahead. Lines 19-21 and the error terms `u5`-`u7` declared in lines 7-9 account for the low

[12] For an introduction to ProbLog syntax see [8].
[13] An exogenous variable is one that is not influenced by any other variable.

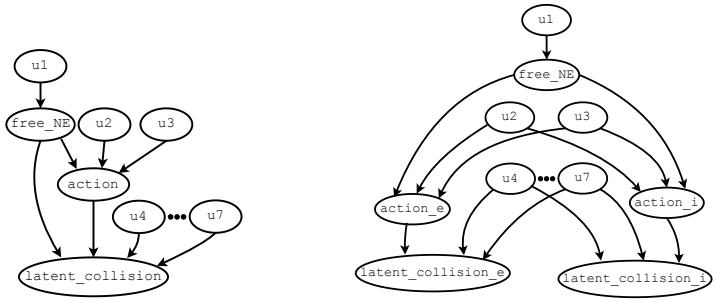


Fig. 5. Example of a simple causal Bayesian Network (left) and its corresponding twin networks model (right).

probability of a potential collision either when the action **keep** is applied or when there is no vehicle ahead.

Listing 1.1. ProbLog encoding of the simple causal Bayesian network.

```
%%% Error terms

0.4618446::u1.
0.3314485::u2(cruise); 0.6685515::u2(keep).
0.9371515::u3(cruise); 0.0628485::u3(keep).
0.5765529::u4.
0.0009990::u5.
0.0838291::u6.
0.0009990::u7.

%%% Rules

free_NE :- u1.

action(V) :- u2(V), \+ free_NE.
action(V) :- u3(V), free_NE.

latent_collision :- u4, action(cruise), \+ free_NE.
latent_collision :- u5, action(keep), \+ free_NE.
latent_collision :- u6, action(cruise), free_NE.
latent_collision :- u7, action(keep), free_NE.
```

A twin networks model consists of two copies of the cBN, including its original structure and parameters of the *endogenous* variables[14] `free_NE`, `action`, and `latent_collision`. The first copy represents the variables in the evidence world, while the second copy stands for the intervention world. The right side of Figure 5 depicts the DAG of the twin networks model. The suffix "_e" identifies

[14] An endogenous variable is one that is influenced by other variables.

endogenous variables in the evidence world, whereas "_i" identifies those in the intervention world. The two copies are linked through error terms u1-u7.

Here, we want to query conditional probabilities such as:

$$p\big(\texttt{latent\_collision\_i} \,\big|\, \texttt{latent\_collision\_e}, \texttt{free\_NE\_e}, \texttt{action\_e}, do(\texttt{action\_i})\big) \tag{1}$$

where $p(\cdot|\cdot)$ is a conditional probability function, latent_collision_e, free_NE_e and action_e are all evidence, and $do(\cdot)$ is the *do*-operator [13], which introduces an intervention over action_i. The objective is to identify a hypothetical action action_i, that minimizes the probability of latent_collision_i given the current state and action. For inference, conditional distribution over error terms given the evidence influences the intervention world.

The package Counterfactuals applies the following probabilistic logic programming pipeline to answer counterfactual queries:

1. assert the interventional action,
2. construct a new program with the logical atoms relevant to the query and the evidence by using an and/or graph,
3. ground facts and rules (in our example, all atoms are already grounded),
4. perform Clark's completion to handle *negation as failure* [4],
5. produce a propositional formula, in conjunctive normal form (*CNF*),
6. use sharpsat-td [10] to compile the CNF formula into a *smooth deterministic decomposable negation normal form* (*sd-DNNF*) [5], and
7. solve the conditional probability query via *weighted model counting* [20] through the aspmc package [6].

3 Evaluation and results

This evaluation qualitatively assesses how well alternative driving actions selected via PLTNs help avoid potential crashes for our self-driving car. For evaluation, a leave-one-out cross-validation (*LOOCV*) design was implemented. First, a unique pair from the state-action space is selected and separated for testing. Denote this pair as (**state**_e, action_e). Second, remove all examples whose **state** component matches the selected **state**_e from the integrated dataset Δ to avoid data leakage. This yields a smaller dataset, denoted Δ_{train}, used for training and containing triplets (**state**, action, latent_collision), where **state** $\neq$ **state**_e, for all the triplets in Δ_{train}. Third, two cBNs with and without edge constraints are trained using Δ_{train} and their corresponding PLTNs constructed, as described in Sections 2.2 and 2.3. Fourth, append latent_collision_e = **T** to (**state**_e, action_e) to obtain the triplet

$$(\textbf{state\_e}, \texttt{action\_e}, \texttt{latent\_collision\_e} = \mathbf{T}).^{15}$$

[15] The assignment latent_collision_e = **T** simulates a potential crash in the observed state-action scenario as alerted by the forward collision warning submodule.

