Data-Driven Probabilistic Causal Inference for Occupant Behavior Modeling

Jinyoung KO, Seungjae LEE*

Department of Civil and Mineral Engineering, University of Toronto, Toronto, ON M5S 1A4, Canada

* Corresponding Author

ABSTRACT

Occupant behaviors and decision-making have significant impacts on indoor environmental quality, energy consumption, and greenhouse gas emissions in buildings. Taking account of occupant behavior in building solutions using data-driven methods is important to reduce carbon emissions and simultaneously satisfy occupants' needs in buildings. Especially, identifying potential causal factors of occupant decision-making is especially imperative to (i) properly intervene in occupant behavior and (ii) improve robustness of building energy solutions. However, investigating underlying causal mechanisms of occupant behavior is challenging. This is because of (i) the difficulty in conducting controlled experiments with real occupants and (ii) the limitation of conventional statistical methods to investigate causality under observational environments. To address such difficulties, this study proposes a probabilistic causal discovery approach based on a Bayesian model comparison and a Monte Carlo method. With the open dataset provided by a thermostat company, the proposed causal discovery method identified four potential causal variables of the household's thermostat decision-making. With the inferred causal knowledge, two models, (i) a causal model including direct causal variables and (ii) a non-causal model involving all available variables, were developed. The prediction performances of the two models were evaluated with two test datasets with and without data shift, where the joint variable distribution of the test dataset is not identical to that of the training dataset. The prediction performances of the two models were similar over the test data without data shift. On the other hand, over the test dataset under data shift, the mean absolute errors of the causal model and the non-causal model were 2.26 ℃ and 3.44 ℃. This showed more robust predictions from the causal model compared with those from the non-causal model, under the data shift.

1. INTRODUCTION

Since energy consumption from the building sector is a source of more than 30 % of global greenhouse gas emissions (IEA, 2022), reducing the carbon footprint of buildings is inevitable to mitigate global climate change. Meanwhile, building systems also need to provide comfortable environments for occupants. Previous literature has presented that occupant behaviors and decisions significantly affect building energy consumption, indoor environments, and greenhouse gas emissions (Khorasani Zadeh et al., 2023; Sarran et al., 2021). The findings have emphasized the importance of considering those as well as occupant comfort and health in building operations to achieve the two goals effectively (Becerik-Gerber et al., 2022; Nagy et al., 2023). There have been various research efforts to investigate occupant behaviors and decision-making processes and integrate those into building solutions using data-driven modeling and machine learning methods (Dong et al., 2022; Kim et al., 2018).

However, some studies started stressing the significance of understanding and considering causal relationships, not merely correlations, between variables associated with occupant behavior and decision-making (Sahoh et al., 2022). They argued that causal knowledge is crucial to properly intervene in or control the direct or indirect causal factors of occupant behavior to maximize building performance and occupant satisfaction (Lee & Karava, 2020). In addition, they discussed potential failures (i.e., limited reliability) of building solutions based solely on conventional data-driven and machine learning methods. This is because the associations discovered by conventional data-driven and machine learning methods from observational data do not guarantee causation (Pearl & Mackenzie, 2018). If a data-driven behavior model is directly incorporated into a controller, the controller may make wrong system-side decisions due to hidden confounders and reverse causation. Moreover, if there exists data shift, difference in the distribution of the

training data and the data encountered during deployment, conventional machine learning models show lower prediction performance (Mehdi Ataei et al., 2021). Incorporating a causal structure in a data-driven model can significantly improve prediction robustness under such data shift, which is common in buildings. In this regard, it is important to identify the underlying causal mechanism of occupant behavior and use the inferred knowledge for occupant behavior modeling and solution development.

Identifying causal mechanisms behind occupant behavior and decision-making in buildings is challenging. This is because (i) it is difficult to conduct controlled experiments in real-world occupants and buildings, and (ii) conventional methods have limited ability to disentangle causal effects from confounding and reverse causal effects. To overcome the above difficulties in real-world environments, causal discovery methods have been developed to infer potential causal structures between variables over a few decades (Heckerman, 1995; Spirtes & Meek, 1995). Recently, there have been a few studies to elaborate causal inference methods for building design and operation. Ko and Lee (Ko & Lee, 2024) have developed a Bayesian causal discovery approach and demonstrated the performance of the method over the synthetic dataset based on the Peter-Clark (PC) algorithm, which is one of the constraint-based causal discovery algorithms. However, the PC algorithm itself is highly order-dependent as it is a greedy search algorithm. The algorithm may not provide a reliable causal structure with a single trial. Also, the causal discovery method needs to be demonstrated over the real-world dataset.

To address the above limitations of the precedented study, this study (i) proposes a more reliable causal discovery approach with the Monte Carlo method and (ii) demonstrates the potential robustness of a causal model developed with the causal discovery result. This paper will first introduce the causal discovery procedure to discover potential causal structures based on the Bayesian model comparison and the Monte Carlo method. The proposed causal discovery method will be implemented over the preprocessed the open dataset from a thermostat company. Consequently, the robustness of the causal model will be demonstrated compared to that of the non-causal model, under the data shift.

2. METHODOLOGY

2.1 Probabilistic Causal Discovery Method

To investigate potential causal relationships between variables in data, various data-driven causal discovery methods have been developed (Glymour et al., 2019). The causal discovery methods aim to discover directed acyclic graphs (DAGs) that represent the causal structure between variables. A constrain-based algorithm, a type of causal discovery algorithm, finds a potential causal graph based on a set of deterministic conditional independence tests. One limitation of the constraint-based algorithm is its inability to quantify the uncertainty associated with the inferred causal structures. To address this issue, Ko and Lee (Ko & Lee, 2024) introduced a probabilistic approach to quantifying the probability of adjacency (which implies the probability of a direct association) between variables based on the concept of Bayesian model comparison. However, the inference of the causal graph was still deterministic and heavily affected by the order between internal processes, which limited the reliability of the causal discovery result (Le et al., 2019). To address these limitations, we propose an improved probabilistic causal discovery method, by applying the concept of Bayesian model comparison and Monte Carlo sampling.

