Composite Graphical Causal Models for Inference on Heterogeneous-Indexed Data

Arne De Temmerman ARNE.DETEMMERMAN@KULEUVEN.BE
and **Mathias Verbeke** MATHIAS.VERBEKE@KULEUVEN.BE
M-Group, Department of Computer Science, KU Leuven; Flanders Make@KU Leuven, Belgium

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Complex real-world systems typically consist of multiple interdependent subsystems, where each subsystem can operate under different sampling references. Consequently, the data collected across these subsystems vary in indexing (e.g., time-, distance-, event-indexed), sampling frequency, or faces index misalignment. To approximately model such kind of systems, surrogate models can be used to serve as a computationally-inexpensive replacement during optimization, sensitivity analysis, uncertainty quantification, or interpretation. Conventional modeling approaches require these datasets to be unified into a single, uniformly-indexed table via preprocessing steps such as aggregation and merging. In this work, we introduce a novel approach, Composite Graphical Causal Models (CGCMs), that preserves the original indexing of each data table during both training and inference. By embedding resampling and aggregation operations directly within a GCM, our method eliminates the need for data homogenization as a preprocessing step. Specifically, a set of GCMs is employed each tailored to a distinct indexing, and connected using aggregation functions to model cross-index dependencies. As validated on synthetic datasets, this design enables a more representative modeling of heterogeneous-indexed processes, improving predictive performance and interpretability.

Keywords: causal inference, composite model, heterogeneous-indexed data, graphical causal models

1. Introduction

Heterogeneous-Indexed Data (HID) refers to a dataset composed of multiple data tables collected from a single system that do not share a common index. The indexes could vary by the type, scale, alignment, or combination. HID is common in complex real-world systems involving multiple stages, and/ or complex, cumulative relations, e.g. in healthcare monitoring, time-series are aggregated to patient-specific and disease-specific parameters and records (Jacobs et al., 2025); in manufacturing, high-frequent time-series sensors are aggregated to broader periods or are linked to the product parameters (Pu et al., 2020); in agriculture daily plant health is derived from high-frequent sensors to finally aggregate them into a weekly harvest estimate (Hanssens, 2015).

1.1. Data indexing based on use

Indexes provide structure by labeling a dataset, allowing ordering, selection, and grouping of the referred data. The choice of index for a specific set of variables is determined by data type, storage constraints, processing requirements, and intended use.

Common data index types include temporal or time-based indexing which is applied in two major cases; (i) The recorded data is time-dependent or (ii) no better index is available to map the data.

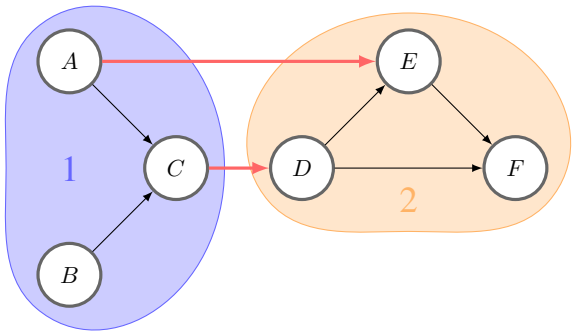


Figure 1: A causal system existing of two regions with three nodes each. The applied indexing varies between regions. The edges in red denote a cross-index causal relation referring to an aggregation function.

A first instantiation of temporal indexing is fixed-frequency in which the data is recorded with a defined interval possibly using a fixed starting point. An extension to temporal indexing is event-based indexing, where a data-point is recorded when a change occurs in relation to time. The referenced event can both occur on the recorded variable itself (on-change) or when a related variable changes (triggered). Instance-based indexing refers to an index chosen based on the instance / group under observation. Similar to event-based indexing, this can be the first value when the instance changes, but other aggregations are also available. One could take a feature (e.g. maximum, mean, last, etc.) of the measurement during the duration of a certain observation of an instance e.g. the highest recorded body temperature of a patient or the average sensor value in a manufacturing batch. Spatial indexing occurs in three main forms: (i) Fixed encoder-based sensors record measurements on a fixed distance-interval i.e. the measurement frequency is based on the production line speed, (ii) moveable sensors that include a position recording alongside measurement, and (iii) multiple identical sensors distributed across locations.

Except for event-based indexing, variations can occur within a single index type in terms of the scale of the indexing. The index can be aggregated to a broader multi-level index e.g. seconds to hours to days, instance to group, or cm to meters. The selection of the scale of the index is based on the rate of change of the variable, measurement frequency of the sensor, and storage limitations. Even if both index type and scale agree, two tables can differ in indexing for both the temporal and spatial index types through index misalignment. While the scale is identical, the resulting index values will be different due to offsets in the start reference (Faircloth, 2014).

1.1.1. MOTIVATIONAL EXAMPLE OF A MULTI-INDEX CAUSAL SYSTEM

Consider the system illustrated in Figure 1 composed of two regions with different indexing schemes. Variables are represented as nodes, and arrows indicate directional causal relations between the nodes forming a Graphical Causal Model (GCM). Region 1 is indexed by time, recording values for nodes A, B, and C at regular intervals. Region 2 is indexed by instance, assigning a single value to nodes D, E, and F per instance. Each instance in Region 2 spans a time interval connected to

Region 1 through a common index, enabling aggregation of time-based data into instance-level features. In this example, the graph defines relations from time-based nodes A and C to instance-based nodes E and D, respectively, crossing index boundaries.

1.1.2. RESAMPLING AND INDEX ALIGNMENT

Most data-driven modeling methods require data to be structured as a row-based dataset, where each row represents a one-to-one correspondence between features and targets. Consequently, HID are typically consolidated into a single dataset during an initial preprocessing step that involves aggregation and merging of multiple tables. This preprocessing is common practice, and datasets released for benchmarking and reproducibility are usually published in this aggregated form, with few exceptions. A widely used strategy for achieving this unified format is resampling, where the data from one region is adjusted to align with the index of another. Considering the motivational example, Upsampling is performed by resampling the lower-frequency data from Region 2 to match the index of Region 1. Depending on the data characteristics, an appropriate interpolation technique is selected to infer missing values. However, a first set of issues arise with this approach: (i) interpolation may introduce patterns or apparent structures that do not reflect true relationships in the data, potentially misleading downstream models; (ii) the data size increases, which can impact computational efficiency; and (iii) the ratio of meaningful data to total data points decreases, potentially diluting signal strength. Alternatively, downsampling reduces the higher-frequency data from Region 1 to match the lower-frequency index of Region 2. This is achieved by grouping values based on the given aggregation index and extracting features from the resulting set. Manual feature definitions are possible, but automated feature extraction and selection can also be applied. A key drawback of downsampling is the loss of information due to the compression of the data during preprocessing.

Limitation 1: Resampling enables merging into a unified table with common index but introduces several challenges, particularly obscuring the relations between variables and the information loss due to aggregation.

1.2. Representative Surrogates Modeling of Real-World Systems

Data-driven surrogate modeling, uses observed data to train regression models to predict system behavior. These models enable efficient exploration of the design space and reduce the number of expensive evaluations required of a system. Surrogate models are applied in optimization, sensitivity analysis, uncertainty quantification, and explainability of complex systems. The objectives Y^k can be defined as a subset of all endogenous parameters Θ^i. The surrogate f_k is defined as an approximation of the objective values Y^k as a function of the features X^l, where X^l is a combination of all remaining endogenous parameters Θ^i and the decision variables V^j:

$$Y^k \subset \Theta^i \tag{1}$$

$$X^l \subset (\Theta^i \setminus Y^k) \cup V^j \tag{2}$$

$$f_k : X^l \to Y^k \tag{3}$$

with $k \leq i$ and $l = i - k + j$.

The surrogate models approximate the real-world objective function f_k based on the features X^l:

$$\hat{f}_k(x) = f_k(x \in \mathbf{X^1}) + \epsilon \tag{4}$$

where ϵ captures the approximation error.

The surrogate can be used for time- and inference-efficient Multi-Objective Optimization (MOO) through an iterative feedback loop between the surrogate and the actual process. Only the decision variables V^j should be selected as features, as they are fully independent and controllable for optimization (De Temmerman et al., 2025). Concerns about the validity of the underlying reasoning can be raised when employing data-driven models as a surrogate and whether the model truly reflects the system response. A challenge arises from the complexity of high-dimensional regressor models used for inference on large datasets. While such models can capture intricate patterns, their robustness to out-of-domain data remains questionable. Moreover, single-model surrogates often suffer from low explainability due to high data correlation, which can obscure meaningful relationships (Gutmann, 2001; Blank and Deb, 2022).

