
CAUSAL DISCOVERY REPORT ON EARTHQUAKES

TECHNICAL REPORT



April 8, 2025

ABSTRACT

This report presents a comprehensive causal discovery analysis of a dataset involving factors influencing earthquake impacts, focusing on variables such as richter, area, region, and deaths. We employed the PCMCI algorithm, a constraint-based method suitable for medium-to-large-scale time-series data, and supplemented our methodology with advanced preprocessing and large language model (LLM) assistance to carefully select hyperparameters. Our findings reveal intricate causal relationships: area influences richter, region affects area, and deaths correlate with area size, indicating that broader affected territories can lead to higher casualties. Additionally, we explored the comparative performance of PCMCI and DYNOTEARs, noting a consensus on key causal edges while highlighting differences in the breadth of causal implications. Our contribution underscores the interconnectedness of seismic factors and their implications for disaster response, providing insights that inform risk management strategies and enhance understanding of earthquake dynamics.

Keywords Causal Discovery, Large Language Model, PCMCI, earthquakes

1 Introduction

Causal discovery and inference are critical components in understanding the underlying relationships within complex datasets. This report focuses on a dataset characterized by various factors that potentially interact with one another. Through advanced analytical techniques, we aim to identify causal relationships and gain insights into how these variables influence each other. By systematically exploring the data and applying established methodologies, our objective is to uncover the true causal structure, which can ultimately guide informed decision-making and enhance our understanding of the phenomena represented in the dataset. We will utilize both graphical and statistical methods to derive conclusions and provide recommendations based on our findings.

2 Background Knowledge

2.1 Detailed Explanation about the Variables

The dataset contains several key variables that are essential for understanding earthquake events. The variable **date** indicates when the earthquake occurred and provides temporal context, while **richter** measures the earthquake's magnitude, reflecting the energy released during the event. The **area** and **region** variables provide geographical classifications, where **area** refers to specific localities and **region** encompasses broader geographical units. Finally, **deaths** quantifies the fatalities resulting from the earthquake, highlighting its human impact and relevance for disaster management.

In addition to these variables, background knowledge in seismology and geology is vital for causal discovery, as it encompasses the principles underlying earthquake measurement, tectonic activity, and the relationship between geological features and seismic occurrences. Socioeconomic factors play a significant role as well, influencing the outcomes of earthquakes based on infrastructure and preparedness. Furthermore, understanding temporal variations in

seismic activity and disaster response mechanisms can enhance causal analyses, allowing for a more comprehensive interpretation of the dynamics at play during such events.

2.2 Possible Causal Relations found by LLM

The following are potential causal relationships suggested by the language model, which are visualized in Figure 1. Please note that only variables present in our dataset are included in the figure.

- **richter → deaths:** Higher Richter scale values indicate stronger earthquakes, leading to more structural damage and potentially resulting in increased deaths.
- **area → deaths:** Larger areas affected by earthquakes may indicate a greater population density, increasing the toll of deaths.
- **region → deaths:** Certain regions may have more vulnerable infrastructure, leading to higher fatalities in the event of an earthquake.
- **richter → area:** Stronger earthquakes tend to affect larger areas due to the intensity and reach of seismic waves.
- **area → richter:** Areas that experience frequently high Richter scale events may develop seismic activity patterns that are unique to those regions.
- **region → richter:** Different regions have varying geological compositions that influence the frequency and intensity of Richter scale readings.

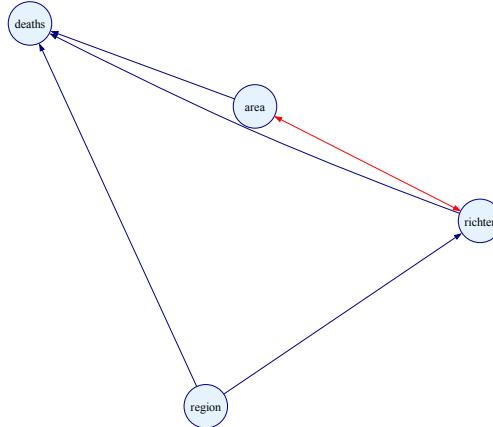


Figure 1: A Causal Graph Suggested by LLM.