Fifth, extend the previous triplet with each of the six available actions as interventions, producing six quartets:

$$(\textbf{state\_e}, \texttt{action\_e}, \texttt{latent\_collision\_e} = \textbf{T}, \texttt{action\_i}),$$

where `action_i` each takes a value in the set

$$\{\textbf{change\_to\_left}, \textbf{change\_to\_right}, \textbf{cruise}, \textbf{keep}, \textbf{swerve\_right}, \textbf{swerve\_left}\}.$$

At this stage, each quartet in the query group shares the same **state**_e, `action_e`, and `latent_collision_e` components, but incorporates a distinct interventional action. Sixth, these six quartets form a unique query group for testing from which the alternative actions will be evaluated. The final step involves minimizing the probability of collision by performing counterfactual querying on each quartet within the query group in the two PLTNs models created. Minimization is formalized as:

$$A^*_{min} = \arg\min_{\texttt{action\_i}} p\big(\texttt{latent\_collision\_i} \big| \texttt{latent\_collision\_e}, \textbf{state}\_\text{e}, \texttt{action\_e}, do(\texttt{action\_i})\big) \tag{2}$$

where A^*_{min} denotes the set of one or more interventional actions `action_i` that minimize the probability of a potential collision within the group and a given PLTNs model. The previous process is repeated for each of the 768 state-action pairs in the sample space and for random samples of size 1%, 50%, and 90% of Δ_{train}, stratified by `action_e`.

Results are summarized in Tables 2, 3 and 4. Table 2 presents a first glimpse of the performance of both PLTNs with and without edge constraints. It shows the total number of safe and unsafe actions suggested with respect to our manual labeling of the sample space. PLTNs with edge constraints never suggested unsafe actions, while PLTNs without edge constraints frequently proposed unsafe actions. Table 3 shows the number of actions tied as optimal within each six-quartet group. The PLTNs with edge constraints show ties ranging from one to five actions, whereas those without constraints exhibit no ties. Finally, Table 4 reports how often each action was selected. It is observed that PLTNs with edge constraints select a broader set of actions, whereas PLTNs without edge constraints primarily suggest **keep**, **swerve_left**, and **swerve_right**. Recall from Table 1 that these actions were originally labeled as safe in the dataset Δ. However, these actions are considered dangerous alternatives when they correspond to the action observed in the original (**state**_e, `action_e`) pair. All such mismatches arise from pairs originally labeled as safe (`latent_collision` = `F`) in the dataset Δ, but designated as unsafe (`latent_collision` = `T`) during testing for collision-risk simulation.

To summarize, Table 5 presents a comparison between the two models with the main characteristics observed in this evaluation. Our initial empirical evidence suggests that in our setting, PLTNs learned without edge constraints have a negative impact on the counterfactual decisions, making them aligned to

Table 2. Number of safe and unsafe actions suggested by the PLTNs, accordingly to our initial labeling of the sample space.

	PLTNs *with* edge constraints			PLTNs *without* edge constraints		
Action label	1%	50%	90%	1%	50%	90%
Safe	1661	1753	1772	538	470	406
Unsafe	0	0	0	230	298	362
Total:	1661	1753	1772	768	768	768

Table 3. Distribution of ties in best-ranked interventions: total number of tied actions[†] and the total number of query groups that generated them.

# of ties	PLTNs *with* edge constraints			PLTNs *without* edge constraints		
(1^{st} place)	1%	50%	90%	1%	50%	90%
1	249/249	199/199	188/188	768/768	768/768	768/768
2	468/234	502/251	508/254	0/0	0/0	0/0
3	609/203	681/227	705/235	0/0	0/0	0/0
4	300/75	336/84	336/84	0/0	0/0	0/0
5	35/7	35/7	35/7	0/0	0/0	0/0
6	0/0	0/0	0/0	0/0	0/0	0/0
Total:	1,661/768	1,753/768	1,772/768	768/768	768/768	768/768

[†]The total number of actions is equal to the # of ties in the 1^{st} place times the total number of groups.

Table 4. Frequency of action selection.

