
CAUSAL DISCOVERY REPORT ON CCS_DATA

TECHNICAL REPORT

ⓘ Causal Copilot

November 4, 2024

ABSTRACT

This report presents a causal discovery analysis of a dataset focused on concrete mix design and its influence on compressive strength, incorporating variables such as cement, aggregates, and hydration age. We utilized the PC algorithm, Greedy Equivalence Search (GES), and NOTEARS for causal structure identification, with hyperparameters optimized through a large language model (LLM) to adapt to the dataset's characteristics. Our findings reveal critical relationships, notably the significant roles of blast furnace slag, superplasticizer, and hydration age in determining concrete strength. The results emphasize the complex interplay of materials in concrete production, contributing to enhanced understanding and optimization of concrete mixes, ultimately providing valuable insights for future applications in construction.

Keywords Causal Discovery, Large Language Model, PC, Ccs_data

1 Introduction

The dataset under consideration is centered on concrete mix design and its impact on compressive strength, a crucial factor in structural integrity. It encompasses a range of variables including the amounts of cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and the age of the concrete. Each of these factors plays a significant role in influencing the overall strength of the concrete, with specific causal relationships that can be identified through analysis. Understanding the hydration processes, material properties, and established industry standards can enhance the models employed in causal discovery, leading to deeper insights into how varying concrete compositions affect their performance over time. This report aims to elucidate the causal mechanisms within this dataset, providing valuable knowledge for optimizing concrete mixes in future applications.

2 Background Knowledge

2.1 Detailed Explanation about the Variables

The dataset includes several key variables that pertain to concrete mix design and its resultant effects on compressive strength, with each variable playing a distinct role in determining the overall performance of the concrete.

- **Cement:** This variable signifies the quantity of cement utilized within the concrete mixture, typically quantified in kilograms (kg) or pounds (lb). Cement serves as the primary binding material, contributing significantly to the strength and stability of the final product.
- **Blast Furnace Slag:** Introduced as a supplementary cementitious material, blast furnace slag is a byproduct from iron production. It can partially replace cement in the mix, potentially enhancing the concrete's durability while mitigating the heat generated during hydration.
- **Fly Ash:** Fly ash, another byproduct, is derived from coal combustion. This material functions as a partial substitute for cement and is known to improve workability, reduce permeability, and bolster long-term strength characteristics of concrete.

- **Water:** The inclusion of water is vital for the hydration process of cement. Its quantity has direct implications for the workability and strength of the concrete, with the water-to-cement ratio being a critical determinant of the hardened product's quality.
- **Superplasticizer:** This agent, also referred to as a water-reducing additive, facilitates enhanced flow properties of concrete while allowing for the use of less water. It is particularly useful in achieving high strength even with lower water-cement ratios.
- **Coarse Aggregate:** Comprising larger particles, such as gravel and crushed stone, coarse aggregate is a fundamental component of the concrete mix. Its size distribution and grading are essential for influencing both the strength and workability of the mixture.
- **Fine Aggregate:** Fine aggregate typically consists of smaller particles, like sand, which fill the voids between coarse aggregates. The nature and composition of fine aggregates significantly affect the overall workability and strength of the concrete mix.
- **Age:** Representing the curing duration of the concrete after mixing and casting, age is a crucial variable measured in days. As concrete hydrates over time, its strength increases, making age a significant factor in compressive strength assessments.
- **Compressive Strength:** This variable quantifies the concrete's load-bearing capacity, usually expressed in megapascals (MPa) or pounds per square inch (psi). Higher values indicate superior performance for structural applications, making this measure essential in evaluating the quality of the concrete mix.

Understanding the characteristics and interactions of these variables provides a foundation for exploring potential causal relationships and developing insights into the factors that influence compressive strength in concrete mixes.