1.3. GCMs as surrogate models

GCMs used as surrogate models for optimization of expensive-to-explore systems bring additional advantages over closed-form models in the form of robust representation due to causal constraints, accurate uncertainty quantification, interpretable model reasoning, and multi-target modelling for multi-objective problems. Maintaining surrogate model that closely resemble the true system dynamics is essential for reliability and interpretability (Koller and Friedman, 2009; D'Souza et al., 2010; De Temmerman et al., 2025). Probabilistic Graphical Models (PGMs) represent conditional dependencies between variables in a directed graph. If the graph is acyclic, forming a Directed Acyclic Graph (DAG), the model is known as a Bayesian Network (BN). GCMs extend PGMs by encoding causal relationships rather than mere probabilistic dependencies. GCMs formalize causal relationships using DAGs, where nodes represent variables and edges denote direct causal influence. Under the causal Markov condition, the joint distribution factorizes as:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{Pa}(X_i))$$

where $\text{Pa}(X_i)$ are the parents of X_i in the graph.

Functional Causal Models (FCMs) extend GCMs by associating each variable with an equation that defines its value as a function of its parents and an exogenous noise term:

$$X_i = f_i(\text{Pa}(X_i), N_i)$$

where f_i is a deterministic function and N_i is an independent noise variable.

Employing GCMs as surrogate for optimization enables robust and interpretable models due to several features such as identification of causal effects, aleatoric noise localization, and counterfactual reasoning (Pearl, 2009; Spirtes et al., 2000; Blöbaum et al., 2024). While GCM mitigates some concerns over close-form regressors, employing them as surrogates still leaves several issues open.

Limitation 2: Some relations can already be known and do not require to be estimated during surrogate building. Current GCM frameworks do not allow embedding known information of node

relations as explicit functions, as is possible with Structural Equation Modelings (SEMs) (Bollen, 1989).

Limitation 3: Current GCM frameworks do not allow incremental or distributed training of the underlying FCMs when working with HID. This brings the disadvantage of requiring data resampling and merging to a unified dataset.

As main contribution, this paper introduces **Composite Graphical Causal Models (CGCMs)**. It extends GCMs with the capability of handling complex, real-world HID by building interpretable surrogates, mitigating the need for lossy preprocessing of resampling data to a uniform indexing is discussed further in Section 3.

The remainder of this paper is structured as follows. Section 2 takes a broader look at handling HID for data-driven modelling, reviewing the limited related work on causal inference with HID. The method is validated using synthetic datasets where Section 4 outlines the experimental setup and Section 5 discusses the results. Section 6 concludes with key insights and future directions. Appendix E provides more details on the synthetic dataset generator for reproducibility.

2. Related Work

Despite extensive research focus on causal inference across both synthetic and real-world data, and in a large range applications domains, the large majority of the current techniques uses single-indexed datasets (Scutari, 2010; Dau et al., 2019; Runge et al., 2020; Krauß et al., 2023). Many datasets originate from heterogeneous-indexed data sources (e.g., ALARM (Beinlich et al., 1989) or barley (Rasmussen et al., 1995)) but are shared only in aggregated form. This is evident from feature names such as `minvol`, which indicate downsampled time-series data. Ideally, original non-aggregated datasets would be available for validation, but this is rarely the case. Single-indexed datasets are preferred due to the lack of HID-capable models and ease of sharing.

Some approaches encode an additional index directly as a node in the causal graph. While this may capture temporal patterns, it often obfuscates underlying relations due to inadvertently added information e.g. evaporation rate linked to time of day instead of air temperature. Such implementations appear in datasets like Mehra (Vitolo et al., 2018) and Hailfinder (Abramson et al., 1996). Encoding indexes as variables can lead to false dependencies and denser graphs, reducing accuracy in causal discovery and inference.

More recently, Pruthi and Jensen (2025) introduced a method that utilizes compositional models to estimate potential outcomes and causal effects within structured systems. This research demonstrates a compositional framework for assessing the impacts of interventions across various real-world heterogeneous indexed systems, achieving promising results with a hierarchical composition model that surpasses several existing baselines. The underlying motivation of this work aligns closely with our own, presenting an alternative methodology and application area. While the work shares our objective of addressing heterogeneous-indexed systems through composition, notable distinctions set the approaches apart. Their methodology emphasizes the estimation of causal effects via a learned hierarchical composition model that aggregates predictions from lower levels to derive higher-level outcomes. However, Pruthi and Jensen (2025) does not tackle the problem of heterogeneous systems, where unobserved variables may fluctuate across different indices (e.g., time of day, spatial location). Their framework treats composition mainly as an architectural decision rather than a mechanism for disentangling causal effects across heterogeneous indices. In contrast, our approach integrates index-aware causal inference mechanisms, enabling more precise modeling

of systems where causal relationships vary with indexing. In addition, their model presumes that composition can be derived solely from data, which poses challenges in high-dimensional, sparsely-indexed systems where certain index combinations may occur infrequently. Our methodology, however, incorporates domain knowledge regarding index structures and process dynamics to inform the decomposition, enhancing robustness against sparse observations across index combinations.

Rubenstein et al. (2017) discussed the idea that a system can be represented at different levels of detail which has an effect on the accuracy of interventions on the system. They introduce the concept of consistency of transformations between SEMs. One of their motivational examples formalize exact transformations between SEMs, using a measurement aggregation map τ and providing conditions under which marginalizing downstream or intermediate mechanisms and aggregating micro-variables preserves interventional distributions on selected targets. Relatedly, Beckers and Halpern (2019) define abstractions of causal models that map low-level variables and interventions to high-level ones, characterizing when high-level counterfactuals are faithful to the underlying system.

3. Composite Graphical Causal Model

This work proposes CGCMs, an approach that ensembles multiple local GCMs into a heterogeneous predictor while retaining core functionalities of causal inference frameworks and mitigating the limitations discussed in Section 1.

3.1. Formal Definition

From the HID, we partition the data into regions defined by unique indexes and associated variables. Let $\mathcal{R} = \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_k\}$ denote the set of regions, where each region $\mathcal{R}_j$ is associated with a subset of variables $\mathbf{V}^{(j)}$ and an index set $\mathcal{I}^{(j)}$. For each region, a region-GCM is constructed with its own DAG and data table extracted using the index region.

$$\mathcal{G}^{(j)} \equiv \left(V^{(j)}, E^{(j)}\right), \quad \text{where} \quad V^{(j)} = \{X_1^{(j)}, \ldots, X_{m_j}^{(j)}\}, E^{(j)} \subseteq V^{(j)} \times V^{(j)}. \tag{5}$$

with $E^{(j)}$, the directed causal edge set and m_j, the number of variables for region $\mathcal{R}_j$.

The disjoint union of all region graphs is defined as:

$$\mathcal{G}^u \triangleq \bigsqcup_{j=1}^{k} \mathcal{G}^{(j)} \tag{6}$$

By construction, these edge sets share no common nodes between each other and are fully disconnected:

$$V^{(j)} \cap V^{(\ell)} = \varnothing, \quad E^{(j)} \cap E^{(\ell)} = \varnothing \qquad \forall\, j \neq \ell. \tag{7}$$

3.2. Construction of region-GCM

Each region-GCM is trained on its respective region by fitting FCMs to the corresponding extracted table. These region-GCMs can be validated and refined individually before integration into the composite, ensuring local accuracy and interpretability. For each region-GCM $\mathcal{G}^{(j)}$, the structural equations follow the standard FCM form:

$$X_i^{(j)} = f_i^{(j)}\left(\mathrm{Pa}\left(X_i^{(j)}\right), N_i^{(j)}\right), \qquad i = 1, \ldots, m_j, \tag{8}$$

where $\mathrm{Pa}(X_i^{(j)})$ are the parents of $X_i^{(j)}$ in $E^{(j)}$, and $N_i^{(j)}$ are mutually independent exogenous noise variables. The induced observational distribution factorizes according to $\mathcal{G}^{(j)}$ (Blöbaum et al., 2024):

$$p\left(\mathbf{X}^{(j)}\right) = \prod_{i=1}^{m_j} p\left(X_i^{(j)} \middle| \mathrm{Pa}\left(X_i^{(j)}\right)\right), \tag{9}$$

3.3. Embedding Explicit Functions in GCMs

Pearl (2009) defines a FCM as a nonlinear, nonparametric generalization of the linear SEM. SEM formalizes (possibly causal) relations between observed and latent variables through explicit functions. Given the conceptual similarities between SEM and GCMs, hybrid approaches are feasible that embed explicit functions within GCMs (Pearl, 2009; Bollen, 1989).