The proposed causal discovery method follows the process of the PC algorithm, a well-known constraint-based causal discovery algorithm. The PC algorithm uses the conditional independence test results between variables to infer a causal graph. The following examples demonstrate the causal discovery process of the PC algorithm. Let's assume the independence test results (Equations 1 and 2) and the true causal structure (Figure 1 (a)) with four variables. First, the PC algorithm begins with the undirected graph with all edges (Figure 1 (b)). Second, as A and B are independent, an adjacent edge between A and B is removed (Figure 1 (c)). Third, adjacent edges between (i) A and D and (ii) B and D are removed as sets of variables are independent given C (Figure 1 (d)). Fourth, as A and B become dependent given C while they are unconditionally independent, the edge directions can be determined from A and B to C (Figure 1 (e)). Finally, as A and D are not dependent given C although there is no adjacent edge between A and D, the direction can be determined to the reverse direction of the edges in the prior step (Figure 1 (f)).

$$H_{A \perp\!\!\!\perp B} = H_{A \perp\!\!\!\perp D|C} = H_{B \perp\!\!\!\perp D|C} = 1, \quad (1)$$

$$H_{A \perp\!\!\!\perp C} = H_{A \perp\!\!\!\perp D} = H_{B \perp\!\!\!\perp C} = H_{B \perp\!\!\!\perp D} = H_{A \perp\!\!\!\perp B|C} = 0, \quad (2)$$

where $H_{x \perp\!\!\!\perp y} = 1$ represents that variable x and y are independent,

$H_{x \perp\!\!\!\perp y} = 0$ represents that variable x and y are dependent,
$H_{x \perp\!\!\!\perp y|c} = 1$ represents that variable x and y are independent given c,
$H_{x \perp\!\!\!\perp y|c} = 0$ represents that variable x and y are dependent given c.

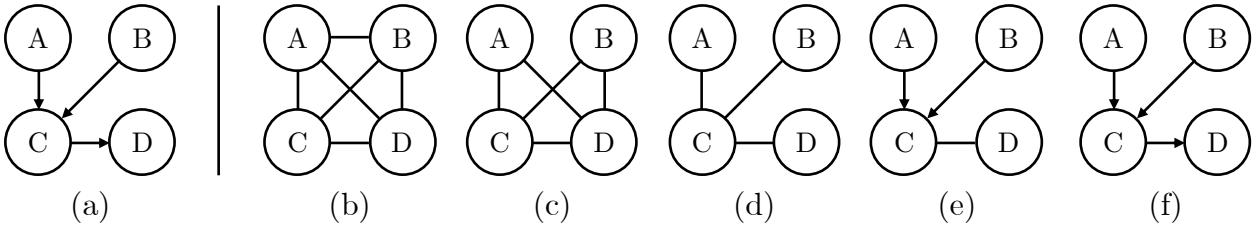

Figure 1: Causal discovery procedure of the PC algorithm.

However, the PC algorithm relies on deterministic independence test results, which cannot consider the uncertainty relevant to the discovered causal structure. To tackle such limitations, we introduce a Bayesian way to estimate the probability of adjacency (corresponding to the probability of dependence) between variables. This allows us to consider the uncertainty behind inferred causal structures. The proposed method first estimates the probability of adjacency between each pair of variables (Ko & Lee, 2024). The unconditional probability of adjacency can be estimated with two graphical models (a) and (b). The graphical representations of model (a) and (b) are described in Figure 2 (a) and (b). Here, the model (a) represents that variable x and y are adjacent, and the probability of y is modelled as:

$$p(\mathrm{y}|\mathbf{x}, \boldsymbol{\beta}_{\mathrm{x}}, \tau) \coloneqq \mathcal{N}(\mathrm{y}|\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}_{\mathrm{x}}, \tau^{-1}),$$

where $\mathcal{N}(\cdot\,|\mu, \tau^{-1})$ is the probability density function of a Gaussian distribution having μ and τ as the mean and precision. The model (b) represents that variable x and y are not adjacent, and the probability of y is modelled as:

$$p(y) \coloneqq \mathcal{N}(y|\beta, \tau^{-1}).$$

After training these models with data, we can quantitatively evaluate which model generalizes data distributions better by comparing the approximates of model evidence (Vehtari & Lampinen, 2002). This can be quantified as model weights based on the stacking of predictive distributions (Yao et al., 2018). Then, the weight for the model (a) refers to the probability of adjacency between x and y without conditioning any variable (i.e., $1 - p_{x \perp\!\!\!\perp y}$). Furthermore, the conditional probability of adjacency also can be estimated by comparing two models: models (c) and (d) (Figures 2 (c) and (d)). The model (c) represents that x and y are adjacent given c. The probability of y is modelled as:

$$p(y|\mathbf{x}, \mathbf{c}, \boldsymbol{\beta}_{\mathrm{x}}, \boldsymbol{\beta}_{\mathbf{c}}, \tau) \coloneqq \mathcal{N}(y|\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}_{\mathrm{x}} + \mathbf{c}^{\mathbf{T}}\boldsymbol{\beta}_{\boldsymbol{c}}, \tau^{-1}).$$

The model (d) refers to that variable x and y are d-separated (i.e., there is no direct edge) given c. The probability of y is modelled as:

$$p(y|\mathbf{c}, \boldsymbol{\beta}_{\mathrm{c}}, \tau) \coloneqq \mathcal{N}(\mathrm{y}|\mathbf{c}^{\mathbf{T}}\boldsymbol{\beta}_{\mathrm{c}}, \tau^{-1}).$$

The model weight of the model (c), which is estimated by model comparison, represents the conditional probability of adjacency between x and y given c (i.e., $1 - p_{x \perp\!\!\!\perp y|c}$). In order to develop graphical models with continuous variables, relationships between variable x, y, and c need to be assumed. We assumed the relationship between each variable is assumed to be a third-order polynomial (n=3). Thus, linear, quadratic, and cubic terms of each variable element are implied in $\mathbf{x}$ and $\mathbf{c}$, where $\mathbf{x} = \{1, x, x^2, \ldots, x^{\mathrm{n}}\}, \mathbf{c} = \{1, c, c^2, \ldots, c^{\mathrm{n}}\}, \boldsymbol{\beta}_{\mathrm{x}} = \{\beta_{\mathrm{x},0}, \beta_{\mathrm{x},1}, \beta_{\mathrm{x},2}, \ldots, \beta_{\mathrm{x,n}}\}$, and $\boldsymbol{\beta}_{\mathrm{x}} = \{\beta_{\mathrm{c},0}, \beta_{\mathrm{c},1}, \beta_{\mathrm{c},2}, \ldots, \beta_{\mathrm{c,n}}\}$.