2.2 Possible Causal Relations among these Variables

- **Cement → Compressive Strength:** Increasing the amount of cement generally increases compressive strength due to more binding material being available.
- **Blast Furnace Slag → Compressive Strength:** Use of slag can enhance strength properties, especially as it may chemically react with calcium hydroxide to form additional cementitious compounds.
- **Fly Ash → Compressive Strength:** Similar to slag, fly ash can contribute to strength, particularly in long-term scenarios, altering the hydration process of the cement.
- **Water → Compressive Strength:** The water-to-cement ratio directly affects strength; too much water can weaken the concrete, while too little can prevent proper hydration.
- **Superplasticizer → Compressive Strength:** Allowing for a lower water-to-cement ratio without sacrificing workability can lead to increased strength.
- **Coarse Aggregate → Compressive Strength:** The quality and size distribution of coarse aggregates can influence the load-bearing capacity of concrete mixes.
- **Fine Aggregate → Compressive Strength:** Similar to coarse aggregates, the properties of fine aggregates affect overall strength as they contribute to the mix's density and bonding.
- **Age → Compressive Strength:** Concrete generally gains strength over time due to ongoing hydration reactions. Higher ages are associated with greater compressive strength.

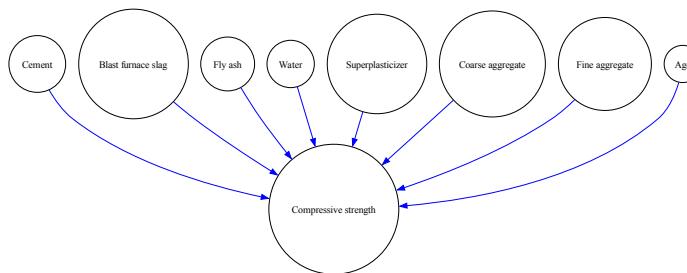


Figure 1: Possible Causal Relation Graph

3 Dataset Descriptions and EDA

The following is a preview of our original dataset.

Table 1: Dataset Preview

Cement	Blast furnace slag	Fly ash	Water	Superplasticizer	Coarse aggregate	Fine aggregate	Age	Compressive strength
540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.99
540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.89
332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.27
332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.05
198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.30

3.1 Data Properties

We employ several statistical methods to identify data properties.

The shape of the data, data types, and missing values are assessed directly from the dataframe. Linearity is evaluated using Ramsey's RESET test, followed by the Benjamini & Yekutieli procedure for multiple test correction. Gaussian noise is assessed through the Shapiro-Wilk test, also applying the Benjamini & Yekutieli procedure for multiple test correction. Time-Series and Heterogeneity are derived from user queries.

Properties of the dataset we analyzed are listed below.

Table 2: Data Properties

Shape ($n \times d$)	Data Type	Missing Value	Linearity	Gaussian Errors	Time-Series	Heterogeneity
(1030, 9)	Continuous		False	False	False	False

3.2 Distribution Analysis

The following figure shows distributions of different variables. The orange dash line represents the mean, and the black line represents the median. Variables are categorized into three types according to their distribution characteristics.

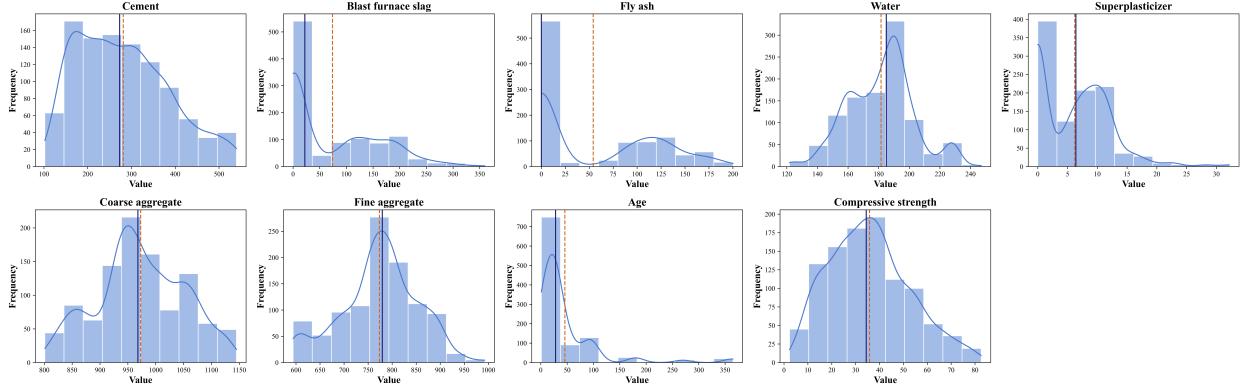


Figure 2: Distribution Plots of Variables

- Slight left skew distributed variables: None
- Slight right skew distributed variables: Blast furnace slag, Fly ash, Age, Compressive strength
- Symmetric distributed variables: Cement, Water, Superplasticizer, Coarse aggregate, Fine aggregate

3.3 Correlation Analysis

In this analysis, we will categorize the correlation statistics of features in the dataset into three distinct categories: Strong correlations, Moderate correlations, and Weak correlations.

