
CAUSAL DISCOVERY REPORT ON DAILYDELHICLIMATE

TECHNICAL REPORT



April 8, 2025

ABSTRACT

This study investigates causal relationships within the Dailydelhiclimate dataset, which encompasses various meteorological variables. Utilizing the PCMCI (Peter and Clark algorithm with Momentary Conditional Independence), we conducted a detailed causal discovery process supported by preprocessing, large language model-assisted algorithm selection, and hyperparameter tuning. Our results reveal intricate interdependencies, indicating bidirectional influences between humidity and mean temperature, as well as the role of wind speed in affecting both mean temperature and humidity levels. The analysis highlights a feedback loop among these variables, emphasizing their interconnected nature in weather dynamics. Our contribution lies in elucidating these relationships through a robust methodological framework, merging statistical rigor with advanced computational techniques to enhance future research in causal inference and meteorology.

Keywords Causal Discovery, Large Language Model, PCMCI, DailyDelhiClimate

1 Introduction

Causal discovery and inference play a critical role in understanding the complex interplay between variables in various fields, from healthcare to social sciences. The dataset at hand presents a unique opportunity to explore these relationships, as it contains rich information that may reveal underlying causal structures. Through the application of appropriate statistical techniques and algorithms, we aim to identify potential causal pathways and assess the strength of these associations. By leveraging both the data at our disposal and the theoretical foundations of causal inference, this analysis seeks to contribute valuable insights that can inform future research and practical applications.

2 Background Knowledge

2.1 Detailed Explanation about the Variables

The dataset includes several key variables related to climate, such as **date**, **meantemp** (mean temperature), **humidity**, **wind speed**, and **meanpressure**. The **date** variable is vital for conducting time-series analyses, allowing researchers to track daily changes in climate. **Meantemp** indicates the average temperature for a given day, significantly impacting weather-related phenomena. **Humidity** measures the moisture in the air and affects comfort levels and precipitation, while **wind speed** is a measure of wind velocity influencing temperature distribution. Lastly, **meanpressure** encapsulates the average atmospheric pressure, which is crucial in forecasting weather patterns.

Understanding background domain knowledge is essential for effective causal discovery. The temporal nature of the data suggests that past weather conditions can influence future observations, necessitating the assessment of time dependencies. Seasonal variations must be accounted for, as climate relationships may differ significantly between summer and winter. Moreover, external influences such as geographical features and broader climate phenomena should be integrated into causal models. Additionally, recognizing the potential confounding factors and distinguishing causality from mere correlation are critical for establishing robust causal relationships in the dataset, guiding researchers in their analytical strategies.

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.

- **humidity → wind speed:** Changes in humidity may affect local weather patterns, which could lead to variations in wind speed as air masses move.

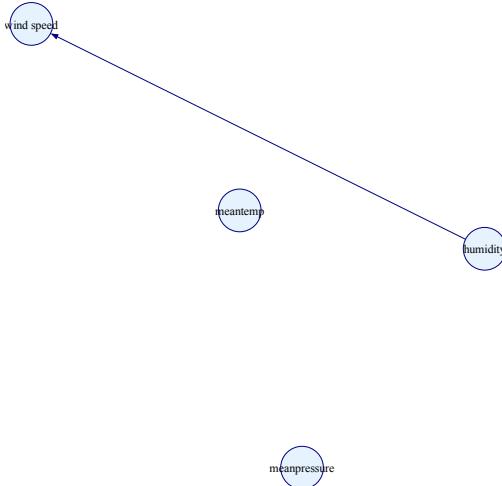


Figure 1: A Causal Graph Suggested by LLM.

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.

meantemp	humidity	wind speed	meanpressure
10.000000	84.500000	0.000000	1015.666667
7.400000	92.000000	2.980000	1017.800000
7.166667	87.000000	4.633333	1018.666667
8.666667	71.333333	1.233333	1017.166667
6.000000	86.833333	3.700000	1016.500000

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
(1462, 4)	Time-series	False	True	False	True	False