Specifically, replacing FCMs estimators with domain-informed explicit functions $g_i^{(j)}$ can improve accuracy, robustness, and computational efficiency. The explicit functions can be extended with a given noise estimate $N_i^{(j)}$, if not the noise can be reconstructed from the observed data. Formally, for a chosen subset $\mathcal{S}^{(j)} \subseteq V^{(j)}$:

$$X_i^{(j)} = g_i^{(j)}\left(\mathrm{Pa}\left(X_i^{(j)}\right)\right) + N_i^{(j)}, \qquad i \in \mathcal{S}^{(j)}. \tag{10}$$

Enforcing explicit functions within GCMs can enhance interpretability and incorporate prior knowledge. However, incorrect specifications of these functions may introduce bias or reduce model flexibility. It is crucial to validate the choice of explicit functions through domain expertise based on the remaining noise estimate.

3.4. Embedding cross-index transformations in GCMs

Although GCMs are constructed for each region, they remain disconnected until cross-region relations are defined. Linking regions requires mapping instances across tables via their respective indexes and specifying transformations between these indexes.

Let $\phi_{j\to\ell} : \mathcal{I}^{(j)} \to \mathcal{I}^{(\ell)}$ be an index mapping from region j to region ℓ. When transitioning from higher to lower temporal or spatial resolution, a downsampling strategy groups high-frequency data and aggregates each group using a total of o specified functions:

$$T_{i,o}^{(\ell)}(s) = \mathcal{T}_{i,o}\left(\{X_i^{(j)}(t) | \phi_{j\to\ell}(t) = s\}\right) \qquad t \in \mathcal{I}^{(j)}, s \in \mathcal{I}^{(\ell)}, i \in \mathcal{S}^{(j)}. \tag{11}$$

where $\mathcal{T}_{i,o}$ is an aggregation operator (e.g. mean, sum). The intermediate node $T_{i,o}^{(\ell)}(s)$ represents one of the o transformed values from starting node $X_i^{(j)}$.

Conversely, the inverse transition defining an upsampling strategy matching low-frequency data to a high-frequency index via interpolation. $\mathcal{T}$ then represents an upsampling interpolation (e.g. linear, spline, nearest-neighbor). $\mathcal{T}$ can both be a predefined operator based on domain knowledge or an estimator trained from data. Appendix C details estimator methods for heterogeneous-indexed data.

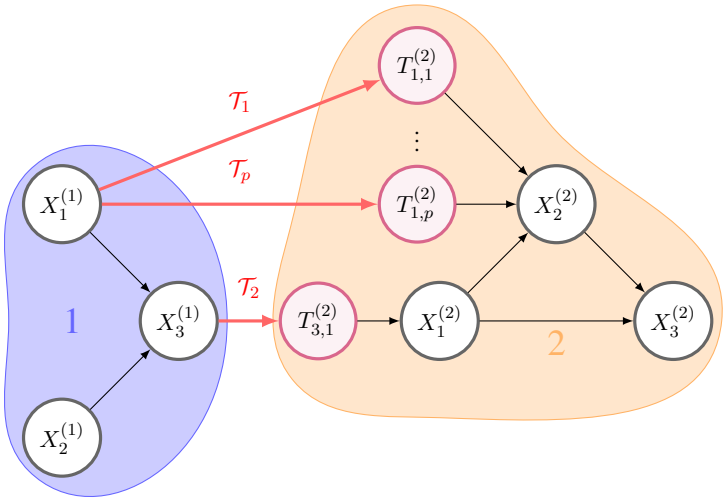


Figure 2: CGCM with two regions: Region 1 (blue) contains three nodes $X^{(1)}$, and a cross-index transformation $\mathcal{T}$ (red) via intermediate linking node $X^{(1)}_{t,3}$ connects to Region 2 (orange).

To represent these transformations explicitly, we introduce a set of intermediate linking nodes:

$$V_t^{(j)} = \{T_i^{(j)}\}.$$

The cross-region edge set then connects original nodes to linking nodes and linking nodes to target nodes:

$$E^{\mathrm{t}} \subseteq \bigcup_{m,n\in[1,k],m\neq n} \left(V^{(m)} \odot V_t^{(n)}\right) \cup \left(V_t^{(n)} \odot V^{(n)}\right) \tag{12}$$

The CGCM augments the disjoint union of region-GCM with cross-region edges induced by index mappings and transformations:

$$\mathcal{G}^c = \left(\bigsqcup_{j=1}^{k} V^{(j)} \cup V_t^{(j)}, \bigsqcup_{j=1}^{k} E^{(j)} \cup E^{\mathrm{t}}\right) \tag{13}$$

The proposed CGCM addresses the aforementioned limitations by enabling distributed learning on HID of GCM while mitigating the drawbacks connected to traditional data resampling strategies during processing. The architecture supports the incorporation of known relationships in the form of embedded explicit functions, thus reducing the reliance on estimation during surrogate building. This allows a design that accommodates construction and validation of local region-GCMs independently, merging them into a unified composite structure through data-driven or explicit aggregations.

4. Numerical Experiments

4.1. Benchmark Selection

Albeit the large amount of benchmark datasets available for causal inference, both synthetic and real-world, in a large range of domains, none are suitable for evaluating our method, as they are

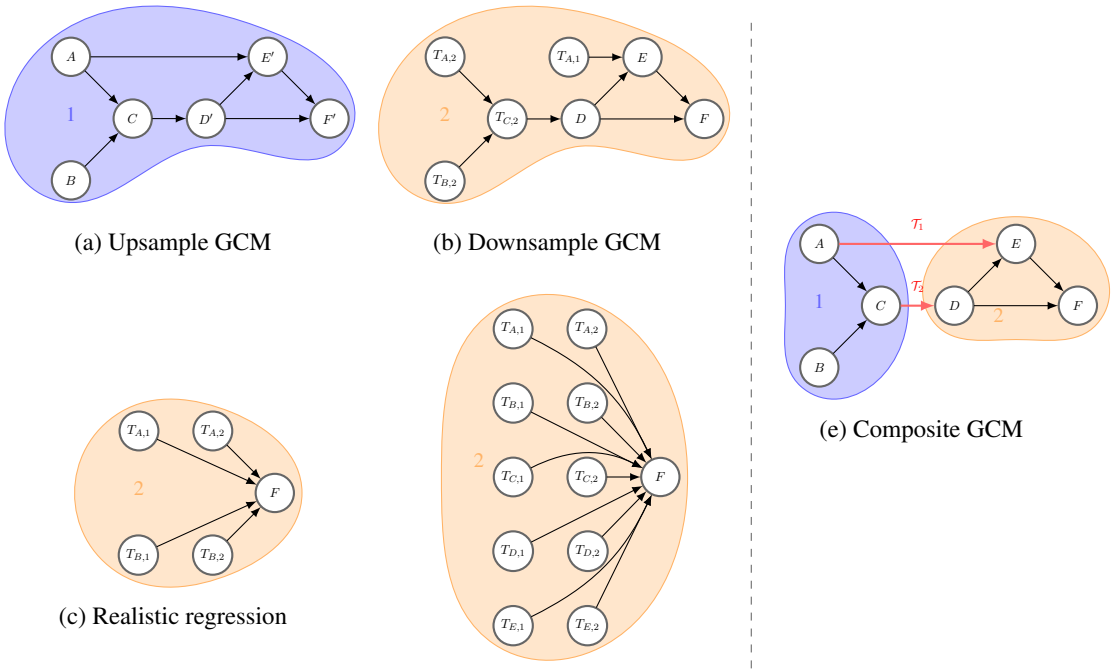


Figure 3: Baseline implementations (left) compared to the proposed CGCM (right) applied to motivational example of Fig 1.

single-indexed (Scutari, 2010; Dau et al., 2019; Runge et al., 2020; Krauß et al., 2023). Our validation uses four synthetic datasets, each featuring unique characteristics as described in Appendix E. These problems differ in causal structure and in the number and type of aggregation functions applied. For each problem, three variants are constructed with increasing complexity of inter-node relations: (i) linear relations only, (ii) quadratic and multivariate relations, and (iii) quadratic and multivariate relations with added Gaussian noise.