	PLTNs *with* edge constraints			PLTNs *without* edge constraints		
Action	1%	50%	90%	1%	50%	90%
change_to_left	80	80	80	0	0	0
change_to_right	0	80	80	0	0	0
cruise	320	320	320	0	0	0
keep	558	524	536	354	223	187
swerve_left	353	421	431	307	218	294
swerve_right	350	328	325	107	327	287
Total:	1661	1753	1772	768	768	768

training data, but potentially introducing dangerous maneuvers without offering safe alternatives. In contrast, incorporating edge constraints expands the array of possibilities, without introducing dangerous maneuvers. The downside of having multiple, equally probable alternatives at hand is that it requires the incorporation of posterior tie-breaking strategies. Despite this extra step, alternatives might help generalize to rare or difficult driving situations, which could lead to effective avoidance if the current chosen action fails. However, these results

should be taken with a grain of salt, since practical validation during deployment of our simulated self-driving car is still required.

Table 5. Comparison of PLTN models.

Characteristics	PLTNs *with* edge constraints	PLTNs *without* edge constraints
Unsafe recommendations	No	Yes
Presence of ties	Yes	No
Action diversity	Yes	Limited (data-driven)

4 Conclusions and future work

We presented ongoing work on developing models for counterfactual reasoning in collision avoidance using probabilistic logic twin networks (*PLTNs*). We compare counterfactual results from PLTNs based on two types of causal Bayesian networks (*cBNs*): one learned with edge constraints and the other entirely learned from data without edge constraints. Our preliminary evaluation shows that PLTNs with edge constraints generate several equally plausible actions to avoid crashes, and their action selection and safer choices are consistently aligned with our criteria about safe actions, compared to those without edge constraints. These findings highlight the importance of structural constraints such as edge restrictions, particularly in safety-critical decision-making domains such as autonomous driving.

As future work, we plan to: (i) evaluate constraint-based strategies such as Peter-Clark or Fast Causal Inference to identify cause-and-effect relationships, (ii) conduct evaluations with additional training-set percentages and repetitions, as well as (iii) testing decision-making in our simulated environment and in the AutoMiny V4, a real scaled autonomous vehicle developed at the Free University of Berlin. The primary goal of this latter test will be to better understand the practical implications of the models developed in this work for safe and efficient autonomous driving.

Acknowledgments. This work was partially funded by UNAM-DGAPA under grant IT102424 and AI Consortium - CIMAT-CONAHCYT.

References

1. Avilés, H., Negrete, M., Kiesel, R., Reyes, A., Rodríguez, V., Machucho, R., Ramírez, G., Pequeño, M., Gracia, I., Luévano, N., Rivera, K., Medrano, J.J.: Crash Avoidance for Autonomous Cars via Probabilistic Logic Counterfactual Reasoning. In: Handbook of Intelligent Robots: Theory, Methods and Applications. Taylor & Francis, CRC Press (2025), to appear.