3.2 Correlation Analysis

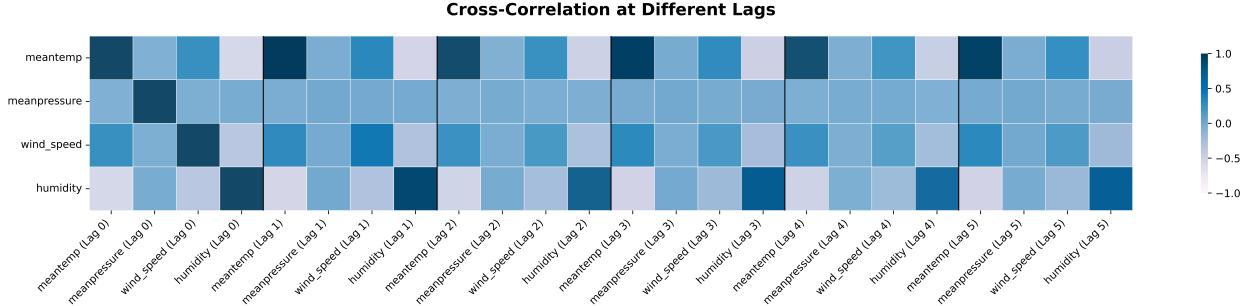


Figure 2: Heatmap of Time-Lagged Correlations Among Variables.

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

- humidity → meantemp at Lag 1: -0.542
- meantemp → humidity at Lag 1: -0.526
- wind speed → humidity at Lag 1: -0.319
- humidity → wind speed at Lag 1: -0.311
- meantemp → wind speed at Lag 1: +0.308

3.3 Time Series Stationarity Analysis

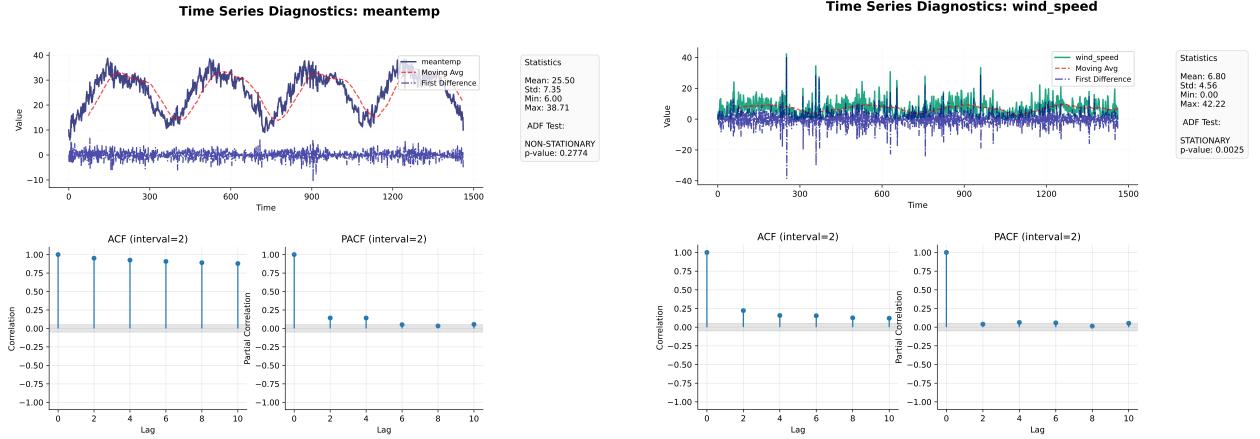


Figure 3: Time Series Diagnostics for two selected variables. Each plot shows the variable with moving average, first difference transform, and ACF/PACF plots.

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 algorithm designed for time-series data. It is flexible in handling both linear and non-linear relationships and can work with both Gaussian and non-Gaussian noise.
- **Justification:** PCMCI is a constraint-based algorithm that is well-suited for time-series data. It is flexible in terms of functional form and noise, making it a good fit for the linear relationships and non-Gaussian errors in the dataset. PCMCI is efficient and can handle medium-to-large-scale datasets, which aligns with the sample size of 1462. It outputs a DAG, which is compatible with the user's acceptance of undirected edges. PCMCI's strong empirical performance and preference for sparse graphs make it a reliable choice for this dataset.