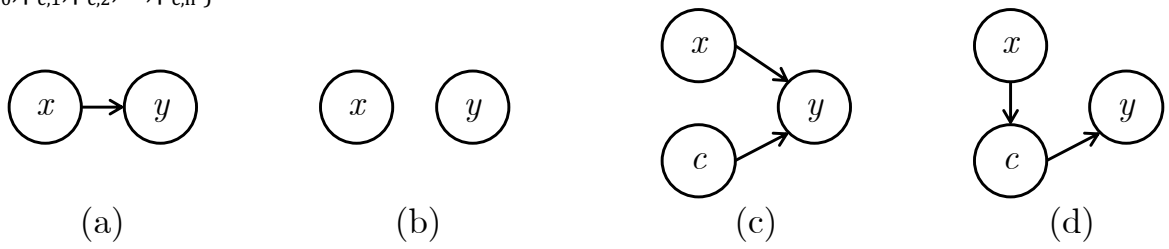

Figure 2: Graphical models.

Based on the probabilities of adjacency between a set of variables, we leveraged the concept of Monte Carlo sampling to infer potential causal structures (Figure 3). Monte Carlo sampling involves repeated iterations to approximate a

target result with samples. In this study, the adjacencies between a set of variables ($H_{x \perp\!\!\!\perp y}$ and $H_{x \perp\!\!\!\perp y|c}$) are determined based on random sampling with the probabilities of adjacency ($p_{x \perp\!\!\!\perp y}$ and $p_{x \perp\!\!\!\perp y|c}$) to infer a causal graph in each iteration. Orders in the causal discovery process are also randomly determined to address the order dependency problem of the PC algorithm. Consequently, in each iteration, a potential causal graph is derived with the idea of the conventional PC algorithm (Spirtes et al., 1993). By having a number of causal graph samples, we can derive the frequencies of directed/undirected edges between each set of variables. The frequencies allow us to identify potential causal variables that have a direct causal relationship to a target variable.

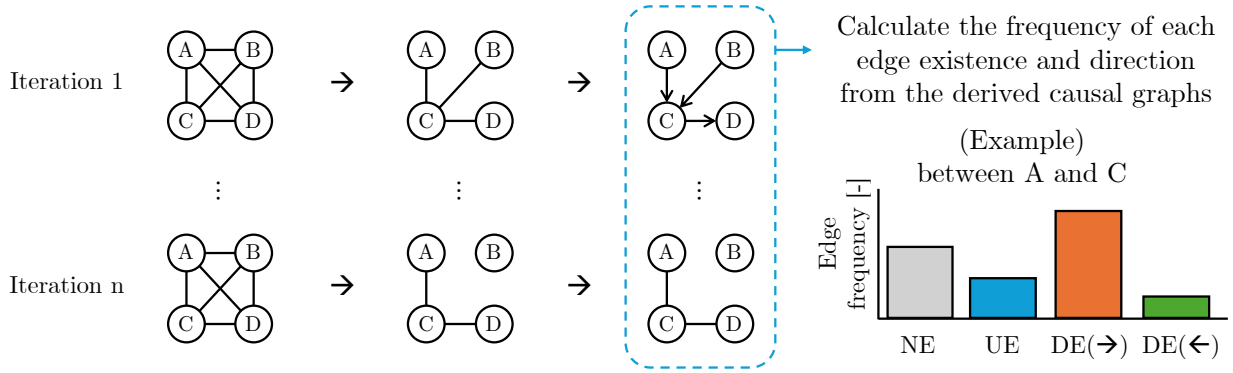

Figure 3: Monte Carlo simulation-based causal discovery method
(NE: no edges, UE: undirected edges, DE(→): directed edges between variables, DE(←): reverse-directed edges between variables).

2.2 Problem Formulation, Data Preparation, and Model Evaluation

We demonstrate the effectiveness of the developed causal discovery method by answering the following research question: "If one changes the cooling setpoint temperature in their house, what are the causal factors affecting the degree of change?"

To this end, a single household, located in San Leandro, in the open dataset provided by a thermostat company is selected. A dataset is prepared by combining data for the household in the dataset and weather data from Oakland, adjacent to San Leandro (Visual Crossing Corporation, 2024). The dataset includes indoor temperature (T_i), indoor relative humidity (RH_i), cooling set point temperature (T_{set}), cooling system run time (rt_c), heating system run time (rt_h), outdoor temperature (T_o), outdoor relative humidity (RH_o), global horizontal irradiance (I_{sol}), and wind speed (v). To account for the effects of past environmental conditions on occupants' thermostat behavior, $T_i, RH_i, rt_c, rt_h, T_o, RH_o, I_{sol}$, and v are re-encoded with 30-minute moving averages, which allows the inclusion of historical information.

The combined data is preprocessed as follows:

- An thermostat goes into Hold mode when an occupant turns on the Hold function. Hence, a setpoint temperature change followed by the activation of the Hold mode within 30 minutes or during the Hold mode is considered a setpoint adjustment by an occupant.
- The setpoint temperature difference before and after an adjustment (ΔT_{set}) is the dependent variable representing the "degree of change" in the research question.
- Consecutive setpoint changes in the time series are considered as a single setpoint adjustment.
- One's previous setpoint adjustment may affect the next setpoint adjustment. Hence, the setpoint temperature change with the previous setpoint adjustment ($\Delta T_{set,prev}$) is included in the analysis. If there was no previous adjustment under the Hold mode during the previous 30 minutes, it is mapped to 0.
- The global horizontal irradiance may not be suitable for investigating the potential effects of solar radiation, such as direct radiation to occupants, radiation to floors and walls, and indirect heat gain from radiation through external surfaces. For better investigation, the solar irradiances on east, south, west, and north-facing vertical walls are estimated with the Erbs solar decomposition models (Erbs et al., 1982) and equations related to the angle of incidences (Duffie & Beckman, 2013).
- Cyclic variables (i.e., time of the day, day of the week, and day of the year) are decomposed into sine and cosine components.

