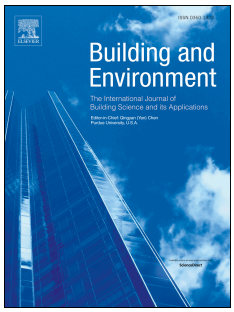
Accepted Manuscript



Exploration of the Bayesian Network framework for modelling window control behaviour

Verena M. Barthelmes, Yeonsook Heo, Valentina Fabi, Stefano P. Corgnati

PII: S0360-1323(17)30466-3

DOI: 10.1016/j.buildenv.2017.10.011

Reference: BAE 5124

To appear in: *Building and Environment*

Received Date: 29 June 2017

Revised Date: 5 October 2017

Accepted Date: 8 October 2017

Please cite this article as: Barthelmes VM, Heo Y, Fabi V, Corgnati SP, Exploration of the Bayesian Network framework for modelling window control behaviour, *Building and Environment* (2017), doi: 10.1016/j.buildenv.2017.10.011.

Exploration of the Bayesian Network framework for modelling window control behaviour

Verena M. Barthelmes*[,a], Yeonsook Heo[b], Valentina Fabi[a], Stefano P. Corgnati[a]

[a]*Department of Energy (DENERG), Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy*
[b]*Department of Architecture, University of Cambridge, 1-5 Scroope Terrace, Cambridge, United Kingdom*

*Corresponding author. Tel.: +39 011 0904507; fax: +39 011 0904499
E-mail address: verena.barthelmes@polito.it (V.M. Barthelmes)

Abstract

Extended literature reviews confirm that the accurate evaluation of occupant energy-related behaviour is a key factor for bridging the gap between predicted and actual energy performance of buildings. One of key energy-related human behaviour is window control actions that have been modelled by different probabilistic modelling approaches. In recent years, Bayesian Networks (BNs) have become a popular representation based on graphical models for modelling stochastic processes with consideration of uncertainty in various fields, from computational biology to complex engineering problems. This study investigates the potential applicability of BNs to capture underlying complicated relationships between various influencing factors and energy-related behavioural actions of occupants in residential buildings: in particular, window opening/closing behaviour of occupants in residential buildings is investigated. This study addresses five key research questions related to modelling window control behaviour: (A) variable selection for identifying key drivers impacting window control behaviour, (B) correlations between key variables for structuring a statistical model, (C) target definition for finding the most suitable target variable, (D) BN model with capabilities to treat mixed data, and (E) validation of a stochastic BN model. A case study on the basis of measured data in one residential apartment located in Copenhagen, Denmark provides key findings associated with the five research questions through the modelling process of developing the BN model.

Key words: Occupant behaviour; Bayesian Networks; window control behaviour; stochastic modelling

1. Introduction

Accounting for uncertainty has become a crucial aspect in the domain of building energy simulation for incorporating human behaviour that impacts building energy performance and comfort expectations. Human behaviour such as occupancy, control of energy systems, occupants' interaction with the building envelope and other comfort criteria settings are considered as key sources of uncertainty in the prediction of building energy use. Indeed, occupant behaviour varies significantly between individuals, which results in large variation of the indoor environmental quality and energy consumptions of the buildings [1][2][3]. Extended literature reviews and state-of-the-art analyses confirm that an accurate modelling of occupant behaviour is a key factor to bridging the gap between predicted and actual energy performance of buildings [4][5][6][7][8].

Frequently, simulation-based design analysis relies on standard use and operation conditions such as fixed schedules for occupancy levels, light switching, ventilation rates and temperature setting. These assumptions often lead to an oversimplification of the human-related variables creating discrepancies between predicted and real energy use of the building. Thus, in recent years, probabilistic modelling approaches have been applied to capture the stochastic nature of energy-related human behaviour when predicting building energy consumptions in dynamic simulation programs [9].

Occupant's action of window opening/closing has an important impact on building energy use and indoor environmental quality (IEQ) by changing the amount of fresh air to the building. Several studies have been carried out to develop stochastic models for predicting the occupant's interaction with the windows. These models are based on statistical algorithms to predict the probability of a specific condition or event, such as the window state or the window opening/closing action, given a set of environmental or other influential factors. Most popularly used methods include logit analysis, probit analysis, and Markov chain processes. Nicol [10] developed a logit regression model to predict the state of windows in a probabilistic manner as the function of indoor and outdoor temperatures. Andersen et al. [11] also used a logistic regression model based on a more comprehensive set of indoor and outdoor environmental variables to infer the probability of opening and closing a window. The study on the basis of the field measurements from 15 Danish dwellings defines four separate models of occupants' window action behaviour patterns for different ownerships and ventilation types. Logit regression models have been also applied in other studies for modelling window control behaviour [12][13][14]. Zhang and Barret [15] developed a probit model for predicting window opening/closing actions in a naturally ventilated office building considering the outdoor temperature as the only independent variable. Haldi and Robinson [16][17] tested different modelling approaches and demonstrated that a discrete-time Markov process approach, which takes into account real dynamic processes, leads to a higher predictive power compared with the logit regression approach. Modelling approaches based on Markov chain processes are used in [18] and [19] to predict window states based on their previous states in office buildings and houses, respectively. As these models consider real dynamic processes by providing transition probabilities between the states of a window, they are limited to capture the dynamic effect of changes in indoor and outdoor environmental conditions on window opening and closing actions.