2. Balke, A., Pearl, J.: Probabilistic evaluation of counterfactual queries, p. 237–254. Association for Computing Machinery, New York, NY, USA, 1 edn. (2022)
3. Chavira, M., Darwiche, A.: On probabilistic inference by weighted model counting. Artificial Intelligence **172**(6-7), 772–799 (2008)
4. Clark, K.L.: Logic and databases, chapter negation as failure. Eds. Plenum Press **5**, 293–322 (1978)
5. Darwiche, A., Marquis, P.: A knowledge compilation map. Journal of Artificial Intelligence Research **17**, 229–264 (2002)
6. Eiter, T., Hecher, M., Kiesel, R.: aspmc: An algebraic answer set counter. In: ICLP Workshops (2021), https://ceur-ws.org/Vol-2970/plppaper1.pdf
7. Eiter, T., Hecher, M., Kiesel, R.: aspmc: New frontiers of algebraic answer set counting. Artificial Intelligence **330**, 104109 (2024). https://doi.org/10.1016/j.artint.2024.104109
8. Fierens, D., Van den Broeck, G., Renkens, J., Shterionov, D., Gutmann, B., Thon, I., Janssens, G., De Raedt, L.: Inference and Learning in Probabilistic Logic Programs using Weighted Boolean Formulas. Theory and Practice of Logic Programming **15**(3), 358–401 (2015)
9. Kiesel, R.: counterfactuals (2023), https://github.com/raki123/counterfactuals
10. Kiesel, R., Eiter, T.: Knowledge compilation and more with sharpsat-td. In: Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning. KR '23 (2023). https://doi.org/10.24963/kr.2023/40
11. Kiesel, R., Rückschloß, K., Weitkämper, F.: "What if?" in Probabilistic Logic Programming. Theory and Practice of Logic Programming **23**(4), 884–899 (2023)
12. Le Cam, L.: Maximum likelihood: An introduction. International Statistical Review / Revue Internationale de Statistique **58**(2), 153–171 (1990). https://doi.org/10.2307/1403464
13. Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, USA, 2nd edn. (2009)
14. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021), https://www.R-project.org/
15. Riguzzi, F.: Foundations of probabilistic logic programming. River Publishers, New York (May 2023)
16. Rodríguez, V., Avilés, H., Machucho, R., Reyes, A., Negrete, M., Ramírez, G., Petrilli, A., Gracia, I., De-La-Garza, G., Rivera, K., Kiesel, R.: Preventing collisions in self-driving cars using probabilistic logic counterfactual reasoning. In: Workshop on Causal Discovery (CaDis). Instituto nacional de Astrofísica, Óptica y Electrónica (INAOE) (2024), available at: https://cadisworkshop.com.mx/wpcontent/uploads/2024/11/preventing-collisions-in-self-driving-cars-usingprobabilistic-logic-counterfactual-reasoning.pdf
17. Ruiz-Tagle, A., Lopez-Droguett, E., Groth, K.M.: A novel probabilistic approach to counterfactual reasoning in system safety. Reliability Engineering & System Safety **228**, 108785 (2022)
18. Scutari, M.: Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software **35**(3), 1–22 (2010)
19. Vennekens, J., Verbaeten, S., Bruynooghe, M.: Logic programs with annotated disjunctions. In: International Conference on Logic Programming. pp. 431–445. Springer (2004). https://doi.org/10.1007/978-3-540-27775-0_30
20. Vlasselaer, J., Kimmig, A., Dries, A., Meert, W., De Raedt, L.: Knowledge compilation and weighted model counting for inference in probabilistic logic programs. In: AAAI Workshop: Beyond NP. vol. 101 (2016)

Scenario optimization with FCMs and MOEAs: problematization of access to public transport in Mérida

Aaron U. Poot Hoil[1], Fernanda Pérez Lombardini[1], Marco A. Rosas[2], Carlos I. Hernández Castellanos[3], and Jesús Mario Siqueiros García[1]

[1] IIMAS , UNAM, Mérida, Yucatán
[2] C3, UNAM, Mexico City, Mexico
[3] IIMAS , UNAM, Mexico City, Mexico
carlos.hernandez@iimas.unam.mx

Abstract. Urban mobility involves intertwined social, infrastructural, and environmental factors. Understanding how policy actions propagate through these relationships is essential for sustainable planning. This work presents a hybrid causal modeling and optimization framework that combines Fuzzy Cognitive Maps (FCMs) with Multi-objective Evolutionary Algorithms (MOEAs) to explore intervention scenarios in the city of Mérida, Yucatán, Mexico. Through a participatory process, stakeholders built an FCM describing 25 urban mobility variables, five of which were treated as controllable interventions. Using the non-dominated sorting genetic algorithm II (NSGA-II), the framework identified Pareto-optimal combinations of actions that jointly improve well-being, access to public transport, and temperature mitigation. Preliminary results indicate that multi-variable interventions, which involve governance, vegetation cover, and transport infrastructure revealed trade-offs and synergies. The study advances on the feasibility of linking expert-based causal models with AI-driven optimization for interpretable, data-informed decision support.

Keywords: Fuzzy Cognitive Maps · Evolutionary Algorithms · Optimization · Urban Mobility.

1 Introducción

Currently, 55% of people live in cities, and the United Nations projects that this figure will rise to 68% by 2050 [4]. In this context, the design and planning of urban public spaces to ensure a good quality of life for their inhabitants is one of the most critical challenges of our time. One of these challenges is associated with good mobility in the city and the adequate design of the public transport system, considering that cities will continue to grow.

Our work focuses on access to public transport in the city of Mérida, Yucatán. For several years, Mérida has been the fastest-growing city in Mexico, which has put significant pressure on improving the public transport system. The efforts of the authorities have been considerable, but much remains to be done, especially

regarding accessibility and the condition of bus stops. Such areas of opportunity result from the city's lack of sound planning and infrastructure necessary to cope with the stress experienced by its users, which climate change exacerbates. For much of the year, daytime temperatures are extreme (33 °C to 45 °C), and the torrential rainy season lasts from May to November.

Based on our case study, this summary presents the methodological approach for scenario optimisation through the coordinated use of FCMs [2] and MOEAs [1] in the context of a participatory process surrounding access to public transport and urban mobility in general in Mérida. The main results were the group identification of the system variables on which it would be preferable to act to improve access and urban mobility in general, as well as the generation of different scenarios prioritizing some variables over others from the set of selected variables to reflect these preferences.