3 Dataset Descriptions and EDA

The following provides a preview of our original dataset. If the dataset contains more than 10 columns, a random subset of 10 columns is displayed for illustrative purposes.

Table 1: Dataset Preview.

richter	area	region	deaths
0.513994	0.658148	-0.645307	-0.219540
-1.007743	1.412670	0.959136	-0.181105
-0.177705	-0.016950	1.242273	-0.198187
-1.837782	-0.771472	1.242273	-0.233775
0.513994	-0.414067	-0.550928	0.022461

3.1 Data Properties

We employed several statistical methods to identify data properties, including:

Basic Data Characteristics

The shape of the data, variable types, and the presence of missing values were assessed directly from the DataFrame. In contrast, properties such as time-series structure and heterogeneity were inferred with LLM based on user queries and DataFrame.

Linearity Testing

We conducted the Ramsey's RESET test to assess linearity between each pair of variables. When the total number of possible variable pairs was fewer than 100, all pairs were tested. If the number exceeded 100, a random subset of 100 pairs was selected for testing to ensure computational feasibility. To account for multiple testing, we employed the Benjamini and Yekutieli procedure, which is robust when dealing with dependent or correlated data. The linearity assumption was considered satisfied only if all tested pairs exhibited linearity; otherwise, it was considered violated.

Normality of Residuals

The assumption of Gaussian (normally distributed) noise was assessed using the Shapiro-Wilk test. The testing approach depended on the outcome of the linearity evaluation. If linearity was satisfied, we fitted ordinary least squares (OLS) models for each variable pair and extracted the residuals for testing. If linearity was not satisfied, we used a flexible non-parametric method—locally weighted scatterplot smoothing (LOWESS)—to model the relationships and obtain residuals. The Benjamini and Yekutieli correction was again applied to control for false discovery under multiple testing.

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
(123, 4)	Time-series	True	False	False	True	False

3.2 Correlation Analysis

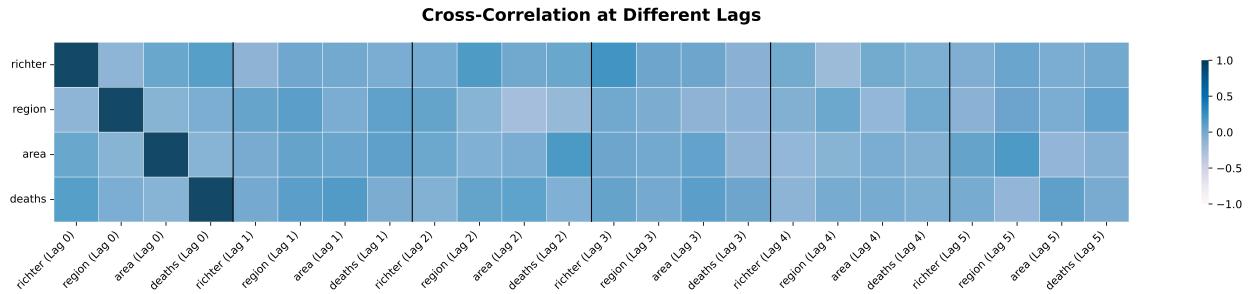


Figure 2: Heatmap of Time-Lagged Correlations Among Variables.

We calculated both time-lagged and instantaneous correlation coefficients between variables. Here we list some variable pairs which have large absolute lagged correlation coefficients.

- region → area at Lag 2: -0.240
- area → deaths at Lag 2: +0.218
- richter → region at Lag 2: +0.192
- region → deaths at Lag 2: -0.159
- area → region at Lag 5: +0.154