This paper investigates the capabilities of the Bayesian Network framework to model occupant behaviour in the context of thermal comfort and building energy analyses in order to bridge the gap between simulations' outcomes and reality. Bayesian Networks (BNs), or rather graphical belief networks, are widely applicable and have become a popular representation for encoding uncertainty in decision-making processes based on incomplete datasets [20]. In recent years, BNs have been used in many fields, from On-line Analytical Processing (OLAP) [21], cancer prognosis and epidemiology [22], the modelling of dwelling fire development and occupancy escape [23], to speech recognition [24]. In the buildings domain, BNs have been introduced to estimate the effects of the indoor climate on the productivity of occupants [25], to investigate

the relationship between indoor environmental parameters, measurements from body sensors and self-reported activities by the occupants [26], to predict occupancy patterns in buildings [27][28], to model energy-related user behaviour for building energy management [29][30], and to predict indoor environmental conditions [31]. So far, these studies based on BN models treat either discrete variables only or continuous variables only.

In comparison to the above-mentioned regression-based models, BN-based approaches are able to flexibly model complex relationships between diverse explanatory variables and window control behaviour by constructing a joint probability distribution over different combinations of the domain variables. Indeed, the BN model permits to easily model joint conditional dependencies of the entire set of variables through a graphical representation of the model structure [32]. The BN model also allows for structuring a variety of explanatory variables and multiple target variables in a hierarchical manner. In addition, BNs are demonstrated to yield good prediction accuracy even with small datasets [33]. They also have capabilities to handle incomplete datasets by using Expectation-Maximization (EM) algorithms [34] in which missing data can be marginalized by integrating over all the possibilities of the missing values. Furthermore, the BN model provides a clear semantic representation of relationships between variables, which facilitates flexibly structuring a model and training it against available data in wider and interdisciplinary research communities.

This paper demonstrates the applicability of the Bayesian Network (BN) framework for predicting window opening/closing behaviour of building occupants based on the measurements in a residential apartment located in Copenhagen, Denmark. In particular, the paper addresses five key research questions related to developing a BN model for predicting window-use patterns. The first set of three research questions addresses general issues relevant to modelling window control behaviour:

A. Which variables are key drivers that determine window control behaviour?
B. What level of correlations resides between variables and should they be captured in the BN model?
C. What is the most suitable target variable of window control behaviour?

Regarding the first question, the Kolmogorov-Smirnov Test (K-S Test) is applied to evaluate which variables are main drivers for window control actions. For the second question, the Kendall Tau correlation coefficient is used to investigate correlations between identified variables and accordingly model them in the BN. The third question (C) investigates different target variables commonly used in the literature (i.e., window opening/closing event and window state) in terms of the modelling accuracy.

The second set of research questions addresses modelling challenges related to the applicability of the BN framework for modelling occupants' window control behaviour:

D. How to handle mixed data in the BN framework?
E. How to validate stochastic BN models?

A key question of this paper addresses how to handle mixed data in the BN framework. Traditional BN approaches to treat either discrete variables or continuous variables are not suited to modelling window control behaviour as datasets typically consist of both continuous variables (e.g., indoor temperature, CO_2 concentration) and non-continuous variables (e.g., binary control actions, time of the day). This study tries to overcome this problem by proposing a modelling procedure that allows for handling mixed data, particularly with use of the bnlearn package [35] in the statistical software R environment [36]. The prediction accuracy of the model is evaluated through a series of methods suitable to validate stochastic models.

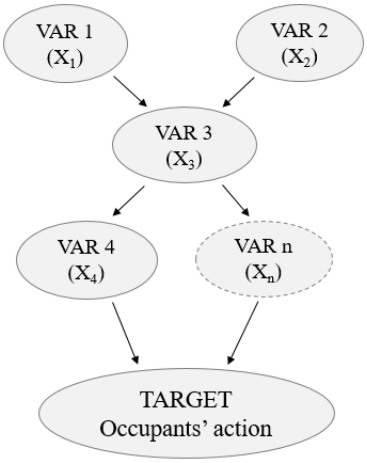
2. The Bayesian Network Framework

2.1 Bayesian Networks

Bayesian Networks are graphical models that represent probabilistic dependencies between discrete or continuous variables (X_i) [37]. In the models, variables are presented by nodes and their relationships are represented by arcs. The direction of arcs determines a hierarchical structure of nodes. Figure 1 shows an example of a Bayesian Network that represents the probabilistic dependencies between an occupant's action and a set of variables (VAR) that potentially impact the action.