4.2. Experimental Setup

The proposed method is compared against three baselines and an oracle approach, illustrated in Figure 3. The aggregation used in the synthetic data generator is predefined and available to both the downsampling baselines and the proposed CGCM.

- Upsample GCM : The lower-frequency data is upsampled or padded by repeating rows to match the index of the higher-frequency region. Upsampled nodes are denoted by $(D \to D')$. A standard GCM is applied after resampling the data into a single table.
- Downsample GCM : The higher-frequency data is aggregated to match the lower-frequency index. Graphs with multiple aggregations apply each function to every upstream node, splitting original nodes into multiple derived nodes (e.g., $A \to T_{A,1}, T_{A,2}$). As with Upsample GCM, a GCM is applied to the single table.

- Oracle regression: All nodes of both the exogenous and the endogenous parameters are selected as features. This method is unrealistic due to having access to information of the endogenous nodes, nodes that contain additional information compared to the exogenous nodes that is unavailable in practice.

- A realistic regression is added using only features from the exogenous nodes as input. This regression only has a more limited access to the available data than the oracle.

For both regression baselines, upstream nodes are split according to each aggregation in the GCM. All methods use a "Bucket of Models" ensemble model, selecting the best performing model and hyperparameters from a linear regressor and Histogram-based Gradient Boosting Regression Tree (Ke et al., 2017) for the FCMs. From the synthetic data generation, we create a training set ($n = 1000$) and a test set ($n = 200$), repeated across 20 semi-random seeds. Predictions from each method are compared to the test set using RMSE normalized by the Interquartile Range (IQR-RMSE). Several other distance metrics where considered, including KL Divergence (Wang et al., 2009), AIC/BIC (Akaike, 1974; Schwarz, 1978), and the F-test (Fisher, 1922). RMSE was chosen because the evaluation focuses on mean prediction accuracy rather than distribution accuracy. The normalization by the IQR of the node data distribution ensures comparability across variables with different scales. Note that we compare mean prediction instead of full prediction distributions for computational efficiency. No significant changes in distribution accuracy are anticipated, as this is primarily determined by the noise model rather than the underlying regressor.

5. Results and Discussion

The CGCM method is compared against three baselines and one oracle method across twelve synthetic datasets (four problems, each with three variants). The aggregated IQR-RMSE scores, computed against the test dataset, are visualized in Figure 4.

The oracle regression achieves near-zero IQR-RMSE score for all problems, with slight increases in the third variant due to added noise. This is an expected outcome since this baseline has access to more information than the other methods. The intermediate node data is given as a feature in the test dataset. In contrast, other methods either estimate these values (upsample, downsample, CGCM) or exclude them entirely (realistic regression). The overall weakest performer is the upsample GCM. This method suffers from a sparse dataset with introduces apparent structure due to the interpolation, disregarding relations inherent across the index regions. Consequently, feature-label alignment is compromised due the chosen interpolation method by requiring predictions without the full feature set available to the estimator. This method is excluded from further discussion.

The first two problems, P and Q, are generally well estimated by all methods, except for P2 and P3. Here, the combination of the used aggregation (sum) with non-linear relations obscures the dependencies between exogenous node A and its child nodes for all baselines. All three baselines are subjected to information loss, either through resampling or by omission of unattainable data. This results in a worse approximation of the final objective and a worse IQR-RMSE score. IQR-RMSE seems mainly dependent on the problem, whereas the added noise does not significantly affect mean prediction accuracy, as seen when comparing variants 2 and 3. In contrast, for problem Q, even with a more complex causal structure, the apparent simpler aggregations allow the classic baseline methods to perform competitively with the oracle and CGCM, albeit with a slight deficit. The CGCM

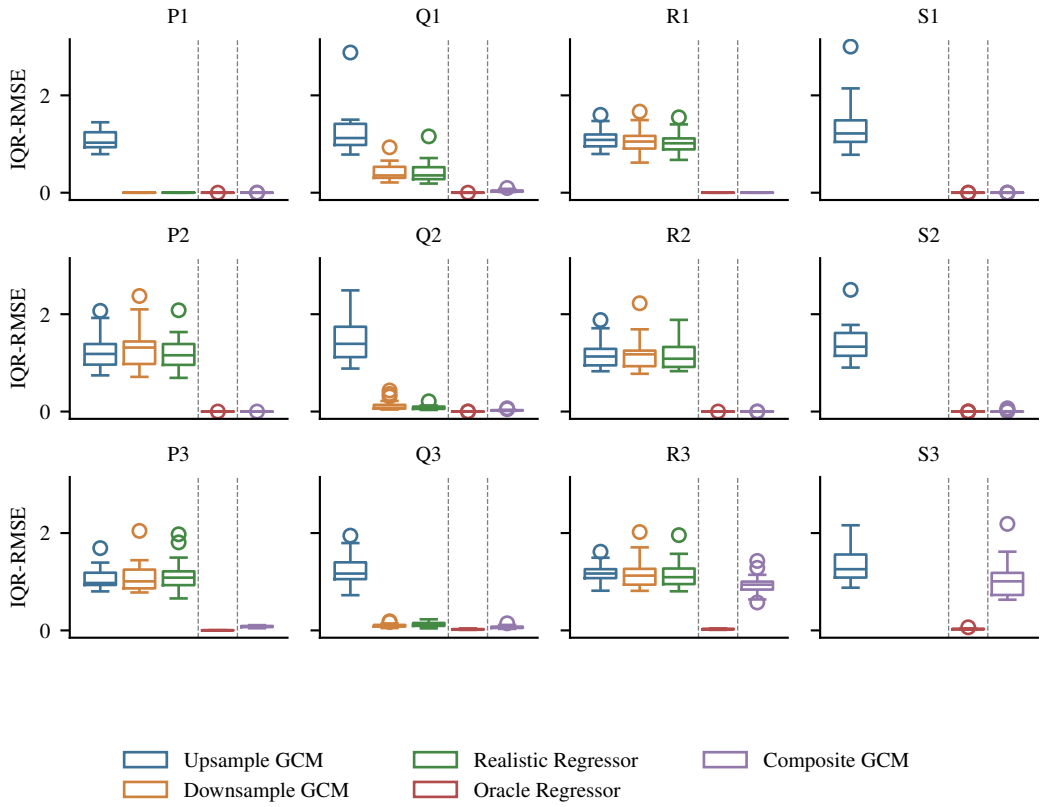


Figure 4: IQR-normalized RMSE between predicted and ground truth values of the end node for three baselines, one oracle, and the proposed CGCM. Graphs are ordered column-wise by problem P-S and row-wise by problem variants 1-3. Evaluation performed on a separate test dataset ($n = 200$). Lower values indicate better performance.

shows a significant improvement over the baselines for the first two variants of the problem R, highlighting the added value of embedding explicit aggregation functions within the GCM. Variant 3 again shows improvement, though less pronounced. Downsample GCM and realistic regressor cannot handle multi-variate aggregation functions due to fundamental incompatibility of the baseline definitions for Problem S. Similar to Problem R, the remaining baseline performs much worse for the first two variants than the oracle and CGCM methods. The proposed method, CGCM equals the oracle in all but two problems, R3 and S3, showing an improved approximation of heterogeneous-indexed dataset compared to standard approaches of resampling the dataset before applying it to model. The result still improves slightly for R3 and S3, compared to the other baselines but does not reach the oracle regression. This worse result can be attributed to effect of the added noise in the first region on the complex aggregation. These slight deviations cause large changes in the second index region, resulting in worse results. Oracle performance remains unaffected since node C is directly available as a feature.