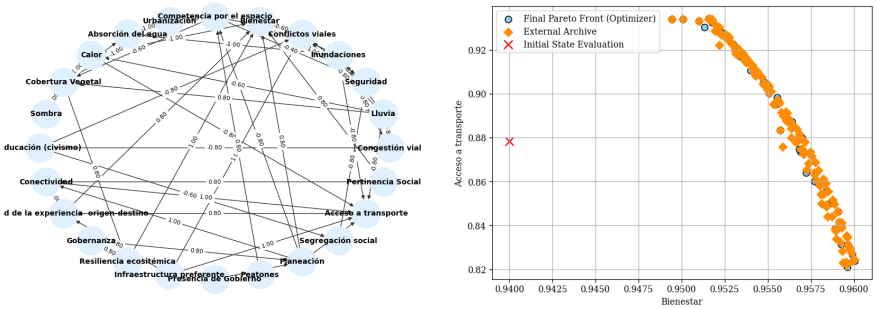
The remainder of this paper is organized as follows: Section 2 introduces the proposed framework that integrates FCMs and MOEAs for causal intervention analysis. Section 3 presents the application of this framework to the case of urban mobility in Mérida, Yucatán, including the experimental setup and main optimization results. Finally, Section 4 summarizes the key findings and outlines future research directions aimed at extending the approach to more dynamic and participatory causal models.

2 Fuzzy Cognitive Maps and Multi-Objective Evolutionary Algorithms

FCMs represent causal knowledge as a signed, directed graph $G = (C, E)$ where each concept C_i denotes a variable of the system and each weighted edge $w_{ij} \in [-1, 1]$ encodes the influence of C_i on C_j. The activation state $A^{(t)}$ of all concepts evolves according to $A^{(t+1)} = f(A^{(t)}W)$, where W is the weight matrix and $f(\cdot)$ a bounded nonlinear activation function (sigmoid or hyperbolic tangent). Once the system reaches a steady state A^*, it represents the predicted outcome of a given intervention.

An intervention is defined by externally fixing the activation of selected variables (analogous to Pearl's do-operator) and simulating its propagation through the causal network. Let $\mathbf{x}$ be the vector of controllable variables and $\mathbf{F}(\mathbf{x}) = [F_1(\mathbf{x}), F_2(\mathbf{x}), F_3(\mathbf{x})]$ the steady-state values of key outcomes such as well-being, access to transport, and temperature. The optimization problem seeks the set of non-dominated solutions that minimize $\mathbf{F}(\mathbf{x})$ under the system dynamics, revealing trade-offs among objectives.

To approximate the Pareto front, we employ NSGA-II [1] implemented in pymoo. Each encodes an intervention vector $\mathbf{x} \in [0, 1]^p$; fitness evaluation uses iterative FCM simulation in pyfcm. The algorithm uses Simulated Binary Crossover (SBX), Polynomial Mutation (PM), population size 40, and 50 generations with an external archive to preserve diversity [3]. The resulting set of Pareto-optimal interventions provides interpretable, causally grounded scenarios for participatory analysis and decision support.



(a) Causal cognitive map (FCM). (b) Pareto-front found by NSGA-II.

Fig. 1: Fuzzy cognitive map and evolutionary optimization results.

3 Results

The proposed framework was applied to a causal model of urban mobility in Mérida, Yucatán, developed through a participatory process with local stakeholders. In this case Education, vegetation cover, urbanization, governance, and preferential public transport infrastructure were selected as decision variables, while well-being, access to public transport, and temperature served as optimization objectives.

Fig. 1 summarizes the algorithmic behavior: (a) shows the collective FCM while (b) presents the final Pareto front compared with the initial state of the system. The front reveals clear trade-offs between well-being and access to public transport, where improved accessibility may slightly increase temperature, emphasizing the nature of the problem.

A representative Pareto-optimal scenario is depicted in Fig. 2. (a–b) illustrate changes in decision and objective variables, while (c–d) show the resulting intervention levels and steady-state activations. Simultaneous improvements in governance, education, and transport infrastructure generated the most favorable outcomes across objectives. Together, Fig. 2 shows how the FCM–MOEA framework uncovers interpretable causal interactions and synergistic multi-variable strategies for sustainable urban mobility planning.