- **VARLiNGAM:**

- **Description:** VARLiNGAM (Vector Autoregressive Linear Non-Gaussian Acyclic Model) is a functional model-based algorithm optimized for linear relationships and non-Gaussian noise. It is designed for time-series data and provides efficient performance on CPU.
- **Justification:** VARLiNGAM is a functional model-based algorithm that is specifically designed for linear relationships and non-Gaussian noise, which matches the dataset's characteristics. It is efficient on CPU and can handle medium-to-large-scale datasets. VARLiNGAM outputs a DAG and is density-robust, making it suitable for the dataset's potential graph density. Its strong empirical performance and ability to handle time-series data make it a suitable choice for this 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:** Parcorr is efficient for linear relationships and continuous data, providing a good balance between accuracy and computational efficiency.
- **Minimum Time Lag:**
 - **Value:** 0
 - **Explanation:** Starting with a minimum lag of 0 allows the algorithm to explore immediate relationships, which is standard unless specific immediate effects are known.
- **Maximum Time Lag:**
 - **Value:** 10
 - **Explanation:** A maximum lag of 10 captures a reasonable range of temporal dependencies while maintaining computational efficiency.
- **Significance Level of PC Algorithm:**
 - **Value:** 0.05
 - **Explanation:** A significance level of 0.05 is appropriate for the sample size, balancing the risk of false positives and negatives.
- **Significance Level for Graph Thresholding:**
 - **Value:** 0.05
 - **Explanation:** An alpha level of 0.05 balances graph sparsity and accuracy, suitable for moderate graph sizes.

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

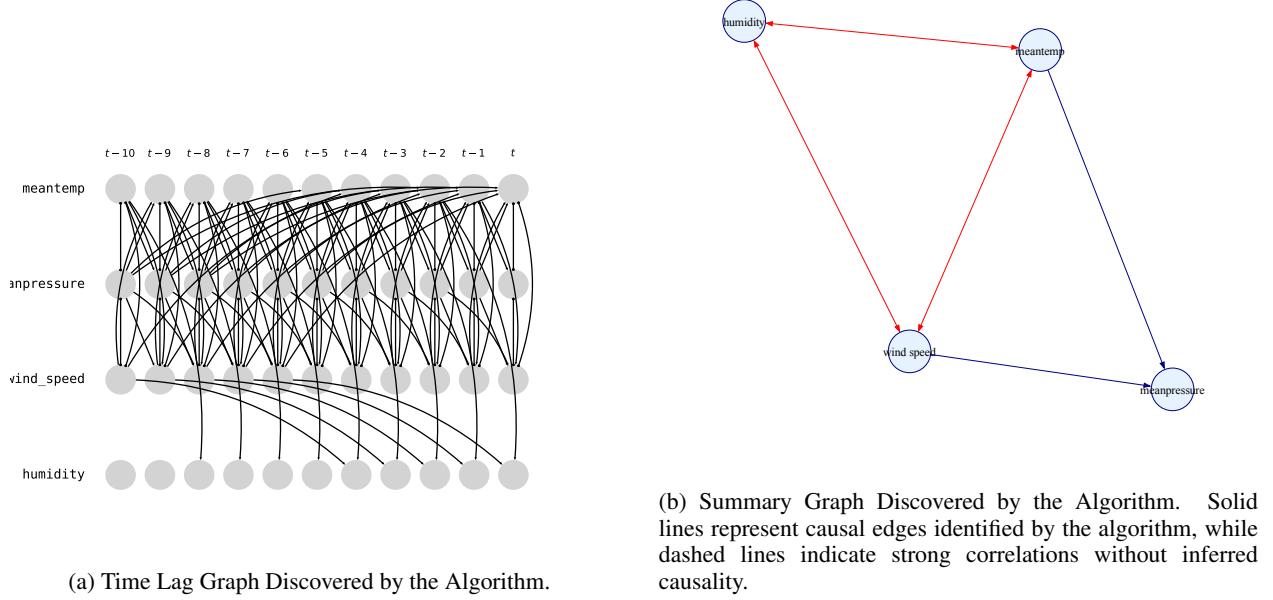


Figure 4: Graphs Discovered by the Algorithm.

The analysis reveals complex interdependencies among the variables involved in weather dynamics. Specifically, humidity and meantemp interact bidirectionally, highlighting that higher humidity can elevate the mean temperature, while increased mean temperature can also lead to higher humidity levels due to the capacity of warmer air to hold more moisture. Wind speed affects both meantemp and humidity by dispersing or promoting thermal mixing, thereby altering temperature and humidity readings. Additionally, wind speed influences mean pressure, suggesting that fluctuations in wind patterns can impact overall atmospheric pressure. The relationships among these variables portray a feedback loop where changes in any one variable can propagate through the others, emphasizing their interconnected nature in weather systems.

- Humidity influences meantemp and vice versa, indicating a reciprocal relationship.
- Wind speed causes changes in both meantemp and humidity, introducing an external factor that modifies temperature and moisture levels.
- Mean temperature is implicated in altering humidity and wind speed, showing its central role in weather dynamics.
- Wind speed also affects mean pressure, indicating the role of air movement in atmospheric pressure variation.
- The relationships demonstrate a feedback mechanism, with each variable affecting and being affected by the others, suggesting a dynamic and interlinked system.