With the dataset, potential causal factors affecting ΔT_{set} are first discovered by the developed causal discovery method. Subsequently, a causal model, which takes only the causal factors as model inputs, is developed. The performance of the causal model is compared with that of a non-causal model, which takes any inputs maximizing prediction performance. Three datasets from three different periods, (i) from May to June, (ii) July and (iii) September, are used for model training and testing. The datasets are labelled as training, test A, and test B. Since July is subsequent to June, the environmental conditions between July and June would not significantly different. In this regard, the test dataset A from July is more likely to have similar joint variable distributions to the training dataset from May to June. On the other hand, the test dataset B from September is distant from the training period. This historical gap between two datasets would lead to difference in environmental conditions and variable distributions in the training and test dataset B. Thus, the test dataset B is more vulnerable to the data shift (i.e., the joint variable distributions in training and test datasets are different) compared with the test dataset A. The number of data points in the training set is 299, and in the test sets are 167 and 104. Table 1 shows the basic statistics (e.g., mean and standard deviation) of variables in the training and test datasets. I_{sol} and cyclic variables are summarized with the statistics before being decomposed.

Table 1: Statistics of variables in training and test datasets.

Variable	Unit	Training (May – June)				Test A (July)				Test B (September)			
		μ	σ	Max	Min	μ	σ	Max	Min	μ	σ	Max	Min
T_i	℃	23	1	26	22	24	1	27	21	25	1	26	23
RH_i	%	49	5	62	38	49	5	61	40	49	6	69	35
T_o	℃	18	4	33	11	18	3	28	14	21	4	30	12
RH_o	%	64	13	88	32	71	13	94	37	66	18	93	23
v	m/s	20	8	45	1	16	8	30	0	15	8	36	0
I_{sol}	W/m^2	410	376	1025	0	385	365	1013	0	275	321	849	0
rt_c	sec	82	115	300	0	118	129	300	0	106	130	300	0
rt_h	Sec	19	50	300	0	21	51	243	0	23	48	203	0
$\Delta T_{set,prev}$	℃	0	1	6	-3.3	0	1	3	-3.4	0	1	6	-1.1
T_{set}	℃	24	2	29	22	25	2	29	21	25	2	29	22
Time of the day	-	15	6	24	1	14	6	24	0	15	6	23	1
Day of the week	-	3	2	6	0	3	2	6	0	3	2	6	0
Day of the year	-	153	18	181	121	199	9	212	184	258	9	273	244
ΔT_{set}	℃	-0.8	2	6	-5	-0.4	2	5	-5.6	0	2	6	-5.6

With the training dataset, two models (i.e., causal and non-causal models) were developed. The causal model refers to a prediction model including only causal variables inferred in Section 3.1 as explanatory variables to predict the setpoint temperature adjustment, ΔT_{set}. Otherwise, the non-causal model is a prediction model that includes all available variables as explanatory variables to predict the setpoint temperature adjustment, ΔT_{set}. Equation 3 describes the posterior distributions of model parameters. Especially for the non-causal model, we assigned automatic relevance determination (ARD) priors ($\boldsymbol{\tau_0}$) over the precision of $\boldsymbol{\beta}$ and let a model choose relevant features by itself (R. M. Neal, 1996). For $\boldsymbol{\tau_0}$, the Gamma distribution having mean and variance close to 1 and 0 are generally used (Miyamoto et al., 2015). We set $\alpha_0 = \beta_0 = 0.001$.

$$\begin{gathered} p(\boldsymbol{\beta}, \boldsymbol{\tau_0}, \tau | \Delta \mathbf{T_{set}}, \mathbf{X}) \propto p(\tau) p(\boldsymbol{\tau_0}) p(\boldsymbol{\beta}|\boldsymbol{\tau_0}) \prod_{n=1}^{N} p(\Delta T_{set,n}|\mathbf{x}_n, \boldsymbol{\beta}, \boldsymbol{\tau_0}, \tau), \\ \text{where } p\left(\Delta T_{set,n}\middle|\mathbf{x}_n, \boldsymbol{\beta}, \tau, \boldsymbol{\tau_0}\right) = \mathcal{N}\left(\Delta T_{set,n}\middle|\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}, \tau^{-1}\right),\ p(\boldsymbol{\beta}|\boldsymbol{\tau_0}) = \mathcal{N}(\boldsymbol{\beta}|0, \boldsymbol{\tau_0}^{-1}), \\ p(\boldsymbol{\tau_0}|\alpha_0, \beta_0) = \Gamma(\boldsymbol{\tau_0}|\alpha_0, \beta_0), p(\tau) = \Gamma(\tau|\alpha_0, \beta_0) \end{gathered} \tag{3}$$

After the model training, each model needs to predict the new setpoint temperature adjustment, $\Delta T_{set,new}$, with new data observations. Equation 4 represents the posterior predictive distributions over the new observations from the test datasets. For the model evaluation, the median of posterior predictive samples of the new setpoint temperature adjustment, $\Delta T_{set,new}$, was used for model evaluation.

$$p\left(\Delta \mathrm{T}_{\mathrm{set,new}} \middle| x_{\mathrm{new}}, \mathbf{X}, \Delta \mathbf{T_{set}}, \tau_0, \boldsymbol{\tau}\right) \propto \int p\left(\Delta \mathrm{T}_{\mathrm{set,new}}, \Delta \mathbf{T_{set}}, \boldsymbol{\beta} \middle| x_{\mathrm{new}}, \mathbf{X}, \boldsymbol{\tau_0}, \tau\right) d\beta$$
$$\text{where } p\left(\Delta \mathrm{T}_{\mathrm{set,new}}, \Delta \mathbf{T_{set}}, \boldsymbol{\beta} \middle| x_{\mathrm{new}}, \mathbf{X}, \boldsymbol{\tau_0}, \tau\right) = \tag{4}$$
$$\left[\prod_{n=1}^{N} p\left(\Delta \mathrm{T}_{\mathrm{set,n}} \middle| x_n, \boldsymbol{\beta}, \boldsymbol{\tau_0}, \tau\right)\right] p(\tau) p(\boldsymbol{\tau_0}) p(\boldsymbol{\beta} | \boldsymbol{\tau_0}) p\left(\Delta \mathrm{T}_{\mathrm{set,new}} | x_{\mathrm{new}}, \boldsymbol{\beta}, \boldsymbol{\tau_0}, \tau\right)$$