A network structure is often explained with a family metaphor; if there is an arc starting from one node to another, the former is a parent of a child (the latter). Extending the metaphor, in a directed chain of nodes, one node is an ancestor of another if it appears earlier in the chain, whereas a node is a descendant of another node if it comes later in the chain. For instance, as shown in Figure 1, as there is an arc from X_1 to X_3, node X_1 is a parent of node X_3. The graphical structure of a Bayesian network, denoted as G=(V,A), is a Directed Acyclic Graph (DAG), where V is the node (or vertex) set and A is the arc (or edge) set. The DAG defines a factorization of the joint probability distribution of $V = \{X_1,X_2,\ldots,X_n\}$, often called the global probability distribution, into a set of local probability distributions, one for each variable [36]. This factorization is based on the assumption that Bayesian Networks have a Markov property [37], which indicates that the state of a random variable X_i depends only on its parents $\prod X_i$. In general, Bayesian network modelling requires the assumption of the Markov property.

Figure 1. Example of a BN: Probabilistic dependencies between occupant behaviour and possible explanatory variables (VAR)/drivers.



In principle, BN models flexibly represent different typologies and handle a mix of various data types. Yet, so far, BN models used in most existing studies are limited to either a discrete case or a continuous case. This limitation is mostly due to the fact that, unfortunately, most software available for developing BN models are applicable to either discrete or continuous data and, thus, do not permit yet to handle a mix of continuous and

discrete datasets in one BN model, particularly when discrete variables are conditional on continuous variables. Equations 1 and 2 define a joint probability distribution for a discrete case and continuous case, respectively:

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i)) \quad \text{(for discrete variables)} \quad (1)$$

$$f(X_1, \dots, X_n) = \prod_{i=1}^{n} f(X_i | Parents(X_i)) \quad \text{(for continuous variables)} \quad (2)$$

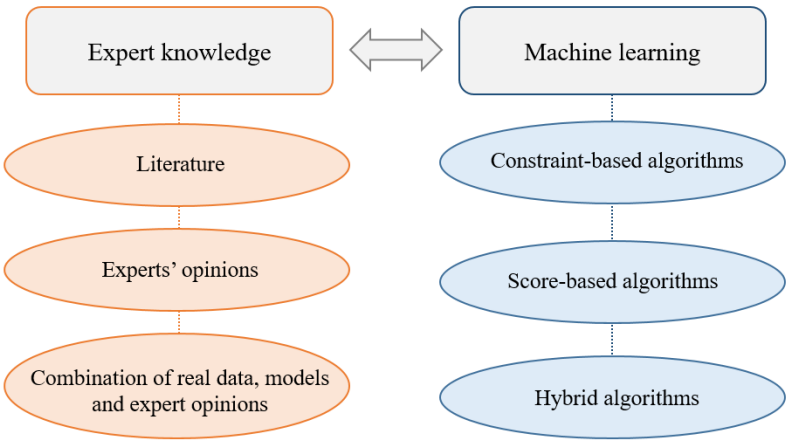
In the discrete case, conditional joint probabilities are represented by the so-called Conditional Probability Tables (CPTs) since all variables are characterized by discrete data. In this case, all intervals for each discrete variable are treated as independent variables, and there is no mechanism to capture the effect of continuous variables such as temperature and relative humidity as a continuous trend. On the other hand, the continuous case assigns each variable X_i with a Gaussian probability density function f (X_i) conditional on the values of its parent nodes. As datasets collected for model development often consist of different data types, many existing studies discretize continuous data for obtaining homogeneous datasets [38][39][28][25]. A key limitation of discretization is a significant loss of information, which has a big impact on the predictive power of resulting BN models and the interpretability of BN models to understand relationships between variables. In fact, data collected for occupant behaviour modelling typically includes both categorical or binary variables (such as window control actions and time-of-day) and continuous variables (such as the indoor/outdoor environmental variables). Hence, it is important to develop a BN framework that

allows for appropriately handling mixed data for occupant behaviour modelling, which will be carefully investigated in Section 4.1.

2.2 Structuring and learning Bayesian Networks

Approaches for developing a BN model can be categorized into two groups. The first group, called as "elicitation", is based on domain experts that rely on expertise to structure a network and quantify probability distributions associated with arcs [40]. This approach can be useful for cases in which field survey data or measurements are not available. The second group is solely based on machine learning algorithms that extract a structure and estimate probability distributions from the dataset [41]. This approach may lead to the model best fit to the training dataset, but whether causal relations between variables derived from the dataset alone are correct needs to be carefully inspected. Alternatively, these two approaches can be combined to fully utilise both expert knowledge and available data; for example, defining the structure of the network based on expert knowledge and learning probability distributions in the BN model from the dataset (Figure 2).

Figure 2. Learning the structure of BNs.