- Strong Correlated Variables: None
- Moderate Correlated Variables: Superplasticizer and Water
- Weak Correlated Variables: None

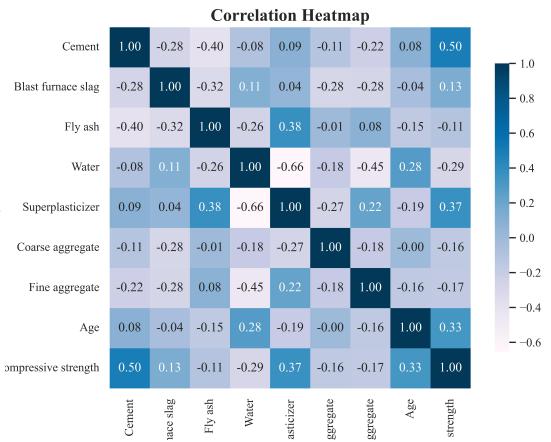


Figure 3: Correlation Heatmap of Variables

4 Discovery Procedure

In this section, we provide a detailed description of the causal discovery process implemented by Causal Copilot. We also provide the chosen algorithms and hyperparameters, along with the justifications for these selections.

4.1 Data Preprocessing

In this initial step, we preprocessed the data and examined its statistical characteristics. This involved cleaning the data, handling missing values, and performing exploratory data analysis to understand distributions and relationships between variables.

4.2 Algorithm Selection assisted with LLM

Following data preprocessing, we employed a large language model (LLM) to assist in selecting appropriate algorithms for causal discovery based on the statistical characteristics of the dataset and relevant background knowledge. The top three chosen algorithms, listed in order of suitability, are as follows:

- **PC:**
 - **Description:** The PC algorithm is a constraint-based method that learns the structure of a causal graph from data by testing conditional independencies between variables. It constructs a directed acyclic graph (DAG) representing the causal relationships.
 - **Justification:** Given that the dataset has a large sample size (1030) and lacks hidden confounders, the PC algorithm is efficient for discovering causal relationships. Its ability to handle continuous data and output a DAG makes it suitable for this analysis.
- **GES:**
 - **Description:** Greedy Equivalence Search (GES) is a score-based causal discovery algorithm that identifies the optimal causal structure by navigating the space of equivalence classes of Directed Acyclic Graphs (DAGs).
 - **Justification:** GES is appropriate for this dataset due to the non-linear relationships and large size. It can operate on non-Gaussian distributions using generalized scores, which aligns with the characteristics of the data provided.
- **NOTEARS:**
 - **Description:** NOTEARS transforms the problem of learning Directed Acyclic Graphs (DAGs) into a continuous optimization problem, allowing for efficient scaling to large datasets.
 - **Justification:** NOTEARS is suitable due to its flexibility in handling high-dimensional data and the possibility of incorporating non-linear relationships. Its ability to directly optimize DAG structures makes it a viable option for this dataset.

4.3 Hyperparameter Values Proposal assisted with LLM

Once the algorithms were selected, the LLM aided in proposing hyperparameters for the chosen algorithm, which are specified below:

- **alpha:**
 - **Value:** 0.05
 - **Explanation:** A significance level of 0.05 is appropriate given the sample size of 1030, which falls in the category of '500-10000.' This balance ensures reasonable sensitivity to detect true causal relationships while minimizing the risk of Type I errors.
- **indep_test:**
 - **Value:** fisherz
 - **Explanation:** Fisher's Z test is suitable for continuous data as indicated in the dataset characteristics. Despite the dataset not following a Gaussian distribution, Fisher's Z is still a commonly used method, as it is efficient and straightforward for analyzing linear correlations.
- **depth:**
 - **Value:** -1
 - **Explanation:** Using -1 for unlimited depth is advisable given the dataset's size and complexity. The dataset consists of 9 features, which does not necessitate depth limitation, thus allowing for a thorough exploration of causal relationships.

4.4 Graph Tuning with Bootstrap and LLM Suggestion

In the final step, we performed graph tuning with suggestions provided by the Bootstrap and LLM.