4 Conclusions and Future Work

This extended abstract presented a proof of concept integrating Fuzzy Cognitive Maps (FCMs) and Multi-Objective Evolutionary Algorithms (MOEAs) to explore causal intervention scenarios in urban mobility. Applied to the city of Mérida, the approach linked participatory causal modeling with AI-based optimization, revealing interpretable trade-offs between well-being, accessibility, and temperature. The results highlight that multi-variable interventions exploiting causal synergies outperform isolated actions, demonstrating the potential of hybrid causal modeling for sustainable planning.

Fig. 2: Representative Pareto-optimal scenario.

Future work will incorporate resource and policy constraints into the optimization process, extend the framework to dynamic or data-driven causal models, and develop interactive visualization tools that allow stakeholders to co-explore Pareto-optimal scenarios, fostering participatory and informed decision-making in complex socio-environmental systems.

Acknowledgments. This research was supported by the Google Academic Research Award and by DGAPA-PAPIIT IA102025.

Disclosure of Interests. The authors have no competing interests.

References

1. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T. A. M. T. (2002). A fast and elitist multi-objective genetic algorithm: NSGA-II. IEEE transactions on evolutionary computation, 6(2), 182-197.
2. Gray, S. A., Gray, S., Cox, L. J., Henly-Shepard, S. Mental Modeler: A Fuzzy-Logic Cognitive Mapping Modeling Tool for Adaptive Environmental Management. In 2013 46th Hawaii International Conference on System Sciences, pp. 965–973. (2013)
3. Schütze, O., Hernández, C. (2021). Archiving strategies for evolutionary multi-objective optimization algorithms (pp. 1-242). Springer.
4. United Nations, Department of Economic and Social Affairs, Population Division: World Urbanization Prospects 2018: Highlights (ST/ESA/SER.A/421) (2019)

The Effects of fNIRS Signal Preprocessing in Effective Connectivity

Samuel Montero-Hernandez[1]

University of Birmingham, United Kingdom s.montero@bham.ac.uk

Abstract. Standard causal discovery methods are rarely applied to fNIRS (functional near-infrared spectroscopy) because their sensitivity to data statistical properties is easily disrupted by preprocessing. We present a study to track how standard fNIRS preprocessing affects key statistical characteristics for causal inference. Using resting-state fNIRS data, we found that preprocessing systematically transforms signal structure, enhancing stationarity but increasing slow temporal dependencies. These preliminary results underscore the need for preprocessing-aware causal analysis and algorithms tailored to the statistical profile of fNIRS data. Final work will integrate additional data and metrics to advance effective connectivity estimation in fNIRS.

Keywords: brain connectivity · causal discovery · fNIRS.

1 Introduction

1.1 From Associations to Causal Graphs

Connectivity analysis in neuroimaging is typically framed in terms of functional connectivity (FC) and effective connectivity (EC). FC quantifies statistical dependencies between signals, whereas EC aims to recover directed relations that are interpretable as causal influences [1]. The distinction is central: FC provides associative structure, while EC offers a model of information flow. Consequently, EC moves beyond correlation to identify directed dependencies consistent with underlying mechanisms [1, 2, 3].

Graphical models provide a natural framework for representing connectivity [4, 5]. Methods such as LiNGAM or DirectLiNGAM aim to recover a single directed graph by leveraging stronger assumptions—linearity, non-Gaussian errors, and acyclicity [6]. These assumptions facilitate identifiability, but verifying their validity in neuroimaging data such as fNIRS is challenging, as the signals are noisy, affected by systemic unobserved confounders, and sensitive to preprocessing choices [7].

1.2 fNIRS in Brain Connectivity Research

Effective connectivity has been extensively investigated using fMRI, where causal graphical models and related methods have provided insights into large-scale

brain networks [2]. However, fMRI imposes significant constraints on experimental design. Functional near-infrared spectroscopy (fNIRS) is an alternative modality that measures the hemodynamic responses from populations and settings where fMRI is impractical [8]. While most fNIRS connectivity research focuses on functional connectivity (FC), which uses statistical dependencies like correlations to find undirected associations [9], it cannot determine the direction of information flow [10, 11]. Consequently, the use of causal discovery methods to establish EC in fNIRS remains largely unexplored.

1.3 Signal preprocessing challenges

The adoption of effective connectivity analysis in fNIRS has been slow due to several challenges. Causal discovery methods are sensitive to the statistical properties of the fNIRS input signals.In addition, preprocessing pipelines can substantially alter the statistical characteristics of the signals, impacting both the identifiability and robustness of causal inference. These factors make it difficult to meet the assumptions of causal discovery methods highlighting the need for systematic evaluation of preprocessing choices. Preprocessing steps, such as band-pass filtering, motion artefact correction, or physiological noise regression, can alter the statistical properties of fNIRS signals and consequently affect the discovery of causal relationships [12]. The track of these changes facilitates the development of new algorithms tailored to the statistical characteristics of fNIRS data, improving interpretability of effective connectivity networks.