3.3 Time Series Stationarity Analysis

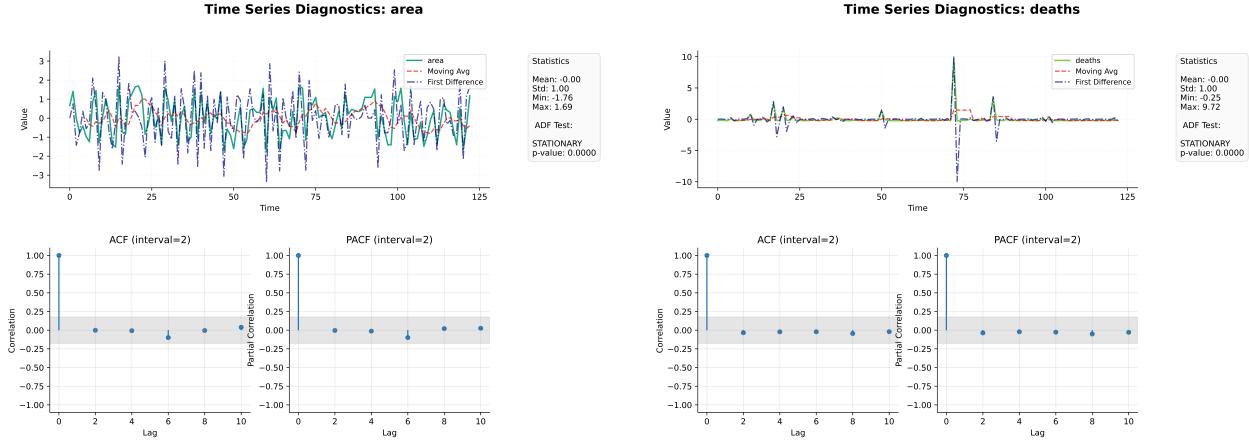


Figure 3: Time Series Diagnostics for two selected variables. Each plot shows the variable with moving average, first difference transform, and ACF/PACF plots.

4 Causal 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 process involved data cleaning, handling missing values, and performing exploratory data analysis to examine variable distributions and inter-variable relationships.

4.2 Algorithm Recommendation assisted with LLM

Following 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:

- **PCMCI:**
 - **Description:** PCMCI (Peter and Clark algorithm with Momentary Conditional Independence) is a constraint-based causal discovery algorithm designed for time-series data. It is flexible in handling both linear and non-linear relationships and can accommodate non-Gaussian noise. PCMCI is efficient and suitable for medium-to-large-scale datasets, providing a DAG output.
 - **Justification:** PCMCI is selected because it is a constraint-based method that is flexible with respect to both functional form and noise, making it suitable for the non-linear and non-Gaussian characteristics of the dataset. It is efficient and can handle medium-to-large-scale datasets, which is appropriate for the user's dataset size. PCMCI's ability to output a DAG aligns with the user's acceptance of CPDAG/PAG formats.
- **NTSNOTEARS:**
 - **Description:** NTSNOTEARS (Non-linear Time Series Nonlinear Optimization with Trace Exponential and Augmented lagRangian for Structure learning) is a continuous-optimization algorithm designed for non-linear time-series data. It is efficient and provides a DAG output, making it suitable for datasets with non-linear relationships and non-Gaussian noise.
 - **Justification:** NTSNOTEARS is selected because it is a continuous-optimization method specifically designed for non-linear time-series data, which matches the non-linear relationships in the dataset. It

is efficient and provides a DAG output, which aligns with the user's acceptance of CPDAG/PAG formats. NTSNOTEARS is also flexible with respect to noise, making it suitable for the non-Gaussian characteristics of the dataset.

Considering data properties, algorithm capability and user's instruction, the final algorithm we choose is PCMCI.

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:

- **Independence Test:**
 - **Value:** parcorr
 - **Explanation:** User specified
- **Minimum Time Lag:**
 - **Value:** 0
 - **Explanation:** Using a minimum lag of 0 ensures that immediate causal relationships are not overlooked, providing a comprehensive analysis.
- **Maximum Time Lag:**
 - **Value:** 10
 - **Explanation:** A maximum lag of 10 balances the need to capture relevant temporal dependencies while maintaining computational efficiency.
- **Significance Level of PC Algorithm:**
 - **Value:** 0.1
 - **Explanation:** A higher significance level increases sensitivity to potential causal relationships, which is beneficial given the smaller sample size.
- **Significance Level for Graph Thresholding:**
 - **Value:** 0.05
 - **Explanation:** This alpha level ensures that the resulting graph is neither too sparse nor too dense, maintaining interpretability and accuracy.

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.