Firstly, we use the Bootstrap technique to get how much confidence we have on each edge in the initial graph. If the confidence probability of a certain edge is greater than 95% and it is not in the initial graph, we force it. Otherwise, if the confidence probability is smaller than 5% and it exists in the initial graph, we change it to the edge type with the highest probability.

After that, we utilize LLM to help us prune edges and determine the direction of undirected edges according to its knowledge repository. In this step, LLM can use background knowledge to add some edges that are neglected by Statistical Methods. Voting techniques are used to enhance the robustness of results given by LLM, and the results given by LLM should not change results given by Bootstrap.

By integrating insights from both Bootstrap and LLM to refine the causal graph, we can achieve improvements in graph's accuracy and robustness.

5 Results Summary

5.1 Initial Graph

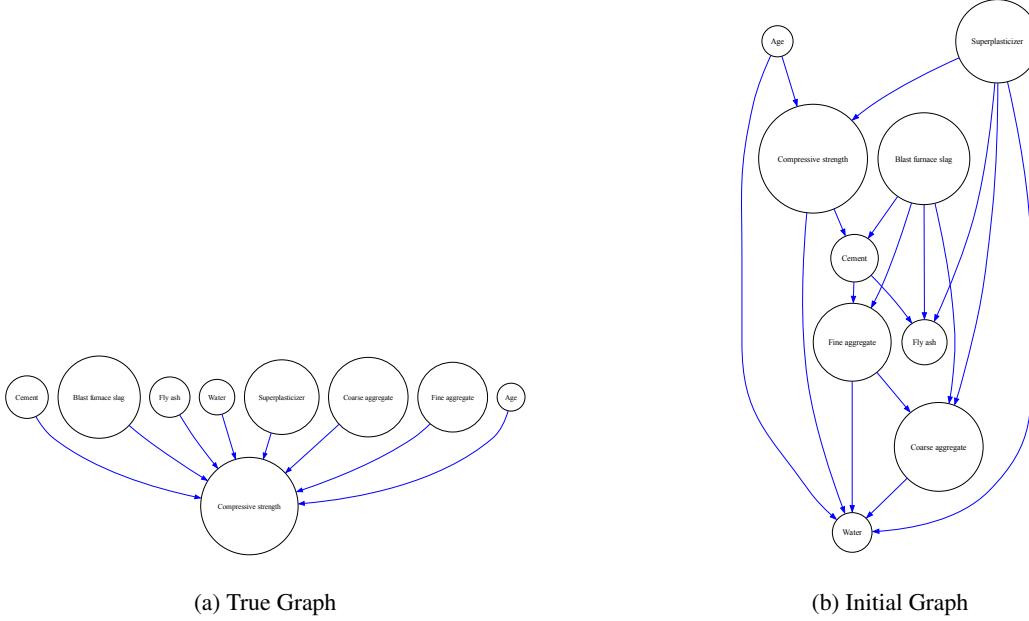


Figure 4: Graphs Comparison of PC

The above is the initial result graph produced by our algorithm.

The analysis reveals a complex interplay within the construction materials used in concrete production, emphasizing the role of various components and their influence on one another. Blast furnace slag is pivotal as it enhances both Cement and multiple aggregate types, including Fly ash, Fine aggregate, and Coarse aggregate, reflecting its contribution to the overall strength and durability of concrete. The Superplasticizer, known for improving workability, directly affects the Water content and the performance metrics of Fly ash, Coarse aggregate, and Compressive strength, indicating its vital role in optimizing mixtures. Moreover, the interaction between aggregates illustrates that Fine aggregate not only influences Coarse aggregate but also has a significant impact on Water, reinforcing the importance of material selection and the timing of hydration processes in achieving the desired concrete properties. Age plays a crucial role as well, affecting both Water and the ultimate Compressive strength, underscoring the evolving nature of concrete as it cures and gains strength over time, while Cement's strength is further influenced by its connection to Compressive strength. This intricate web of causal relationships highlights the critical considerations in material choice and optimization aimed at enhancing concrete performance in construction applications.

5.2 Revised Graph

By using the method mentioned in Section 4.4, we provide a revised graph pruned with Bootstrap and LLM suggestion. Pruning results are as follows.

Bootstrap doesn't force or forbid any edges.