Several machine learning algorithms have been developed to extract a BN structure directly from the dataset. Constraint-based algorithms (conditional independence learners) are all optimized derivatives of the Inductive Causation algorithm [42]. These algorithms use the conditional independence tests to detect the Markov blankets of the variables and accordingly identify causal relationships among variables in a BN network. The main drawback of constraint-based algorithms is that they are not robust to correctly define independencies among variables when they are highly correlated. Another group of algorithms, called search-and-score searches over possible Bayesian Network structures to find the best factorization of the joint distribution [21]. These score-based learning algorithms are general purpose heuristic optimization algorithms which rank testing network structures with respect to a goodness-of-fit score. One of the most commonly used measures in this process is Bayesian Information Criterion score (BIC score) [36], which

measures the model predictability with evaluation of the value of adding more variables into the model. This measure defined in Equations 3 and 4 for the discrete and continuous cases represents a useful tool for optimizing the model in terms of both its predictive power and complexity.

$$BIC = \sum_{i=1}^{n} logP_{X_i}\left(X_i | \prod X_i\right) - \frac{d}{2}\log n$$

$$BIC = \sum_{i=1}^{n} logf_{X_i}\left(X_i | \prod X_i\right) - \frac{d}{2}\log n \qquad \text{(for continuous variables)} \qquad (4)$$

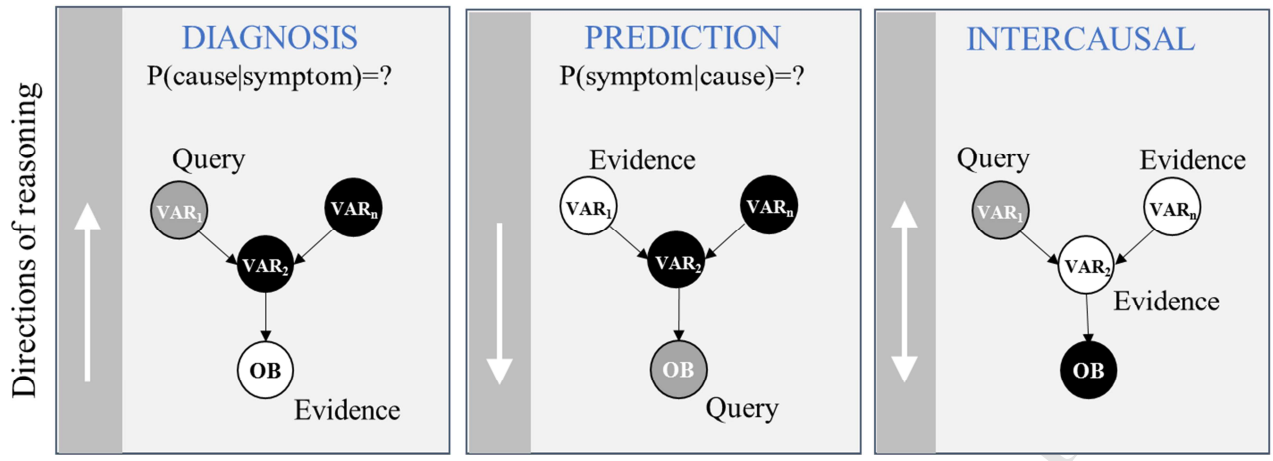
where d is the number of variables included in the BN network and n is the sample size.

Hybrid algorithms are developed to determine a BN network based on both conditional independence tests and network scores. Several commercial software such as Hugin [43], BayesiaLab [44] and Netica [45] provide these algorithms for users to obtain a BN structure directly from a given dataset. These algorithms are also available in statistical computing environments such as R (bnlearn, deal, catnet, pcalg, gRbase, gRain) [36], Matlab (Bayes Net Toolbox) [46], Java [47] or Python [48].

2.3 Reasoning with Bayesian Networks

Bayesian Networks provide full representations of probability distributions over their variables and supporting different types of reasoning. Figure 3 summarizes the main directions of reasoning with BNs in the context of occupant behaviour analysis. BN models permit to perform diagnostic reasoning to understand which variables (VARs) influence occupants' actions (OB) and in which manner: for example, "What specific environmental conditions trigger occupants to open windows?". This type of reasoning occurs in the opposite direction to network arcs to understand what causes certain actions. Another type of reasoning is predictive reasoning, which is typically the major objective of developing occupant behaviour models: for example, "If the outdoor temperature is around 21°C, what is the probability that occupants will open windows?". In this case reasoning follows the direction of the network arcs to predict occupant's action given expected environmental conditions. Another form of reasoning is called as intercausal reasoning that involves reasoning about mutual causes of a common effect [37]: for instance, reasoning about the relationships between several response variables, such as indoor and outdoor environmental variables. Since any nodes in BNs may be query nodes (target variables) and any may be evidence nodes (explanatory variables), sometimes the reasoning does not fit neatly into one of the types described above, and these types of reasoning can be combined in any way. Further detailed information about BNs can be found in [37][49][50][51].

Figure 3. Types of reasoning with Bayesian Networks.



3. Structuring a statistical model for predicting window opening behaviour

This section address the first set of key research questions associated with the process of modelling the window control behaviour introduced in Section 1:

A. Variable selection
B. Correlation between variables
C. Target definition

The modelling process is based on measurements of one natural-ventilated, rented two-persons apartment located in Copenhagen, Denmark [11]. Table 1 summarises measurements related to the indoor and outdoor environment conditions, occupants' interaction with the windows, and time-related factors such as the time of the day or the day of the week. These measurements were collected in 10-minutes intervals continuously for approximately 3 months (February–May). The outdoor environmental measurements were acquired from a meteorological measuring station located near the apartment. The same time resolution was used for analysis.