We first applied the Bootstrap method to estimate the confidence level associated with each edge in the initial graph. Specifically:

- If an edge not present in the initial graph exhibited a Bootstrap confidence greater than 90%, we added it to the graph.
- Conversely, if an existing edge had a confidence lower than 10%, we removed it.
- For edges with moderate confidence (between 10% and 90%), we consulted the LLM to assess their validity and directionality, drawing on its extensive background knowledge.

The LLM contributed by:

- Reintroducing plausible edges that may have been overlooked by statistical methods;
- Removing or redirecting edges that appeared statistically valid but were conceptually implausible.

To improve the robustness of LLM-generated suggestions, we employed a voting mechanism. Importantly, LLM recommendations were not allowed to override high-confidence decisions made by the Bootstrap procedure. By integrating insights from both of Bootsrap and LLM to refine the causal graph, we can achieve improvements in graph's accuracy and robustness.

5 Causal Graph Estimation Summary

5.1 Causal Graph Discovered by the Algorithm

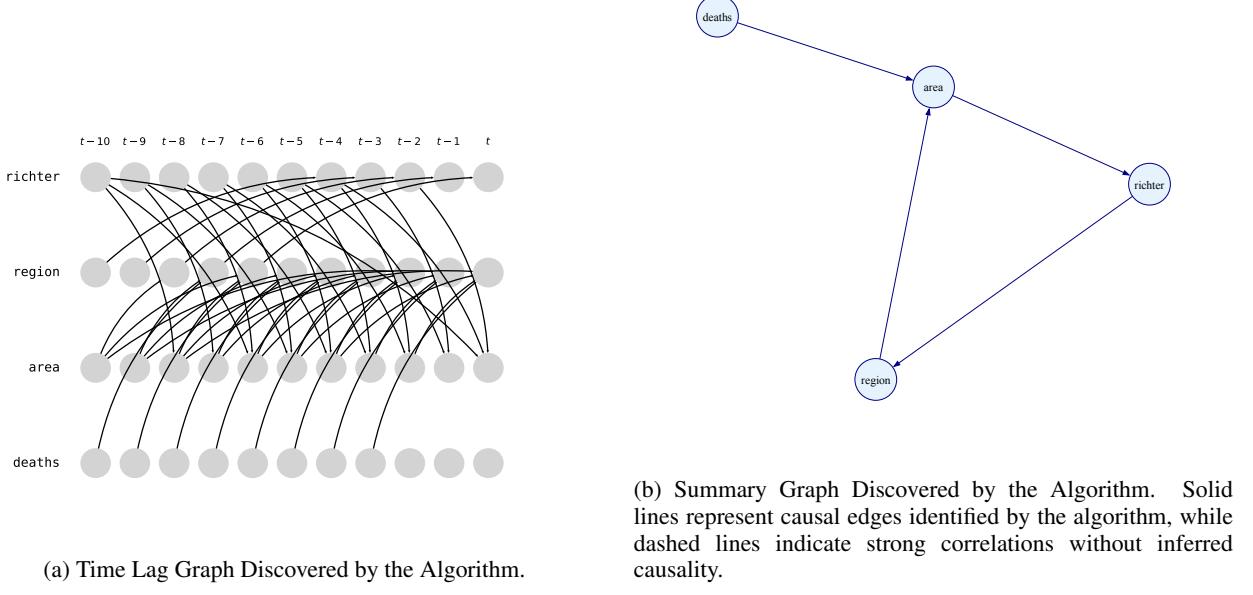


Figure 4: Graphs Discovered by the Algorithm.

The analysis of the causal relationships among the variables richter, area, region, and deaths reveals a complex interplay that influences how seismic events affect populations. The variable area, which represents the geographical extent impacted by an earthquake, has a direct causal influence on richter, indicating that the scale or magnitude of seismic activity is affected by the area of the event. Additionally, area is affected by region, suggesting that the characteristics of a specific geographical region can impact the extent of land affected by an earthquake. The variable deaths is in turn influenced by area, indicating that larger affected areas can correlate with higher casualties during seismic events. Interestingly, richter affects region, which implies that the magnitude of the earthquake can also shape the understanding and categorization of the affected geographical region.

- The area influences the richter, suggesting that larger earthquakes may arise from more extensive affected areas.
- Region affects area, indicating that geographical characteristics can determine the extent of an earthquake's impact.
- Deaths are caused by area, showing that larger affected regions tend to have higher casualties.
- Richter influences region, meaning the magnitude of the earthquake can impact how we define and understand affected areas.