3. RESULT

3.1 Causal Discovery

Potential causal variables were inferred based on the probabilistic causal discovery approach. To discover potential causal graphs, (i) unconditional and conditional probabilities of edge adjacency between each set of variables were firstly estimated based on Bayesian model comparison. Based on Monte Carlo sampling, every single iteration in the causal discovery process infers a single causal graph. In each iteration, the edge adjacencies were probabilistically determined, and the order of the causal discovery was randomly assigned. A causal graph corresponding to each iteration was inferred based on the constraint-based causal discovery method. The total iterations were 10,000. Figure 4 presents the frequency of different types of edges appearing in the causal graph, between each variable and $\Delta \mathrm{T_{set}}$. The notable frequencies of edge appearing in the causal graph in the plots (i.e., blue, orange, and green bars in the plots) imply that $\mathrm{T_{set}}$, $\mathrm{rt_c}$, $\mathrm{rt_h}$, and $\mathrm{T_o}$ are potential causal factors that affect the setpoint temperature adjustment, $\Delta \mathrm{T_{set}}$.

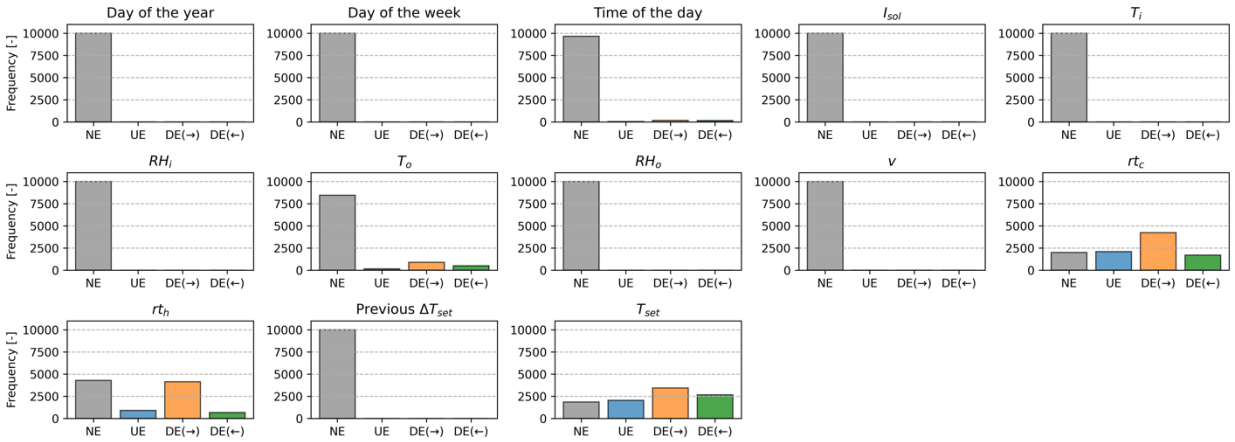

Figure 4: Potential edge between each variable and $\Delta \mathrm{T_{set}}$.
(NE: no edged, UE: undirected edges, DE(→): directed edges from each variable to $\Delta \mathrm{T_{set}}$, DE(←): directed edges from $\Delta \mathrm{T_{set}}$ to each variable)

Preliminary hypotheses about the causal relationships between variables and $\Delta \mathrm{T_{set}}$ were discussed as follows:

- High frequency in the directed edge from the current setpoint temperature, $\mathrm{T_{set}}$, to the setpoint temperature adjustment, $\Delta \mathrm{T_{set}}$, can imply that $\mathrm{T_{set}}$ is a potential causal variable for the setpoint temperature adjustment. The first potential hypothesis is concern about energy consumption. For example, when $\mathrm{T_{set}}$ is low, the occupant may worry about potential energy consumption and try to increase $\mathrm{T_{set}}$. The second potential hypothesis is that the occupant may perceive the current setpoint temperature itself as a guide for decision-making. For instance, the occupant may have one's own preferred $\mathrm{T_{set}}$ range in one's mind. If the current $\mathrm{T_{set}}$ deviates from the range, the occupant may want to make $\mathrm{T_{set}}$ in the range.
- High frequency in the directed edge from the cooling system run time, $\mathrm{rt_c}$, to the setpoint temperature adjustment, $\Delta \mathrm{T_{set}}$, can imply that $\mathrm{rt_c}$ is a potential causal variable for the setpoint temperature adjustment. A potential hypothesis is that an occupant is likely to turn off the cooling by increasing $\mathrm{T_{set}}$ as the occupant perceives that cooling was sufficiently run already.
- High frequency in the directed edge from the heating system run time, $\mathrm{rt_h}$, to the setpoint temperature adjustment, $\Delta \mathrm{T_{set}}$, can imply that $\mathrm{rt_h}$ is a potential causal variable for the setpoint temperature adjustment. A potential hypothesis is that an occupant may attribute discomfort to $\mathrm{rt_h}$. Thus, the occupant may want to decrease $\mathrm{T_{set}}$ in order to turn cooling on when the occupant perceives that heating has been running for a long time.
- High frequency in the directed edge from outdoor temperature, $\mathrm{T_o}$ to the setpoint temperature adjustment, $\Delta \mathrm{T_{set}}$, can imply that $\mathrm{T_o}$ is a potential causal factor of the setpoint temperature adjustment. Hypothetically,

T_o may affect the external heat gain of the residential house and influence internal wall temperature. The changed internal wall temperature may affect a space's mean radiant temperature, which can consequently affect the occupant's thermal comfort and setpoint adjustment decision-making.