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:** Meantemp causes meanpressure, Wind speed causes meanpressure
- **Low Confidence Edges:** Humidity causes meantemp, Wind speed causes meantemp, Meantemp causes humidity, Wind speed causes humidity, Meantemp causes wind speed, Humidity causes wind speed

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.

LLM doesn't forbid any edges.

The following are directions added by the LLM:

- **meantemp → meanpressure:** Higher temperatures can correspond with lower atmospheric pressure, as warmer air is less dense and tends to rise, thus causing a drop in mean pressure associated with changing weather patterns;
- **humidity → meanpressure:** Increased humidity levels can lead to a decrease in mean pressure due to the condensation of water vapor in the atmosphere, which contributes to the formation of clouds and precipitation;

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

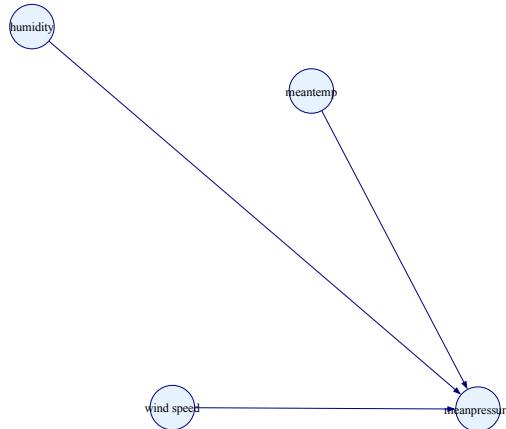


Figure 5: Revised Graph by LLM.

5.3 Graph Reliability Analysis

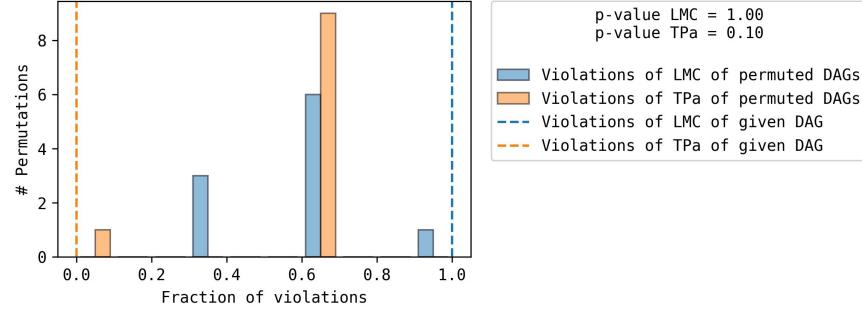


Figure 6: Refutation Graph.

The results from the graph refutation test indicate that the given Directed Acyclic Graph (DAG) exhibits significant limitations in its informativeness and reliability. With only 1 out of 10 permutations falling within the Markov equivalence class of the DAG (p-value: 0.10), it suggests that the model fails to adequately represent the underlying causal relationships. Moreover, the DAG violates all six Local Markov Conditions (LMCs), indicating that the model poorly reflects the data's structure, with a p-value of 1.00 showing that it is worse than 100% of the permuted DAGs. Given these findings and the predefined significance level of 0.05, we conclude that the DAG is not informative and therefore, we do not reject it, acknowledging its limitations in accurately capturing the causal dynamics of the system being studied.

5.4 Conclusion

In this causal discovery report, we analyzed the Dailydelhicleimate dataset to explore the intricate relationships among weather variables, particularly focusing on humidity, mean temperature, wind speed, and mean pressure. Utilizing a combination of sophisticated algorithms, including PCMCi and VARLiNGAM, we meticulously processed the data, ensuring its integrity and suitability for analysis while implementing statistical techniques to identify potential causal pathways.

The results of our analysis uncovered dynamic interdependencies and feedback loops among the variables, highlighting the reciprocal influences of humidity and mean temperature, as well as the role of wind speed in impacting temperature and atmospheric pressure. Our contributions lie in advancing the understanding of these weather systems through the revision of causal graphs, enhanced by a novel integration of bootstrap confidence estimates and suggestions from a large language model (LLM). This dual approach not only refines the graph's accuracy but also enriches the insights into how these variables interact, providing a significant foundation for future research and applications in environmental monitoring and forecasting.