In conclusion, the relationships among these variables highlight how the magnitude of earthquakes, the area affected, and the regional context are interlinked, ultimately affecting casualty outcomes. Understanding these connections is crucial for improving disaster response and risk management strategies.

5.2 Causal Graph after Revision with Bootstrap and LLM

5.2.1 Bootstrap Probability

To evaluate the confidence associated with each edge in the causal graph, we employed a bootstrapping procedure to estimate the probability of existence for each edge. From a statistical perspective, we categorize these probabilities into three levels:

- **High Confidence Edges:** None
- **Moderate Confidence Edges:** area causes richter, region causes area, deaths causes area, richter causes region
- **Low Confidence Edges:** None

5.2.2 LLM Pruning

By using the method mentioned in the 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 relationships are forbidden by LLM:

- **richter → area:** The magnitude of an earthquake, represented by the Richter scale, does not influence or alter the geographical area where the earthquake occurs;

The following are directions added by the LLM:

- **richter → region:** The Richter scale measures earthquake magnitude, and higher magnitudes can have a significant impact on specific regions, causing destructive effects that can directly alter the level of damage and response needed in those areas.
- **area → region:** The geographical area significantly influences the characteristics of a region, including its socio-economic status, natural resources, and population density, which can subsequently affect various aspects of development;
- **deaths → area:** The causal relationship indicates that certain areas may have higher mortality rates due to factors such as environmental hazards or lack of access to healthcare, thereby influencing the number of deaths reported in those regions;
- **region → deaths:** The region can significantly influence the number of deaths due to varying factors such as population density, healthcare access, and environmental conditions, which directly affect mortality rates;

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

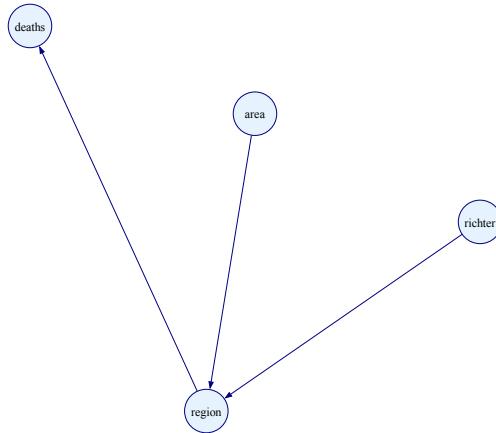


Figure 5: Revised Graph by LLM.

5.3 Graph Reliability Analysis

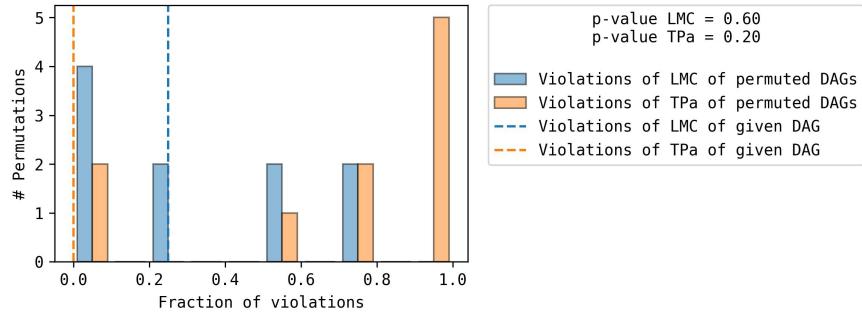


Figure 6: Refutation Graph.

The results of the graph refutation test indicate that the analyzed Directed Acyclic Graph (DAG) fails to provide significant inferential power, as evidenced by the low proportion of permutations (2 out of 10) that reside within its Markov equivalence class, yielding a p-value of 0.20. Furthermore, the DAG violates 1 out of 4 Local Markov Conditions (LMCs), but it still outperforms 40% of the permuted graphs, with a corresponding p-value of 0.60. Given the predetermined significance level of 0.05, the lack of informative power and the failure to meet the criteria for rejection lead us to retain the DAG despite its limitations. Overall, the evidence suggests that the graph does not adequately represent the underlying causal structure, prompting the need for further investigation and potential model refinement.