- Although indoor temperature, T_i, and relative humidity, RH_i, are well-known factors for the occupants' thermal comfort and can be strong candidates as direct causal factors of the setpoint temperature adjustment, ΔT_{set}, there is no frequency of the adjacent edges between T_i and RH_i. This supports that two variables may have no direct causal relationships to ΔT_{set}. In the training dataset, variable ranges of T_i and RH_i are [22 °C, 26°C] and [38 %, 62 %], lying in the comfort zone recommended by ASHRAE (ASHRAE Research, 2020, pp. 55–2020). This might be a reason that these two variables and ΔT_{set} seem not to be associated, which can hypothesize that T_i and RH_i are not important for this specific household's thermostat decision-making.

3.2 Causal Model Development and Evaluation

Two occupant behavior models—one was causal, and the other was non-causal—were developed for a comparative evaluation, following the modeling procedure described in Section 2.2. The causal model was constructed with the four potential causal factors discovered (Section 3.1). Since the edge frequency of the time of the day was significantly low compared to the others, it was not included in the causal model. The non-causal model was allowed to automatically choose necessary inputs from the dataset.

Figure 5 shows the prediction performance of the causal and non-causal models by comparing predicted setpoint temperature changes $\Delta T_{set,pred}$ with the monitored $\Delta T_{set,true}$ over the two test datasets with or without data shift. The predictions by both models are biased. This is because (i) the complexity of the polynomial relationships is not sufficiently complex to explain the target variable, ΔT_{set}, and (ii) the training and test sets had different data distributions. The dotted trendlines (blue and orange) have 1 as their slope and the mean error (ME) between $\Delta T_{set,pred}$ and $\Delta T_{set,true}$ as their intercept. The distances between trendlines and the diagonal (black) line show how much the predictions by each model are biased compared to the true setpoint temperature changes.

Figure 5 (a) presents the prediction performance of two models over the test dataset A without data shift. The predictions by both causal and non-causal models show similar deviations from the true setpoint temperature changes $\Delta T_{set,true}$ compared to the predictions $\Delta T_{set,pred}$. The ME of the non-causal model, 0.86 °C, was 26.5% lower than that of the causal model, 1.17 °C. The mean absolute error (MAE) of the non-causal model, 1.82 °C, was 3.2 % lower than that of the causal model, 1.88 °C. These results are because the non-causal model has all available information from association with a higher number of variables than the causal model. Thus, the non-causal model can generalize the data better without data shift, when the data distributions between training and test datasets are similar.

Figure 5 (b) shows the prediction performance of two models over the dataset B under data shift. The predictions by the non-causal model more deviate from the true setpoint temperature changes $\Delta T_{set,true}$ compared to the predictions $\Delta T_{set,pred}$ by the causal model. The ME of the non-causal model, 3.29 °C, was 183% higher than that of the causal model, 1.16 °C. The MAE of the non-causal model, 3.44 °C, was 52.2% higher than that of the causal model, 2.26 °C. These outcomes indicate that the causal model, based on potential causal relationships, outperforms the non-causal model, merely association-based, when there is a change in the data distribution. This aligns with the previous literature showing higher robustness of causal models to data shift (Scholkopf et al., 2021).

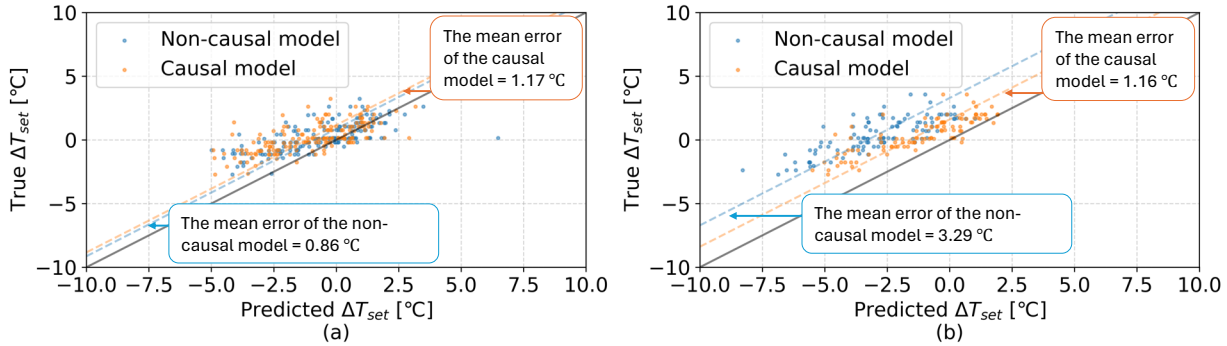

Figure 5: Prediction robustness of each model over (a) the test dataset A without data shift and (b) the test dataset B with data shift.

3.3 Causal Effect Analysis

Figure 6 shows the posterior predictive distribution of setpoint temperature change ΔT_{set}. To visualize the distribution with 2-D figures, for each graph, the input variables, except for the one on the x-axis, were fixed at single values (T_{set} = 25 °C, T_o = 20 °C, rt_c = 100 sec, rt_h = 0 sec). Each figure represents how much each variable influences setpoint temperature change ΔT_{set}. Orange-colored points are the data points in the training dataset.

Figure 6 (a) shows the potential causal effects of the previous setpoint temperature T_{set} on setpoint change ΔT_{set}. The result shows that this household is likely to lower the thermostat setpoint temperature when the current setpoint is high, and vice-versa.

Figure 6 (b) shows the potential causal effects of outdoor temperature T_o on setpoint change ΔT_{set}. The predictive distribution (the solid line and uncertainty band) does not vary with respect to the outdoor temperature. This represents that the change of outdoor temperature is not likely to affect setpoint change significantly, i.e., low direct causal effect from T_o. This is aligned with the causal discovery result (Fig. 3) representing that the direct dependency between ΔT_{set} and T_o is unlikely. However, the high uncertainty suggests further investigations.

Figure 6 (c) shows the potential causal effects of cooling system run time rt_c on setpoint change ΔT_{set}. As rt_c increases from 250 seconds to 300 seconds, ΔT_{set} also increases. A plausible explanation is that the occupants adjusted the setpoint temperature to manage the cooling system operation. When the occupants wanted cooling, if the system was not running, they decreased the setpoint temperature to start the system; conversely, when they did not want cooling, they increased the setpoint temperature to stop cooling.