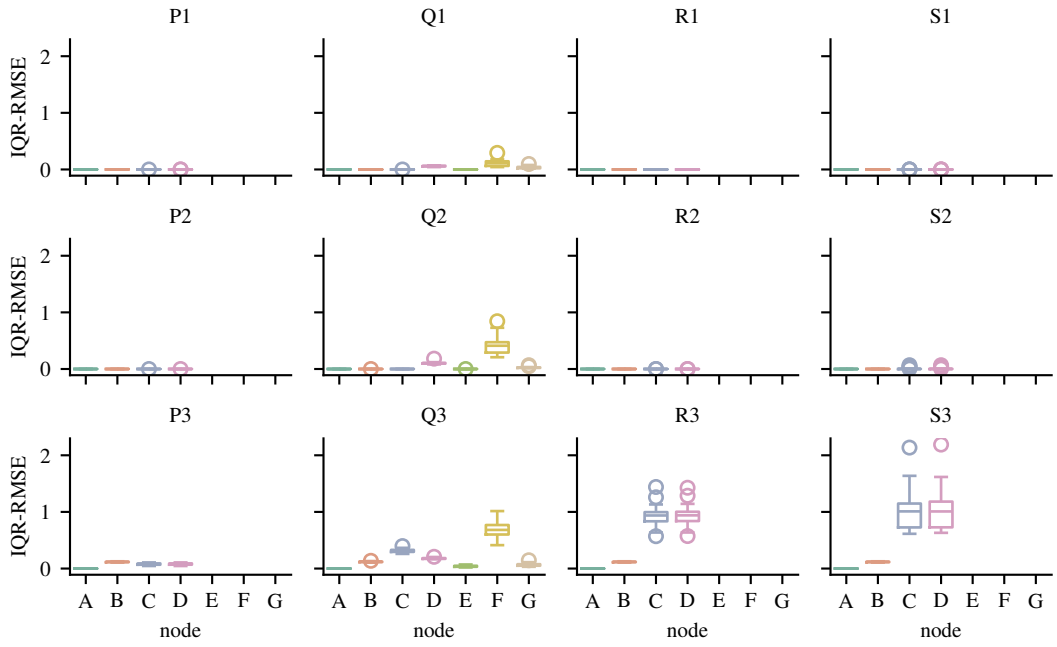


Figure 5: The IQR-RMSE between predicted and ground truth for all nodes in CGCM. Graphs are ordered for problems P-S with variants 1-3. Lower values indicate better performance.

Figure 5 show the deviations of the predicted value for each node in the CGCM compared to the ground truth. Notably, nodes C in variants R3 and S3 exhibit higher errors compared to upstream nodes (e.g., B), suggesting a bad predictive performance after the aggregation function. In practice, this issue could be mitigated by decomposing complex aggregation functions into sequential nodes, enabling finer-grained error analysis and improved interpretability.

6. Discussion and conclusion

Over the range of compared methods and synthetic problems, we observed that the proposed Composite Graphical Causal Model (CGCM) consistently outperforms traditional resampling-based approaches for handling Heterogeneous-Indexed Data (HID). These results indicate that an information loss occurs in traditional methods due to prior resampling that does not appear in the proposed method, which directly incorporates the heterogeneous structure of the data into the model. The causal structure that these earlier methods use is not representative for the aggregated features after resampling, as the conditional dependencies do not hold after the resampling step.

$$P(Y|X) \neq P(\mathcal{T}(Y)|\mathcal{T}(X)) \tag{14}$$

where $\mathcal{T}$ is the resampling transformation. Appendix B provides an error analysis showing the mechanism behind this discrepancy. This discrepancy leads to suboptimal performance in surrogate modeling tasks, as evidenced by the higher accuracy scores observed in baseline methods compared to the Composite Graphical Causal Model (CGCM). By explicitly modeling the relationships

and transformations between different index regions, the CGCM preserves the underlying causal structure and enhances predictive accuracy.

This work addresses the challenge of heterogeneous data and its prevalence in real-world applications for system modeling for optimization and control. We demonstrated that current resampling techniques introduce significant limitations, reducing the accuracy and interpretability of surrogate models. To address these challenges, we proposed an enhancement to Graphical Causal Models (GCMs) in the form of a CGCM by embedding known relations, inspired by Structural Equation Modeling (SEM), for integrating aggregation functions across multi-indexed regions. Our approach introduces a modular, piecewise construction of a composite model, offering an intuitive framework for handling complex heterogeneous datasets for extracting actionable insights and driving process optimization. These improvements resulted in notable gains in robustness and predictive accuracy for heterogeneous-indexed systems.

Acknowledgments

All code to reproduce the research results is available through gitlab.kuleuven.be/m-group-campus-brugge/dtai_public/publications/compositeGCM.

Funding This research is supported by Internal Funds KU Leuven (STG/21/057).

Author contributions statement A.D.T conceptualized and formulated the research goal, designed the methodology, conducted the experiments, and analyzed the results. All authors reviewed the manuscript.

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

References

Bruce Abramson, John Brown, Ward Edwards, Allan Murphy, and Robert L. Winkler. Hailfinder: A Bayesian system for forecasting severe weather. International Journal of Forecasting, 12(1): 57–71, March 1996. ISSN 0169-2070. doi: 10.1016/0169-2070(95)00664-8.

H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723, December 1974. ISSN 1558-2523. doi: 10.1109/TAC.1974.1100705.

Sander Beckers and Joseph Y. Halpern. Abstracting Causal Models. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):2678–2685, July 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33012678.

Ingo A. Beinlich, H. J. Suermondt, R. Martin Chavez, and Gregory F. Cooper. The ALARM Monitoring System: A Case Study with two Probabilistic Inference Techniques for Belief Networks. In O. Rienhoff, D. A. B. Lindberg, Jim Hunter, John Cookson, and Jeremy Wyatt, editors, AIME 89, volume 38, pages 247–256, Berlin, Heidelberg, 1989. Springer Berlin Heidelberg. ISBN 978-3-540-51543-2 978-3-642-93437-7. doi: 10.1007/978-3-642-93437-7_28.

Julian Blank and Kalyanmoy Deb. GPSAF: A Generalized Probabilistic Surrogate-Assisted Framework for Constrained Single- and Multi-objective Optimization, April 2022.

Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A. Mastakouri, and Dominik Janzing. DoWhy-GCM: An Extension of DoWhy for Causal Inference in Graphical Causal Models. Journal of Machine Learning Research, 25(147):1–7, 2024. ISSN 1533-7928.

Kenneth A. Bollen. Structural Equations with Latent Variables. Structural Equations with Latent Variables. John Wiley & Sons, Oxford, England, 1989. ISBN 978-0-471-01171-2. doi: 10.1002/9781118619179.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The UCR time series archive. IEEE/CAA Journal of Automatica Sinica, 6(6):1293–1305, November 2019. ISSN 2329-9274. doi: 10.1109/JAS.2019.1911747.

Arne De Brabandere, Tim Op De Beéck, Kilian Hendrickx, Wannes Meert, and Jesse Davis. TSFuse: Automated feature construction for multiple time series data. Machine Learning, 113(8):5001–5056, August 2024. ISSN 1573-0565. doi: 10.1007/s10994-021-06096-2.

Arne De Temmerman, Matthias De Ryck, and Mathias Verbeke. Handling uncertainty with parametric surrogate-assisted optimization for dynamic multi-objective problems. Neural Computing and Applications, June 2025. ISSN 1433-3058. doi: 10.1007/s00521-025-11359-3.

Rio G. L. D'Souza, K. Chandra Sekaran, and A. Kandasamy. Improved NSGA-II Based on a Novel Ranking Scheme, February 2010.

Jeremy Faircloth. Chapter 4 - databases. In Jeremy Faircloth, editor, Enterprise Applications Administration, pages 131–173. Morgan Kaufmann, Boston, 2014. ISBN 978-0-12-407773-7. doi: 10.1016/B978-0-12-407773-7.00004-1.

R. A. Fisher. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222(594-604):309–368, 1922. doi: 10.1098/rsta.1922.0009.

H.-M. Gutmann. A Radial Basis Function Method for Global Optimization. Journal of Global Optimization, 19(3):201–227, March 2001. ISSN 1573-2916. doi: 10.1023/A:1011255519438.

Jochen Hanssens. Decision Support for Tomato Growers Based on Plant Responses, Modelling and Greenhouse Energy Consumption. PhD thesis, Ghent University, 2015.