In this work, we investigate how preprocessing pipelines affect effective connectivity analysis of fNIRS signals in the resting-state paradigm. We focus on quantifying the impact of different preprocessing steps on the statistical properties of the signals and on the structure of the inferred causal networks. This allows to identify opportunities for developing algorithms tailored to the fNIRS characteristics.

2 Methods

Data and Participants FNIRS data were acquired from 12 subjects using a NIRSport device (NIRx Medical Technologies, USA), with 48 channels (760 nm and 850 nm) at 5.09 Hz. Channels covered the frontal, parietal, motor, and temporal regions bilaterally. Measurements corresponded to a resting-state paradigm during 12 minutes of recording.

Preprocessing Pipeline The preprocessing pipeline included standard procedures in fNIRS analysis. First, raw intensities were converted to optical density. Next, motion artefacts were corrected using temporal derivative distribution repair (TDDR) [13], followed by band-pass filtering to retain slow oscillations (0.01–0.2 Hz) and reduce systemic noise. Finally, signals were converted to oxy- and deoxy-haemoglobin concentrations using a differential pathlength factor of 6. Each preprocessing step was treated as a transformation stage to assess how

the statistical structure of the signals evolved along the pipeline. All the preprocessing steps were implemented in Python using the Cedalion framework [14].

Statistical Characterisation of Preprocessing Stages To quantify the impact of each preprocessing operation, we computed a set of statistical descriptors for every channel and wavelength at each stage. These descriptors capture properties directly linked to common assumptions of causal discovery algorithms. Distributional properties were evaluated through the Shapiro–Wilk test for normality, skewness, and kurtosis [15]. Temporal structure was examined using the Augmented Dickey–Fuller test for stationarity and the autocorrelation coefficient at lag 1 [16]. All metrics were computed separately per wavelength and tracked through each preprocessing stage. For this proof-of-concept, only one participant was analysed, allowing an exploratory view of within-subject transformations. Preliminary comparisons were based on visual and descriptive trends—for instance, examining whether normality or stationarity improved after filtering or motion correction.

3 Preliminary results

The effects of preprocessing on the distributional and temporal characteristics of fNIRS signals were evaluated in raw intensity, optical density conversion, motion artefact correction, band-pass filtering, and haemoglobin concentration computations. Results are presented in Figure 1.

Regarding distributional characteristics, the Shapiro–Wilk test yielded extremely low p-values ($log_{10}(p) < 10^{-8}$) for the raw signals, confirming pronounced deviations from normality. Across preprocessing stages, kurtosis values initially fluctuated above and below a value of three and gradually converged towards zero, suggesting a reduction in heavy-tailed behaviour and a trend toward more symmetric distributions. Skewness exhibited oscillatory patterns between positive and negative values across stages, further reflecting evolving asymmetry in the signal amplitude distributions. Temporal characteristics revealed a clear transition toward greater stationarity and reduced autocorrelation. As shown in the ADF test results (Fig. 1), p-values below the 0.01 significance threshold ($\log_{10}(p) \leq -2$) indicate that the signals progressively transitioned from non-stationary to stationary, particularly after band-pass filtering and haemoglobin concentration computation. Additionally, lag-1 autocorrelation values tended to increase after these stages, indicating an increase in slow temporal dependencies.

4 Conclusion

This proof-of-concept analysis demonstrates the feasibility of tracking the evolution of statistical properties of fNIRS signals throughout preprocessing. The results highlight that both distributional and temporal characteristics are systematically altered by standard preprocessing steps, with notable effects on normality, stationarity, and temporal dependency. However, exploration of additional