Table 1. Available target and explanatory variables.*

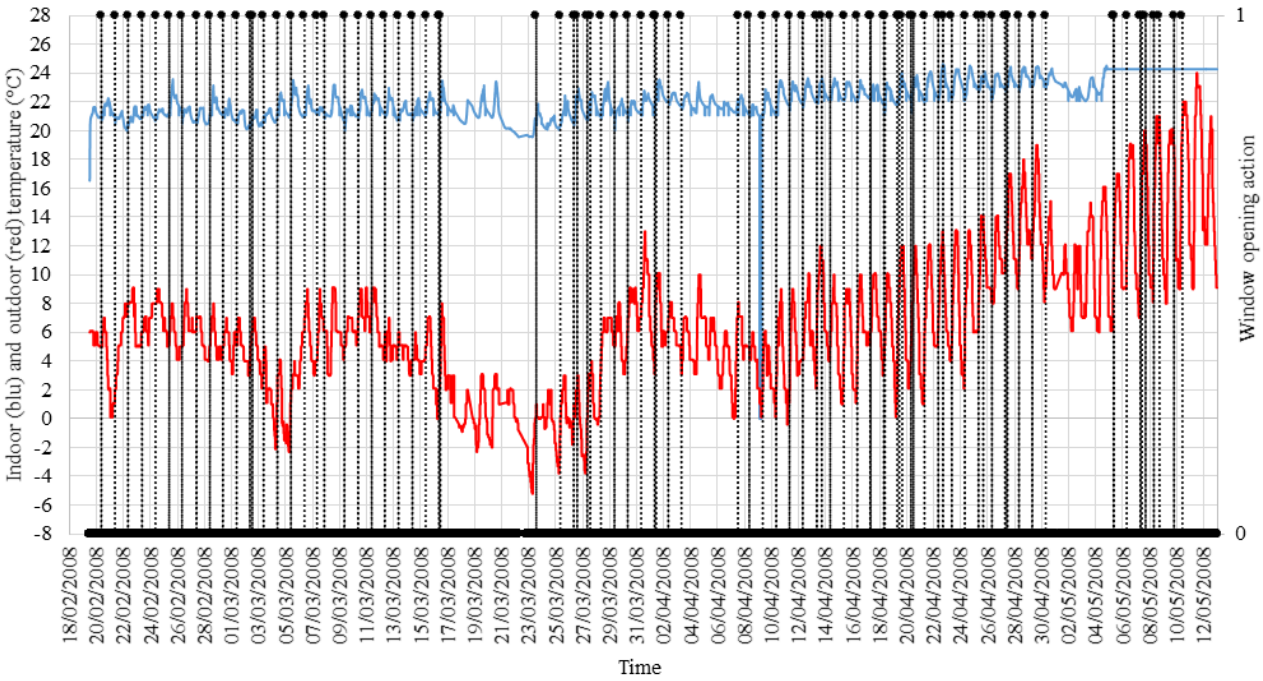
Potential VARs	Abbreviation	Unit	Min	Max	Mean	Median	St. Dev.
Indoor Environment							
Dry bulb temperature	T_{in}	°C	12.1	25	21	21	3
Relative humidity	RH_{in}	%	26	66	38	38	5
Illuminance	Lux	lux	1	8360	95	43	171
CO_2 concentration	$CO_{2,in}$	ppm	101	2261	608	580	161
Outdoor Environment							
Air Temperature	Tout	°C	-5	24	7	6	5
Relative humidity	RH_{out}	%	25	100	73	74	18
Wind speed	Wind	m/s	0	13	3	2	2
Global solar radiation	SR	W/m^2	0	904	184	63	230

Occupant Behaviour		**Range of values**
Window position/state	WS*	0/1 (closed/open)
Window opening/closing action	WOA*/WCA*	0/1 (no action/action)
Other		**Range of values**
Time of the day	Hour	1-24
Weekday	WD	Monday-Sunday