Michiel Jacobs, Lode Vuegen, Tom Verresen, Marie Schouterden, David Ruttens, and Peter Karsmakers. Exploring Model Architectures for Real-Time Lung Sound Event Detection. In ESANN 2025 Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2025.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

Daphne Koller and Nir Friedman. Probabilistic Graphical Models: Principles and Techniques. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2009. ISBN 978-0-262-01319-2.

J. Krauß, T. Hülsmann, L. Leyendecker, and R. H. Schmitt. Application Areas, Use Cases, and Data Sets for Machine Learning and Artificial Intelligence in Production. In Mathias Liewald, Alexander Verl, Thomas Bauernhansl, and Hans-Christian Möhring, editors, Production at the Leading Edge of Technology, pages 504–513, Cham, 2023. Springer International Publishing. ISBN 978-3-031-18318-8. doi: 10.1007/978-3-031-18318-8_51.

Judea Pearl. Causality. Cambridge University Press, September 2009. ISBN 978-0-521-89560-6.

Purva Pruthi and David Jensen. Compositional Models for Estimating Causal Effects. In Proceedings of the Fourth Conference on Causal Learning and Reasoning, pages 1365–1404. PMLR, June 2025.

Yuanyuan Pu, Alicja Szmigiel, and Derek B. Apel. Purities prediction in a manufacturing froth flotation plant: The deep learning techniques. Neural Computing and Applications, 32(17):13639–13649, September 2020. ISSN 1433-3058. doi: 10.1007/s00521-020-04773-2.

Ilse Ankjær Rasmussen, Kristian Kristensen, and Sten Stetter. Growing malting barley without the use of pesticides: Integrated Crop Protection - Toward Sustainability? BCPC Symposium Proceedings, 63:431–438, 1995.

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal Consistency of Structural Equation Models, July 2017.

Jakob Runge, J Muñoz-Marí, G Mateo, and G Camps-Valls. CauseMe: An online system for benchmarking causal discovery methods., 2020.

Gideon Schwarz. Estimating the Dimension of a Model. The Annals of Statistics, 6(2):461–464, March 1978. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176344136.

M. Scutari. Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software, 35(3):1–22, 2010.

Peter Spirtes, Clark N. Glymour, and Richard Scheines. Causation, Prediction, and Search. MIT Press, 2000. ISBN 978-0-262-19440-2.

Claudia Vitolo, Marco Scutari, Mohamed Ghalaieny, Allan Tucker, and Andrew Russell. Modeling Air Pollution, Climate, and Health Data Using Bayesian Networks: A Case Study of the English Regions. Earth and Space Science, 5(4):76–88, 2018. doi: 10.1002/2017EA000326.

Qing Wang, Sanjeev R. Kulkarni, and Sergio Verdu. Divergence Estimation for Multidimensional Densities Via k-Nearest-Neighbor Distances. IEEE Transactions on Information Theory, 55(5): 2392–2405, May 2009. ISSN 1557-9654. doi: 10.1109/TIT.2009.2016060.

Appendix A. Scalability of CGCM

Composite Graphical Causal Model (CGCM) affects scalability compared to baseline GCMs in several ways. As the number of regions increases, the composite graph may grow rapidly, as is the case with any graphical model. Below, we discuss the benefits and limitations of this approach in terms of data and computational resources.

Benefits

- Causal Discovery: The proposed method improves scalability during graph construction and learning of larger systems, as the local causal structures within each region can be learned independently before being integrated into the composite graph.
- Causal Inference: By maintaining the optimal index for each region, we ensure that the number of Functional Causal Model (FCM) evaluations remains minimal during causal inference. Specifically, the number of inferences is reduced from being proportional to the total number of nodes and the global frequency to being adaptive to the number of regions and their local frequencies. This enables more efficient inference in systems with varying temporal dynamics across regions.

Limitations

- Causal Discovery: As the number of regions increases, the number of potential cross-region interactions may grow exponentially, leading to a rapid increase in the number of edges in the composite graph. This paper does not examine the discovery of the causal graph structure in depth, as we recognize that this constitutes a substantial challenge on its own. Extending existing time-series causal discovery methods with automated feature extraction is an interesting direction for future research. However, the computational cost of time-series feature extraction and selection is relatively high. Combining time-series causal discovery, which is already computationally intensive, with feature extraction methods may lead to practical limitations in terms of runtime and memory usage. Possible solutions include more efficient algorithms based on sparse representations, dimensionality reduction techniques, or optimal search strategies (De Brabandere et al., 2024).

Appendix B. Dependencies during Resampling for Problem P2

This appendix provides a detailed analysis of the errors introduced during resampling, complementing the main results. The focus is on how the relationships between nodes change during the resampling process and how these changes affect the overall error. As an illustration, we consider Problem P2, where the baselines exhibit notably worse performance compared to the oracle and CGCM methods, as shown in Figure 4.

Problem P2 is a four-node problem with a quadratic relationship between A and B, a sum aggregation from B to C, and a linear transformation from C to D. The data are generated with $n = 1000$ samples and $m = 20$ aggregation groups, each containing $k = n/m = 50$ values. The data generation process is defined as follows:

$$A_i \sim \mathcal{N}(0, 1) \text{ for } i = 0, \ldots, n-1 \tag{15}$$

$$B_i = 2A_i^2 \tag{16}$$

$$C_j = \sum_{i \in g_j} B_i \text{ with } j = 0, \ldots, m-1,\ g_j = [k * j, k * j + 1, \ldots, k * (j+1) - 1] \tag{17}$$

$$D_j = 4C_j \tag{18}$$

A sample is illustrated in Figure 6, showing the general shape of each variable's series.

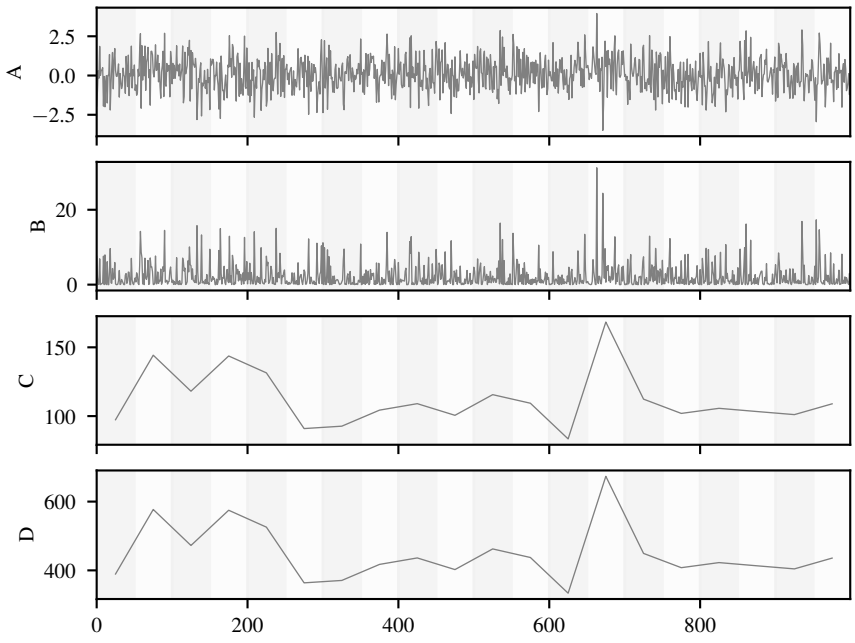


Figure 6: Sample Data for each node in Problem P2 (n=1000)

B.1. Upsampling

As explained in Section 1.1.2, upsampling creates new samples by interpolating between existing ones. Nodes C and D are resampled from 20 to 1000 points to match the number of samples in A and B. Equation 19 describes this process: each new sample is assigned the value of its corresponding group in the original data. Consequently, all samples within a group share the same value after upsampling, producing a piecewise constant structure.

$$C_i = C_j \tag{19}$$
$$D_i = D_j \tag{20}$$
$$\text{for } i \in g_j,\ j = 0, \ldots, m-1, g_j = [k*j, k*j+1, \ldots, k*(j+1)-1]$$

As shown in Figure 7, upsampling introduces an apparent square-wave structure with a period of 50 samples. This artifact is not present in the original data and is a direct consequence of the upsampling method.