Figure 6 (d) indicates the potential causal effects of heating system run time rt_h on setpoint change ΔT_{set}. The figure implies that there may be positive effect from the heating system run time. A plausible explanation is that the household did not want cooling to be turned on when heating is on. However, it may be due to observational bias and multicollinearity in the data. In the training dataset, the distribution of heating system run time rt_h is biased to the lower range, and higher rt_h are highly associated with higher T_{set}. These facts suggest considering different ways of data transformation and preprocessing to reach a more concrete conclusion.

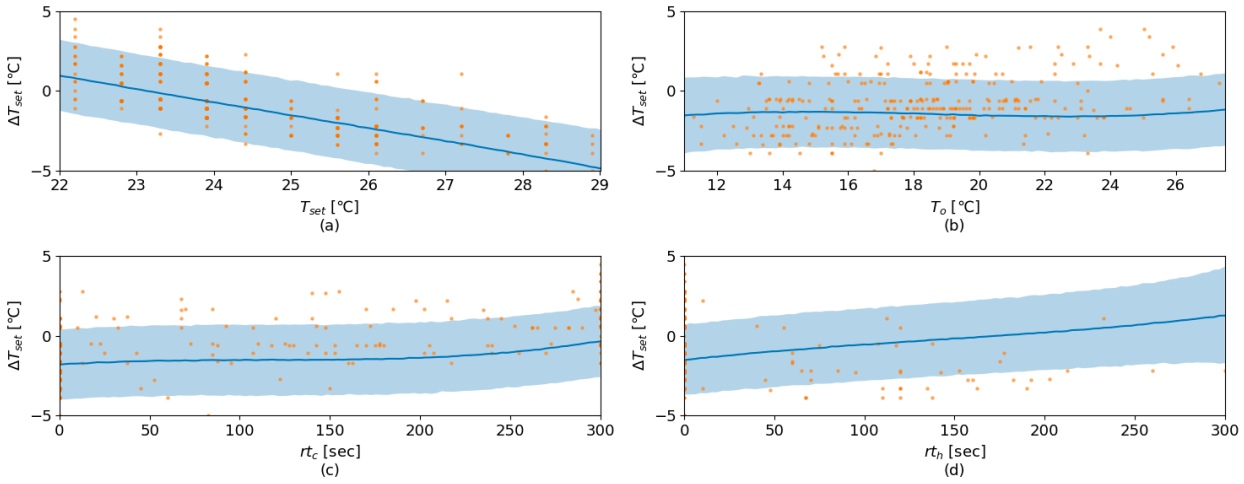

Figure 5: Posterior predictive distribution with variable adjustment
(Solid line and shaded region each refer to the median and 95 % credible regions)

4. DISCUSSION AND CONCLUSION

In this paper, a new probabilistic causal discovery method was proposed with a Monte Carlo method. A Bayesian model comparison was leveraged to estimate the probability of adjacency between each set of variables. Based on the quantified probability of adjacency and a constraint-based causal discovery procedure, the proposed method can infer a set of causal graphs from observational data. To demonstrate the proposed method, with the open dataset provided by a thermostat company, potential causal variables of a household's setpoint temperature adjustment behavior were

inferred. The potential causal variables were used to develop a causal model; on the other hand, all available variables were used to develop a non-causal model. Over the test data without data shift, a causal model and a non-causal model indicated similar prediction performances. However, under the data shift, the causal model showed improved prediction robustness compared with a non-causal model, which emphasizes the importance of developing an occupant behavior model based on causal knowledge.

The proposed causal discovery method has potential limitations in regard to (i) Bayesian model comparison and (ii) causal discovery algorithm. Since the proposed method relies on the Bayesian model comparison result, the model comparison procedure can influence the method's reliability. The Bayesian model comparison needs to assume a specific functional relationship between variables to estimate a probability of adjacency between variables. If a model complexity is too limited to generalize data, the model comparison can provide less reliable results for causal discovery. To tackle such limitation, recent studies have started to introduce the need for more complex functional forms in causal discovery methods such as neural networks (Ke & Bauer, 2022). Furthermore, the model comparison result is influenced by data quantity and quality since the model comparison is inherently data-driven. If observed data cannot sufficiently represent variable spaces of interest (e.g., available data are only a few or biased), this can result in unreliable model comparison results.

In addition, the limitation of the constraint-based causal discovery algorithm also affects the reliability of inferred causal structures. The constraint-based algorithm is sensitive to erratic independence test results as the algorithm is prone to remove edges between variables rather than maintain (Triantafillou et al., 2014). If a single independence result is not accurate, the causal discovery result may become less reliable. Moreover, the constraint-based algorithms rely on the faithfulness assumption (B. Neal, 2020); in other words, independence tests themselves sufficiently can represent the graphical structure. In some cases, the faithfulness assumption can be violated, which can prevent us from discovering true causal relationships.

Despite potential limitations in the proposed method, this paper shows a great potential of employing causal knowledge for occupant behavior modeling in terms of prediction robustness. Especially, for occupant-centric building energy solutions, occupant behavior modeling based on causal knowledge should be implemented to to properly intervene in occupant behaivor to maximize building energy performance and occupants' satisfaction. The following studies need to demonstrate the effectiveness and reliability of a building operational strategy based on a causal occupant behavior model, compared with that based on an association-based behavior model.

REFERENCES

ASHRAE Research. (2020). *ANSI/ASHRAE Standard 55-2020*.

Becerik-Gerber, B., Lucas, G., Aryal, A., Awada, M., Bergés, M., Billington, S. L., Boric-Lubecke, O., Ghahramani, A., Heydarian, A., Jazizadeh, F., Liu, R., Zhu, R., Marks, F., Roll, S., Seyedrezaei, M., Taylor, J. E., Höelscher, C., Khan, A., Langevin, J., … Zhao, J. (2022). Ten questions concerning human-building interaction research for improving the quality of life. *Building and Environment*, *226*, 109681. https://doi.org/10.1016/j.buildenv.2022.109681

Dong, B., Markovic, R., Carlucci, S., Liu, Y., Wagner, A., Liguori, A., van Treeck, C., Oleynikov, D., Azar, E., Fajilla, G., Drgoňa, J., Kim, J., Vellei, M., De Simone, M., Shamsaiee, M., Bavaresco, M., Favero, M., Kjaergaard, M., Osman, M., … Kang, X. (2022). A guideline to document occupant behavior models for advanced building controls. *Building and Environment*, *219*, 109195. https://doi.org/10.1016/j.buildenv.2022.109195

Duffie, J. A., & Beckman, W. A. (2013). *Solar Engineering of Thermal Processes*.