The following are force results given by LLM:

- **Water → Cement:** The amount of water used in the concrete mix directly influences the hydration of cement. A proper water-to-cement ratio is essential, as too much water can weaken the concrete, while too little can prevent proper hydration, affecting the overall strength of the cement.
- **Superplasticizer → Cement:** Superplasticizers improve the workability of concrete mixes by allowing for a lower water-to-cement ratio without sacrificing fluidity. This results in a more efficient hydration process for cement, leading to greater compressive strength.
- **Coarse Aggregate → Cement:** The properties and size distribution of coarse aggregates can significantly affect the overall strength of the concrete mix. Properly graded coarse aggregates work in conjunction with cement to provide a robust structural framework, enhancing the load-bearing capacity of the concrete.
- **Age → Cement:** As concrete cures, it continues to gain strength through ongoing hydration reactions. The age of the concrete is crucial as it reflects the time available for the cement to hydrate fully, thereby influencing its compressive strength.

This structured approach ensures a comprehensive and methodical analysis of the causal relationships within the dataset.

5.3 Graph Reliability Analysis

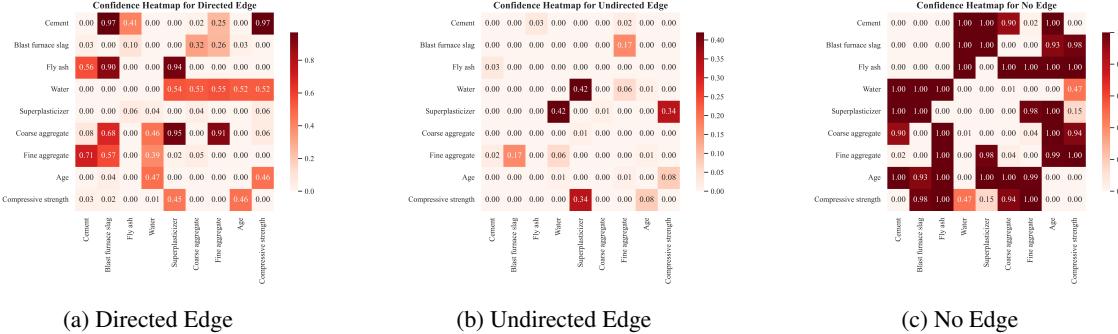


Figure 6: Confidence Heatmap of Different Edges

The above heatmaps show the confidence probability we have on different kinds of edges, including directed edge (\rightarrow), undirected edge ($-$), No Edge, and probability of no edge. The heatmap of bi-edges is not shown because probabilities of all edges are 0. Based on the confidence probability heatmap and background knowledge, we can analyze the reliability of our graph.

From the statistics perspective, we have high confidence to believe that the edges Cement → Fly Ash (0.41), Coarse Aggregate → Water (0.46), Fine Aggregate → Water (0.39), and Age → Water (0.47) exist, as they all exhibit moderate bootstrap probabilities indicating a potential causal influence. Conversely, edges such as Blast Furnace Slag →

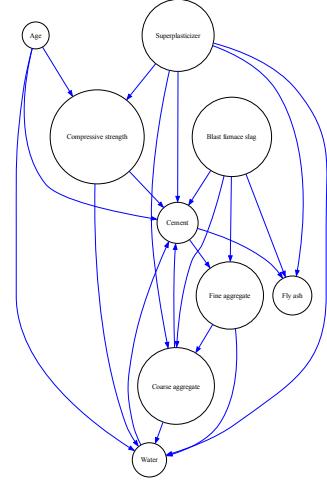


Figure 5: Revised Graph

Cement (0.03), Superplasticizer → Water (0.04), and Compressive Strength → Water (0.01) demonstrate low bootstrap probabilities, implying that there is insufficient confidence to affirm these relationships as causal.

However, based on expert knowledge, we know that the edges Cement → Compressive Strength, Water → Compressive Strength, and Age → Compressive Strength are well-established in the literature and practice concerning concrete mix design, reinforcing the likelihood of their existence. On the other hand, the edges such as Blast Furnace Slag → Cement and Superplasticizer → Coarse Aggregate are known to be less straightforward and are likely weak or indirect relationships, which aligns with their low bootstrap probabilities.

Therefore, the result of this causal graph is partially reliable. While some relationships are strongly supported by both statistical evidence and domain expertise, others require careful consideration and validation before drawing definitive conclusions about their causal nature.