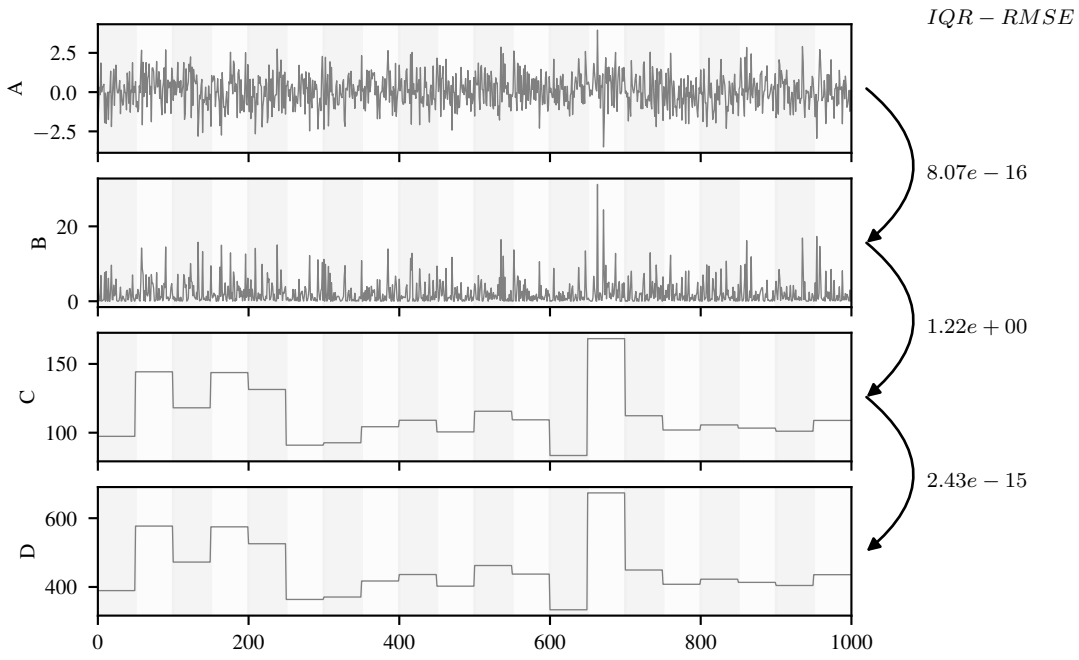


Figure 7: Original data (A and B) with upsampled data (C and D) for Problem P2 (n=1000)

The upsample GCM method models the relationship between B and C as a one-to-one mapping:

$$C_i = f(B_i) \text{ for } i = 0, \ldots, n-1 \tag{21}$$

This formulation is insufficient to capture the true relationship between B and C, because the remaining samples in B that belong to the same group as B_i are ignored. Figure 7 displays the RMSE normalized by the Interquartile Range (IQR-RMSE) of a Generalized Additive Model (GAM) for

each relation, serving as a metric for non-monotonic correlation between nodes. The GAM is a flexible regression model capable of capturing non-linear relationships. The IQR-RMSE values confirm that the upsampled data exhibits a significantly higher error between B and C, as expected from the information loss during upsampling.

B.2. Downsampling

Conversely, the downsampling process aggregates samples from A and B to produce new samples aligned with C and D. The aggregation method is a simple sum, where each new sample equals the sum of the corresponding group in the original data, as described by Equation 22.

$$C_j = \sum_{i \in g_j} C_i \tag{22}$$

$$D_j = \sum_{i \in g_j} D_i \tag{23}$$

$$\text{for } j \in 0, \ldots, m-1, g_j = [k * j, k * j + 1, \ldots, k * (j+1) - 1]$$

In Figure 8, the downsampled data for A and B show a significant loss of detail compared to the originals. The downsampled series are much smoother and fail to capture the variability present in the original data.

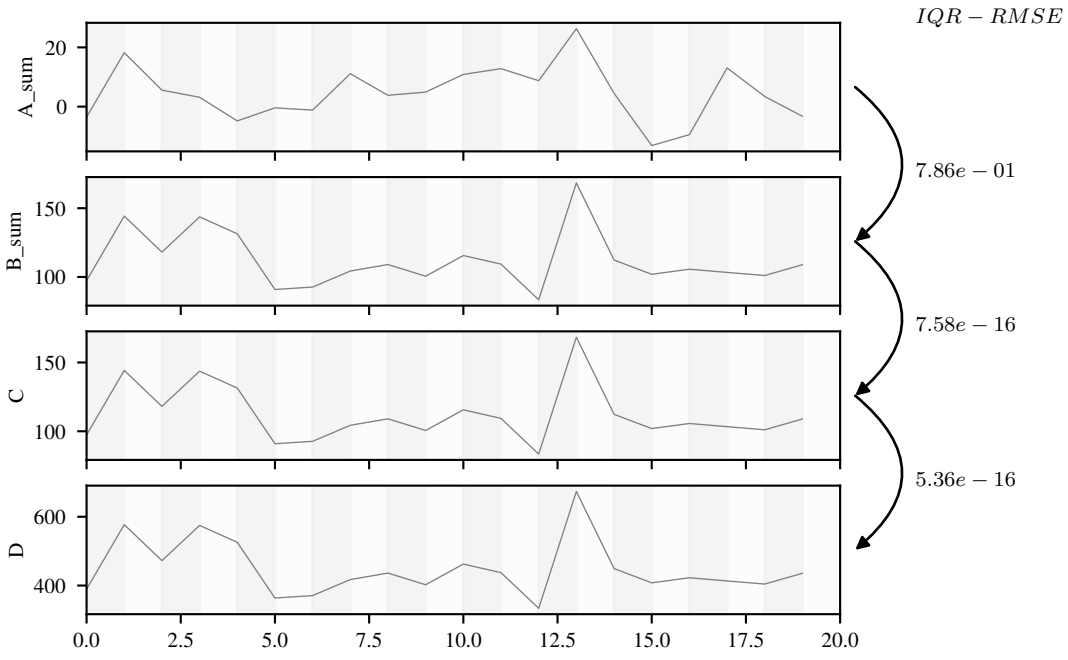


Figure 8: Original data (C and D) with downsampled data (A and B) for Problem P2 (n=20)

Compared to the upsampling method, the downsampled data yield a lower Root Mean Square Error (RMSE) between B and C, indicating a better fit of the GAM model. This is because the sum

feature extracted during downsampling is inherent to the function from B to C. In this example, no additional transformation is applied, so B is in fact equal to C.

$$C_j = f(\sum_{i \in g_j} B_i) \text{ for } j = 0, \ldots, m-1 \tag{24}$$

The error in the downsampling approach arises between A and B, where the downsampled data exhibit a higher RMSE compared to the other node pairs. The aggregation over groups leads to an information loss about individual samples and their pairwise relationships, because the original dependency does not hold after resampling. Specifically, the sum aggregation fails to preserve the non-linear (quadratic) relationship between A and B, leading to a significant increase in modeling error, as reflected by the higher RMSE for this relation.

$$P(Y|X) \neq P(\sum_{i \in g_j} Y_i | \sum_{i \in g_j} X_i) \tag{25}$$

The scatter plot in Figure 9 illustrates this loss of information: the original data clearly exhibit the quadratic relationship between A and B, whereas the downsampled data show no discernible dependency. This discrepancy highlights the limitations of the downsampling approach for capturing true relationships in heterogeneously indexed data, resulting in suboptimal surrogate modeling performance.

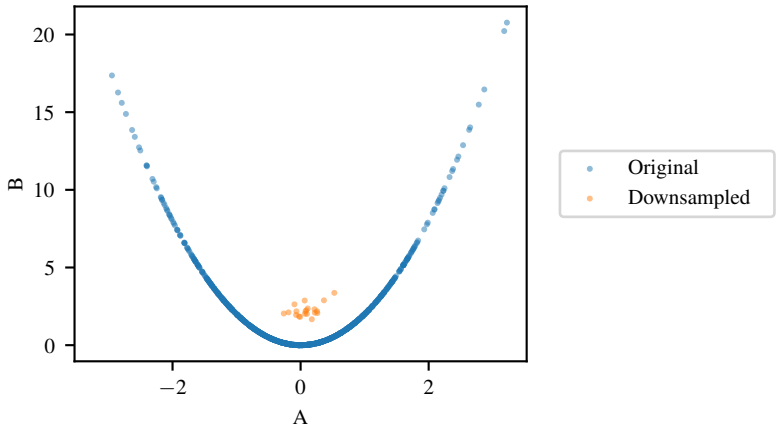


Figure 9: Scatter plot of original and downsampled data between A and B for Problem P2 (n=20)

The same analysis applies to the two remaining baselines, the oracle and realistic regression, and is therefore not repeated here. The error does not visibly affect the oracle method, because additional information from succeeding nodes compensates for the loss in preceding nodes, as discussed in Section 5.