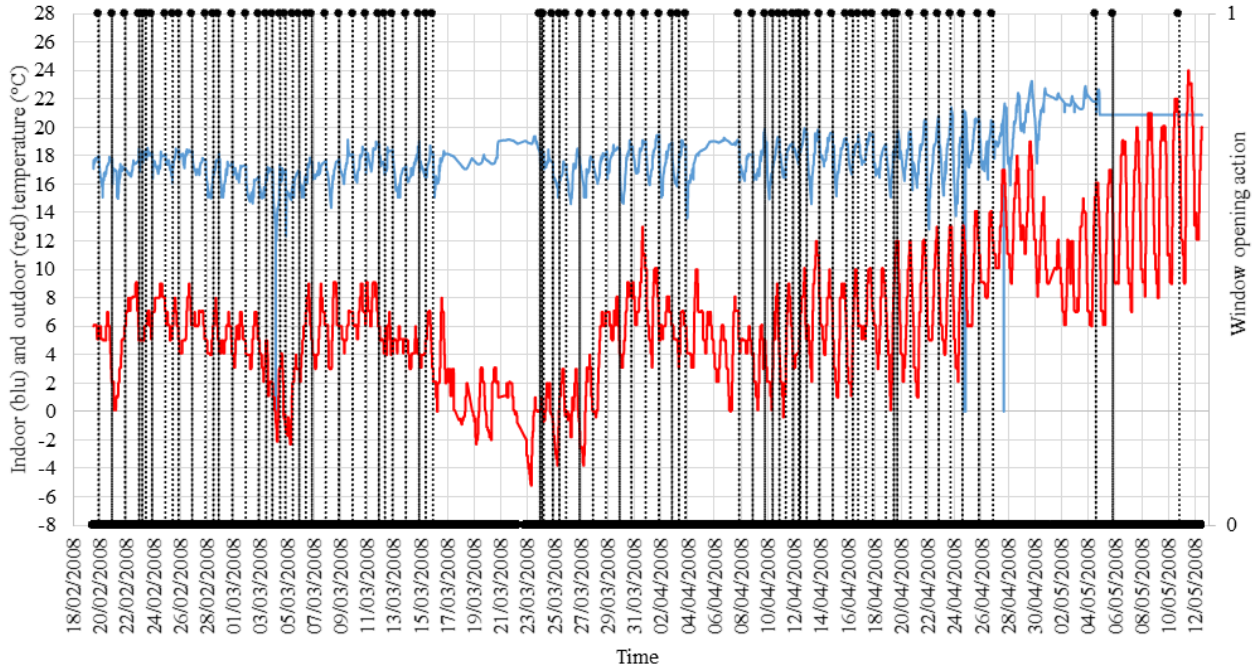
Figure 4 shows the measured window opening actions throughout the monitoring period (vertical black dotted lines), plotted against measured indoor temperatures (blue) and outdoor temperatures (red) in the living room and the sleeping room. One thing to point out is that this study treats window states as a binary variable (0=closed, 1=open) and does not take into account the degree of opening (angle of the shutter with respect to the window frame). Windows are a two-wing window type, manually controlled by the building occupants. Detailed information on window types and measurement instruments can be found in [11]. In total, the occupants performed 215 window opening actions during the monitoring phase.

Figure 4. Window opening actions (black dotted lines) and indoor (blue line) and outdoor (red line) temperature in the living room (a) and the sleeping room (b).

(a)



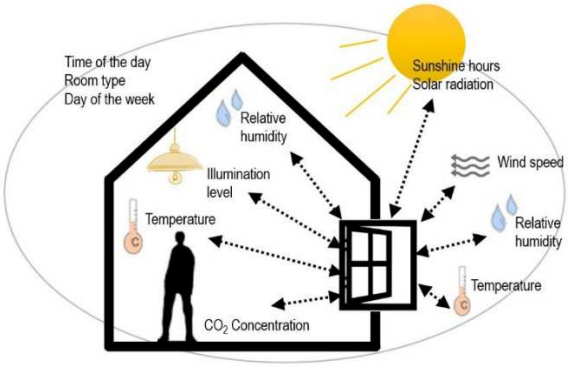
(b)



3.1 Question A: Variable selection

The first question addresses a variable selection step that identifies key explanatory variables that influence window opening/closing behaviour (Figure 5). Borgeson and Brager [52] provide an extensive summary of the literature on modelling studies for predicting occupants' window control behaviour. In most models studied by [52], temperature is considered as the most important driver [53][54], although there is no consensus about whether indoor or outdoor temperature is dominant in determining the behaviour. Other models use time-related factors such as the time of the day and season or the current window state as key variables to predict window control actions [55][17][14]. Review of the existing literature confirmed that the dataset used for this study include key explanatory variables that are found to impact window control behaviour.

Figure 5. Definition of explanatory variables.



As the next step, a two-sample Kolmogorov-Smirnov test (K-S test) was used to test which variables are main drivers that trigger window control actions. The two-sample K-S statistic quantifies a distance between the empirical distribution functions of two samples to evaluate whether two samples come from the same probability distribution function [56]. This method is useful to test whether a certain explanatory variable impacts window control actions by comparing the distribution of variable values when window opening or closing actions is different from that in the entire dataset. First, the entire dataset, including all explanatory variables and window control variable, was generated as a baseline. Then, from (i) the entire dataset, two subsets were generated depending on the window control action: (ii) data only when window opening actions were monitored and (iii) only data when window closing actions were monitored. Hence, (i) provides the distribution of explanatory variable values regardless the window control action, while (ii) and (iii) provide the specific distribution depending on the window control action (opening and closing, respectively). Then, the two-sample K-S test was applied to a pair of samples – (i) and (ii) for the window opening behaviour and (i) and (iii) for the window closing behaviour - for each environmental and time-related variable to examine how different the two samples are. For instance, if the distribution of the indoor air temperature substantially differs between the samples (i) and (ii), it indicates that the indoor air temperature has a significant impact on window opening actions. The statistical significance of differences between the two samples is represented by the p-value; the lower the p-value is, the more the two samples differ. The significance threshold of the p-value is typically 0.05, which is also used in this study to exclude unimportant variables from further analysis.