Erbs, D. G., Klein, S. A., & Duffie, J. A. (1982). Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation. *Solar Energy*, *28*(4), 293–302. https://doi.org/10.1016/0038-092X(82)90302-4

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, *10*, 524. https://doi.org/10.3389/fgene.2019.00524

Heckerman, D. (1995). A Bayesian approach to learning causal networks. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 285–295.

IEA. (2022). *Buildings—Tracking report*. https://www.iea.org/reports/buildings

Ke, N. R., & Bauer, S. (2022). *Causality and Deep Learning: Synergies, Challenges and the Future*. International Conference on Machine Learning (ICML) Tutorial, Hall F. https://icml.cc/virtual/2022/tutorial/18442

Khorasani Zadeh, Z., Ouf, M., Gunay, B., Delcroix, B., Larochelle Martin, G., & Daoud, A. (2023). Development of prediction models for thermostat override behavior in direct load control events. *Energy and Buildings*, *301*, 113707. https://doi.org/10.1016/j.enbuild.2023.113707

Kim, J., Schiavon, S., & Brager, G. (2018). Personal comfort models – A new paradigm in thermal comfort for occupant-centric environmental control. *Building and Environment*, *132*, 114–124. https://doi.org/10.1016/j.buildenv.2018.01.023

Ko, J., & Lee, S. (2024). *Bayesian Causal Inference for Occupant-centric Building System Operation*. 2024 ASHRAE Winter Conference, Chicago, IL.

Le, T. D., Hoang, T., Li, J., Liu, L., Liu, H., & Hu, S. (2019). A Fast PC Algorithm for High Dimensional Causal Discovery with Multi-Core PCs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *16*(5), 1483–1495. https://doi.org/10.1109/TCBB.2016.2591526

Lee, S., & Karava, P. (2020). Towards smart buildings with self-tuned indoor thermal environments – A critical review. *Energy and Buildings*, *224*, 110172. https://doi.org/10.1016/j.enbuild.2020.110172

Mehdi Ataei, Murat Erdogdu, Sedef Akinli Kocak, Shai Ben-David, Shems Saleh, Ali Pesaranghader, Andrew Alberts-Scherer, George Sanchez, Saeed Pouryazdian, Ahmad Ghazi, Jennifer Nguyen, Karim Khayrat, & Bo Zhao. (2021). *Understanding Dataset Shift and Potential Remedies* [A Vector Institute Industry Collaborative Project]. Vector Institute. https://vectorinstitute.ai/wp-content/uploads/2021/08/ds_project_report_final_august9.pdf

Miyamoto, A., Watanabe, K., Ikeda, K., & Sato, M.-A. (2015). Variational Inference With ARD Prior for NIRS Diffuse Optical Tomography. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(5), 1109–1114. https://doi.org/10.1109/TNNLS.2014.2328576

Nagy, Z., Gunay, B., Miller, C., Hahn, J., Ouf, M. M., Lee, S., Hobson, B. W., Abuimara, T., Bandurski, K., André, M., Lorenz, C.-L., Crosby, S., Dong, B., Jiang, Z., Peng, Y., Favero, M., Park, J. Y., Nweye, K., Nojedehi, P., … Vellei, M. (2023). Ten questions concerning occupant-centric control and operations. *Building and Environment*, *242*, 110518. https://doi.org/10.1016/j.buildenv.2023.110518

Neal, B. (2020). *Introduction to Causal Inference*. https://www.bradyneal.com/causal-inference-course

Neal, R. M. (1996). *Bayesian Learning for Neural Networks* (Vol. 118). Springer New York. https://doi.org/10.1007/978-1-4612-0745-0

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Sahoh, B., Kaewrat, C., Yeranee, K., Kittiphattanabawon, N., & Kliangkhlao, M. (2022). Causal AI-Powered Event Interpretation: A Cause-and-Effect Discovery for Indoor Thermal Comfort Measurements. *IEEE Internet of Things Journal*, *9*(22), 23188–23200. https://doi.org/10.1109/JIOT.2022.3188283

Sarran, L., Gunay, H. B., O'Brien, W., Hviid, C. A., & Rode, C. (2021). A data-driven study of thermostat overrides during demand response events. *Energy Policy*, *153*, 112290. https://doi.org/10.1016/j.enpol.2021.112290

Scholkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward Causal Representation Learning. *Proceedings of the IEEE*, *109*(5), 612–634. https://doi.org/10.1109/JPROC.2021.3058954

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search* (Vol. 81). Springer New York. https://doi.org/10.1007/978-1-4612-2748-9

Spirtes, P., & Meek, C. (1995). Learning Bayesian Networks with Discrete Variables from Data. *KDD-95 Proceedings*. https://cdn.aaai.org/KDD/1995/KDD95-048.pdf

Triantafillou, S., Tsamardinos, I., & Roumpelaki, A. (2014). Learning Neighborhoods of High Confidence in Constraint-Based Causal Discovery. In L. C. van der Gaag & A. J. Feelders (Eds.), *Probabilistic Graphical Models* (pp. 487–502). Springer International Publishing. https://doi.org/10.1007/978-3-319-11433-0_32

Vehtari, A., & Lampinen, J. (2002). Bayesian Model Assessment and Comparison Using Cross-Validation Predictive Densities. *Neural Computation*, *14*(10), 2439–2468. https://doi.org/10.1162/08997660260293292

Visual Crossing Corporation. (2024). *Visual Crossing Weather (2017)* [Data service]. Retrieved from https://www.visualcrossing.com/.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions. *Bayesian Analysis*, *13*(3). https://doi.org/10.1214/17-BA1091