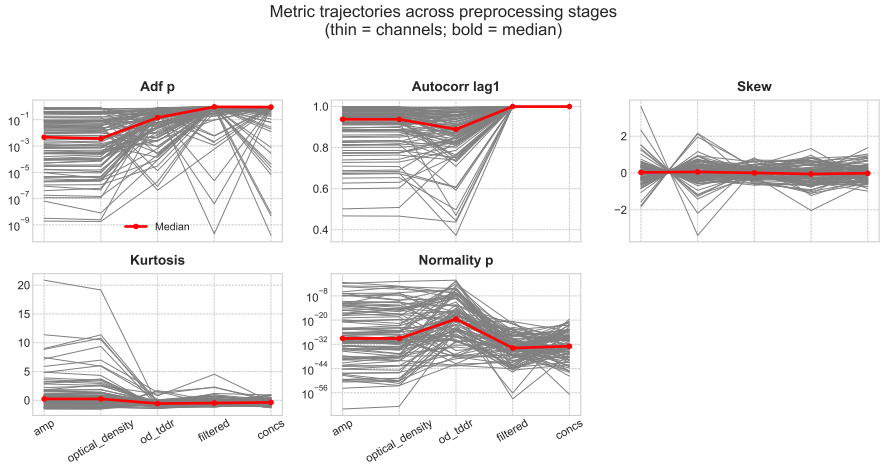


Fig. 1. Distributional and temporal evolution of fNIRS signals through preprocessing. Normality and shape metrics; Shapiro-Wilk test p-value (Normality_p), Skewness (Skew), and Kurtosis. Temporal structure metrics; Augmented Dickey-Fuller stationary test p-value (Adf_p), autocorrelation functional at lag 1 (Autocorr_lag1). X-axis indicates the signal preprocessing steps explored. (raw signals= amp, optical densities = optical_density, motion artefacts correction = od_tddr, band-pass filtering = filtered, and haemoglobin concentration calculation = concs.) Grey and red lines represent individual channels and median values.

metrics such as cross-channels dependencies, linearity, and homoscedasticity, and methods to correct the potential deviations from key statistical assumptions, as well as data from remaining participants will be incorporated in the final version of the paper to strengthen these findings and extend the analysis toward effective connectivity estimation.

References

[1] Karl J Friston. Functional and effective connectivity: a review. *Brain Connectivity*, 1(1):13–36, 2011.

[2] Anil K Seth, Adam B Barrett, and Lionel Barnett. Granger causality analysis in neuroscience and beyond. *Journal of Neuroscience*, 35(23):8193–8199, 2015.

[3] Karl J Friston. The effective connectivity of fmri is a causal effect. *NeuroImage*, 79:250–256, 2013.

[4] Judea Pearl. *Causality*. Cambridge university press, 2009.

[5] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81. MIT press, 2000.

[6] Shohei Shimizu, Takeshi Inazumi, Kirk Smith, Aapo Hyvärinen, Yoshinobu Kawahara, and Takashi Washio. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12:3863–3877, 2011.

[7] Meltem A Yücel, Stefan K öhler, and Scott B Kask. Functional near-infrared spectroscopy fnirs: an introductory review of the technology, analysis pipelines, and machine learning applications. *Neurophotonics*, 8(1):013003, 2021.

[8] Felix Scholkmann, Susanne Kleiser, Andreas J Metz, Roman Zimmermann, Javier Mata Pavia, Ursula Wolf, and Martin Wolf. A review on the processing and analysis of functional near-infrared spectroscopy fnirs data. *NeuroImage*, 85:69–83, 2014.

[9] Ahmet Yilmaz, Izzet Güler, and Emre Özbey. A review of functional near-infrared spectroscopy fnirs in brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics*, 26(2):681–692, 2022.

[10] Martina Grumm, Robert J Huster, and Simon Schäfer. A review of functional connectivity in fnirs studies. *BioRxiv*, page 417961, 2018.

[11] Jaron R Cohen and Mark D'Esposito. Statistical analysis of functional connectivity mri using graph theory. *NeuroImage*, 52(3):766–775, 2010.

[12] Sara Brigadoi, Robert J Cooper, Alberto D'Andrea, Adele Fasano, Fabrizio Fava, Michele Gervasoni, Edoardo Molteni, Virginia Pellizzi, Paola Pinti, Simone Schiavi, et al. Motion artifact removal in functional near-infrared spectroscopy: a comparison of different methods. *Journal of biomedical optics*, 19(10):105008, 2014.

[13] Frank A. Fishburn, Ruth S. Ludlum, Chandan J. Vaidya, and Andrei V. Medvedev. Temporal derivative distribution repair (tddr): A motion correction method for fnirs. *NeuroImage*, 184:171–179, 2019.

[14] Ibs-Lab et al. Cedalion: Causal evaluation data analysis library for neuroimaging. https://github.com/ibs-lab/cedalion, 2024. Accessed: 2025-10-12.

[15] A. Ghasemi and S. Zahediasl. Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, 10(2):486–489, 2012.

[16] David A. Dickey and Wayne A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431, 1979.