Figure 6. K-S test: Definition of the samples.

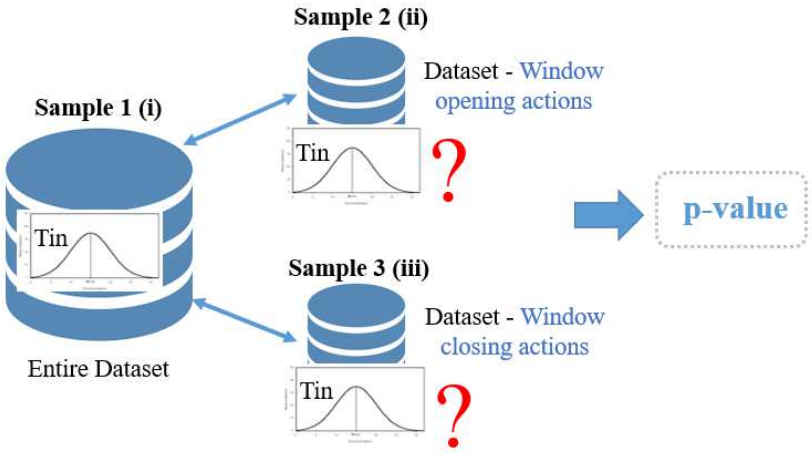


Table 2 shows the K-S test results used to rank the most influencing variables for the window opening and closing behaviour. For the window action behaviour, the results highlight that the six most influencing variables in the case study analysed are the time of the day, CO_2 concentration, solar radiation, indoor and outdoor air temperature and indoor relative humidity. All the variables with a p-value higher than 0.05 were excluded from the analysis (darker grey boxes). One thing to note is that the day of the week (WD) does not influence the window opening action (WOA) at all (p-value =1). Furthermore, the K-S test results reveal that the six most influencing variables are identical for the window opening and window closing actions, while their ranking varies slightly. The most important variable is the time of the day for both actions. Indeed, exploratory data analyses also showed that the windows were opened and closed in certain times of the day (morning and late afternoon hours). The window closing actions were also influenced by the wind speed and the illuminance level.

Table 2. K-S test: Variable selection.

	WINDOW OPENING ACTION (WOA)		**WINDOW CLOSING ACTION (WCA)**	
Rank	**VAR**	**p-value**	**VAR**	**p-value**
1	Hour	5.754×10^{-12}	Hour	2.2×10^{-16}
2	$CO_{2,in}$	8.668×10^{-11}	SR	2.2×10^{-16}
3	SR	3.226×10^{-6}	$CO_{2,in}$	0.000102
4	T_{in}	0.0001399	T_{in}	0.0001399
5	T_{out}	0.005	Tout	0.003193
6	RH_{in}	0.008602	RH_{in}	0.008602
7	Lux	0.15	Wind	0.01012
8	Wind	0.2212	Lux	0.03478
9	RH_{out}	0.335	RH_{out}	0.335
10	WD	1	WD	1

3.2 Question B: Correlations between variables

The second question investigates correlations between the explanatory variables identified in section 3.1 as correlations between the variables need to be carefully treated in development of statistical models to correctly quantify the effect of individual variables on occupants' actions. This study uses the Kendall rank correlation coefficient to relatively evaluate the importance of correlations between the measured variables and accordingly structure the arcs between the explanatory variables in the BN model in an efficient manner. In particular, The Kendall rank correlation coefficient, commonly referred to as Kendall's tau coefficient, is a statistic used to measure the ordinal association between two measured quantities [57].

Kendall τ coefficient is calculated as follows; let (VAR_{x1}, VAR_{y1}), (VAR_{x2}, VAR_{y2}), …, (VAR_{xn}, VAR_{yn}) be a set of observations of the joint random variables VAR_X and VAR_Y respectively, such that all the values of (VAR_{xi}) and (VAR_{yi}) are unique. Any pair of observations (VAR_{xi}, VAR_{yi}) and (VAR_{xj}, VAR_{yj}) are said to be concordant if the ranks of both variables agree; that is, if both $VARx_i > VARx_j$ and $VARy_i > VARy_j$ or if both $VARx_i < VARx_j$ and $VARy_i < VARy_j$. Otherwise, they are said to be discordant. Equation 4 defines the Kendall τ coefficient and n is the total number of combinations:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2} \quad (5)$$

Table 3 shows the ranking of the most correlated variables with the six important drivers and associated Kendall coefficient values. Overall, highly strong correlations between the selected variables are not observed. Mild correlations are observed among the indoor air temperature (T_{in}), the outdoor air temperature (T_{out}), and the solar radiation (SR). As expected, correlations are found between the indoor temperature and relative humidity (T_{in} and RH_{in}) and the outdoor temperature and relative humidity (T_{out} and RH_{out}). Furthermore, minor correlations are found between the time of the day (Hour), the outdoor air temperature (T_{out}), and the solar radiation (SR). These correlations between the selected variables will be represented in the BN model by adding arcs between the identified pairs with correlations. It is worth noting that this analysis intends to evaluate all the correlations between the variables in a relative manner without specific numerical thresholds to define the importance of correlation.