5.4 Result Graph Comparision

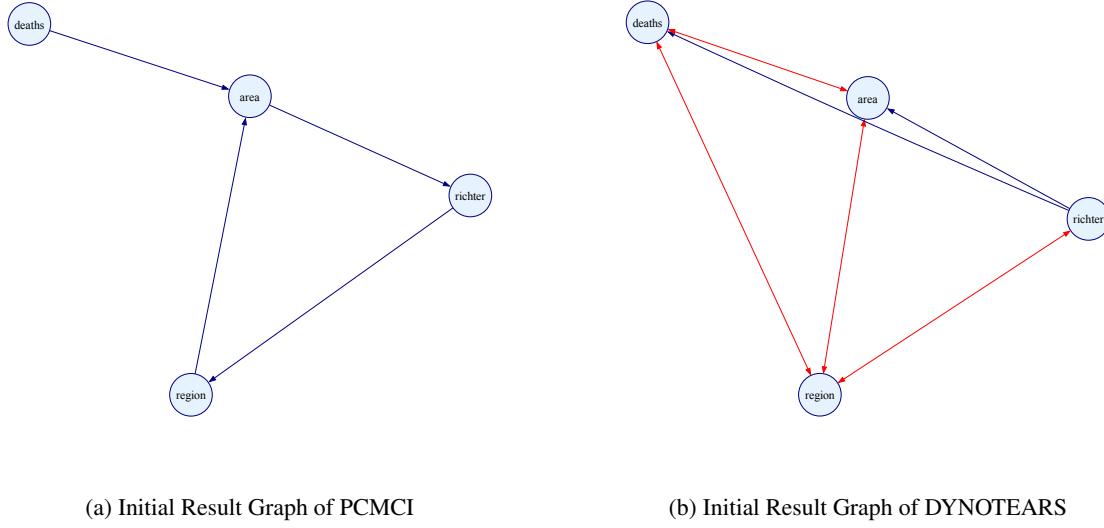


Figure 7: Result Graph Comparision of Different Algorithms.

When comparing the causal graphs generated by the PCMCI and DYNOTEARNS algorithms, several differences and similarities can be observed in the directed edges they propose.

The PCMCI algorithm identifies four directed edges: area causes richter, region causes area, deaths causes area, and richter causes region. In contrast, DYNOTEARNS presents a more extensive set of causal relationships, including region causes richter, richter causes area, region causes area, deaths causes area, richter causes region, area causes region, deaths causes region, richter causes deaths, area causes deaths, and region causes deaths.

Among the edges proposed by both algorithms, the common edges include region causes area, deaths causes area, and richter causes region. This indicates a consensus between the two algorithms on the causal influence of these particular relationships, suggesting a stronger reliability for these causal assertions.

However, considering the number of edges identified by DYNOTEARNS compared to PCMCI, the edges suggested by DYNOTEARNS may provide a more comprehensive understanding of the causal relationships present in the data. The additional edges identified by DYNOTEARNS could be seen as more informative, but the reliability of these edges would depend on the underlying assumptions and the data structure. Common edges identified by both algorithms can be regarded as more reliable due to their agreement, while edges unique to DYNOTEARNS warrant further scrutiny to validate their causal claims.

5.5 Conclusion

In this report, we analyzed a dataset related to earthquakes, exploring variables such as magnitude (richter), geographical area, region, and casualty figures (deaths) to uncover their causal relationships. Utilizing a comprehensive methodology that included data preprocessing, algorithm selection through a large language model (LLM), and robustness assessment via graph tuning with Bootstrap and LLM suggestions, we applied the PCMCI and DYNOTEARNS algorithms to derive insights from the dataset.

Our findings revealed critical causal connections, with the area impacting the magnitude of earthquakes, the region affecting the extent of impact, and a direct correlation between area and casualty figures. Notably, these insights indicate the intertwined nature of the variables, emphasizing that understanding the causal structures can significantly enhance disaster response and risk management strategies. Our contributions lie in the systematic application of advanced causal discovery methodologies to a complex dataset, showcasing the importance of integrating statistical techniques and domain knowledge for improved understanding of natural disaster dynamics.