Appendix C. Series-aggregation approximation

As discussed in Section 3.4, cross-index transformations between regions can be defined using explicit functions or learned estimators. This section details three approaches of these cross-index transformations: (1) explicit transformations based on domain knowledge, (2) explicit extractions combined with learned estimators, and (3) bag of extractions combined with learned estimators.

1. The first approach involves defining explicit transformations based on domain knowledge. For instance, if the relationship between two nodes across regions is known to be a specific mathematical function (e.g., mean, sum,max), this function can be directly embedded into the GCM.

2. The second approach combines explicit extractions with learned estimators. In this case, specific features are extracted from the source region based on domain knowledge identical to the first approach, but a learned estimator is trained to predict the target node in the destination region using these features. This allows for more flexibility than the first approach while still leveraging domain expertise.

3. The third approach utilizes a "bag of features", a list of available features, combined with learned estimators. Here, a wide range of features are extracted from the source region without strong assumptions about their relevance. A learned estimator is then trained to select and utilize the most informative features for predicting the target node in the destination region. This approach is particularly useful when domain knowledge is limited or when the relationships between nodes are complex and cannot be captured by explicit functions.

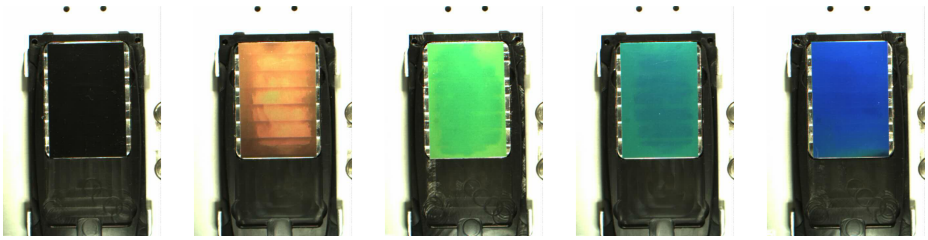


Figure 10: Five samples captured by the control system, sorted by increasing temperature.

Appendix D. Application case of Composite Graphical Causal Model (CGCM)

The proposed problem of heterogeneous data is prevalent in manufacturing due to the stepwise, distributed, and often parallelized structure of production and assembly processes. One application case study involves using a CGCM as a surrogate model for real-time, process multi-objective optimization for decision support.

The production process consists of three major steps: assembly, heating of the product, and quality control. It resembles other processes in several sectors, such as: food processing; plastic extrusion or injection molding; metal annealing, hardening, or sintering; and composite, polymer, or adhesive curing.

The problem has a two-level structure: a fast, time-indexed oven control loop and order-level product properties, with unknown physical relations and an unknown cross-index aggregation. The CGCM builds a composite model combining (i) a temporal GCM or dynamic Bayesian Network (BN) for setpoint T_s, oven temperature T_o, and binary heater state H with explicit hysteresis, and (ii) an order-level regression from time-series features of T_o to product temperature T_p, product appearance, and an explicit pass/fail cost. The appearance of the product is checked with an inline quality control system with samples shown in Figure 10.

D.1. Causal Representation

A causal graph for the system can be constructed using expert knowledge, as shown in Figure 11. Two distinct indexing levels can be discovered: (i) a time-based control loop for the heating process controlled through a closed-loop controller, and (ii) a higher-level record of product properties and quality parameters.

The time-based indexing data relates to the inline heating process. Three variables can be connected through a simple temperature control system: setpoint temperature T_s, measured oven temperature T_o, and binary heater on/off state H.

Based on the current oven temperature, we trace the causal path toward the objective value, product value V_p, derived from product color. This path includes product temperature T_p, product color expressed in HSV components, and finally the objective V_p.

D.2. Discussion

Several explicit relations can be embedded in the model based on system knowledge and human reasoning, such as internal oven control logic, the physical relation where energy equals the integral of power, and the cost function of the product based on color appearance. These relations are shown

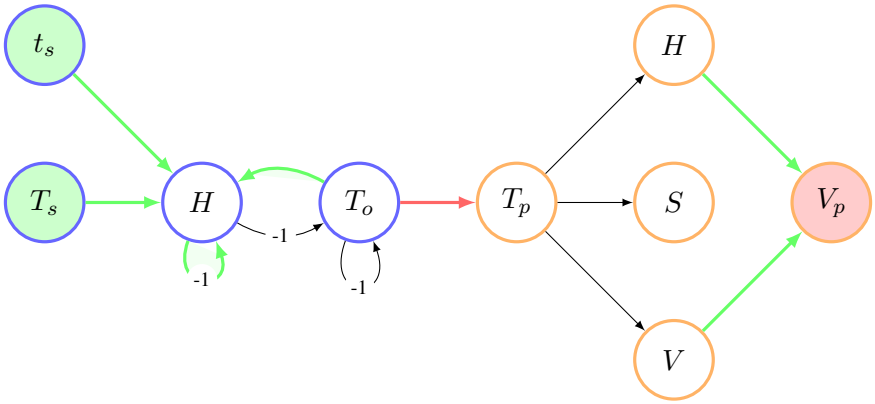


Figure 11: A causal graph of the case study system with eleven variables shown as nodes with the causal relations shown by edges between the nodes. Two decision variables (T_s,t_s) are highlighted by a green node background. The objective V_p is highlighted by a red background. The varying indexing is denoted through the edge color of the nodes for time-based (blue) and order-based (orange). The edges in red denote a cross-index causal relation referring to an aggregation function whereas green edges represent explicit relations. Numbered edges refer to relations that included in timelag between the relations where the number corresponds to the number of shifts based on the index.

as green edges in Figure 11. These explicit functions allow us to expand the data-based model with latent nodes, creating a hybrid representation of measured and derived values.

The advantage of handling Heterogeneous-Indexed Data (HID) is not to be understated in this case, as the time-based data and order-based data are not directly comparable due to the varying window sizes. The CGCM method allows us to combine these two data sources in a single model, without the need for complex data preprocessing or feature engineering to align the data. This is particularly beneficial in real-world applications where data may be collected at different frequencies or may have missing values. A CGCM is capable of capturing the complex structure required for the case study, combining known and unknown relations with an unknown aggregation function. The propagation of uncertainty, or compounding error, poses a significant challenge to the scalability and reliability of models representing real-world systems. Deviations resulting from both model error and embedded model noise, accumulated along the topological order of the causal graph, provide valuable insights into the reliability of the model's predictions and the robustness of the decision support system. The CGCM method allows one to analyze and mitigate this compounding error by providing a structured framework for understanding how errors propagate through the model and identifying potential sources of uncertainty.

Appendix E. Synthetic datasets and variations

Table 1: The Directed Acyclic Graph (DAG) colored by index region, the fixed aggregation functions are displayed for each cross-index relation.

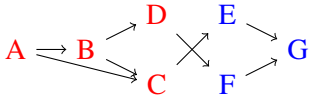
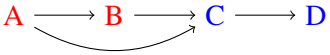
Problem	DAG	Aggregation function
P	A ⟶ B ⟶ C ⟶ D	$c = \sum \left(\{b_i\}_{i=1}^{n}\right)$
Q		$e = \min\left(\{c_i\}_{i=1}^{n}\right)$ $f = \text{med}\left(\{d_i\}_{i=1}^{n}\right)$
R	A ⟶ B ⟶ C ⟶ D	$c = \text{med}\left(\left\{\sum_{j=1}^{i} \sin\left(b_j^2\right)\right\}_{i=1}^{n}\right)$
S		$c = \text{med}\left(\left\{\sum_{j=1}^{i} \sin\left(\frac{a_j}{b_j^2}\right)\right\}_{i=1}^{n}\right)$

Table 2: Three variations in the complexity of the relations between the nodes.

Version	Relation characteristics
1	linear
2	quadratic, multi-variate
3	quadratic, multi-variate, random normal noise