Table 3. Kendall's Tau: Nonlinear correlation between the most influencing variables on window action behaviour and the other variables.

Ranking	Hour		$CO2_{in}$		SR		T_{in}		T_{out}		RH_{in}	
1	T_{out}	0.16	T_{out}	0.17	Luxin	0.38	RH_{in}	0.39	RH_{out}	0.40	T_{in}	0.39
2	SR	0.16	Hour	0.13	T_{out}	0.36	T_{out}	0.36	T_{in}	0.36	RH_{out}	0.25
3	$CO_{2,in}$	0.13	RH_{in}	0.10	T_{in}	0.21	RH_{out}	0.23	SR	0.36	$CO_{2,in}$	0.10
4	T_{in}	0.11	Lux_{in}	0.08	Wind	0.20	SR	0.21	$CO_{2,in}$	0.17	SR	0.09
5	Wind	0.10	WD	0.02	Hour	0.16	Hour	0.11	Hour	0.16	Wind	0.09
6	RH_{in}	0.06	SR	0.02	$CO_{2,in}$	0.02	Lux_{in}	0.03	Wind	0.09	Lux_{in}	0.07
7	Lux_{in}	0.03	Wind	0.02	WD	-0.01	Wind	0.02	RH_{in}	0.06	WD	0.06
8	WD	0.01	T_{in}	0.02	RH_{in}	-0.09	$CO_{2,in}$	0.02	WD	0.04	T_{out}	0.06
9	RH_{out}	-0.22	RH_{out}	0.02	RH_{out}	-0.45	WD	0.01	Lux_{in}	0.00	Hour	0.06

3.3 Question C: Target definition

The third question examines the suitability of different target variables to predict occupants' window control actions. Previous models compute the probability of windows being open or closed [16] or to the probability of occupants taking window opening or closing actions [11][13]. Figures 7 and 8 show results from using these two variables as a target variable predicted as the function of the indoor air temperature. Figure 7 depicts the counterintuitive trend of the probability of windows being open increasing as the indoor air temperature decreases. This misrepresentation is due to strong bi-directional interactions between the indoor environmental variables and the window state. When the window state is 1 (open window), cool air flows into the room, lowering the indoor air temperature and the CO_2 concentration. Hence, using the window state as a target variable may lead to unreliable outcomes indoor environmental variables are used as explanatory variables. Andersen et al. [11] also pointed out that it is problematic to infer the window state based on indoor environment conditions (e.g. indoor temperature) since these are directly influenced by the state of the window. Figure 8 highlights that using the window opening action (WOA) as a target variable instead of the window state overcomes this problem by taking into account the values of the indoor environmental variables only when the window is actually being opened (or closed). It is worth mentioning that using the WOA rather than the WS may lead to weaker arc strengths in the BN model since much less data is used for training the model (e.g., 215 data points when WOA took place out of the entire set of 65335 data points).

Figure 7. Probability of an open window (WS) depending on the indoor air temperature.

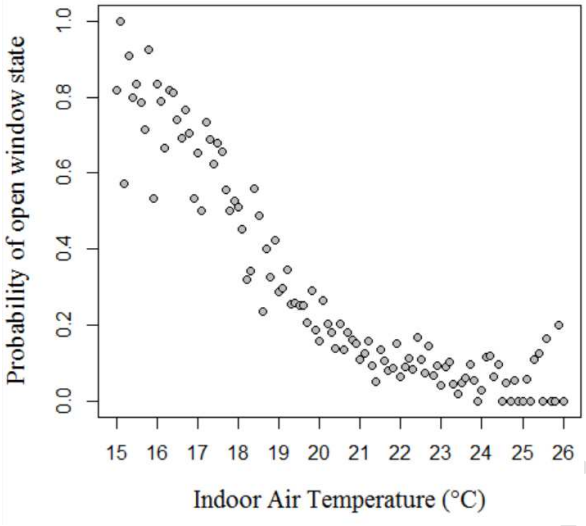
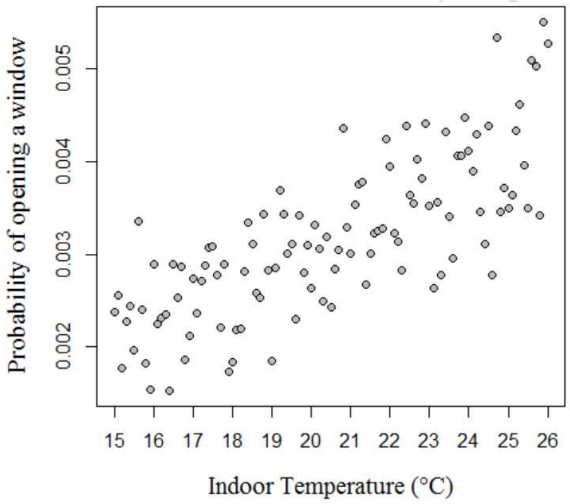


Figure 8. Probability of an open window action (WOA) depending on the indoor air temperature.

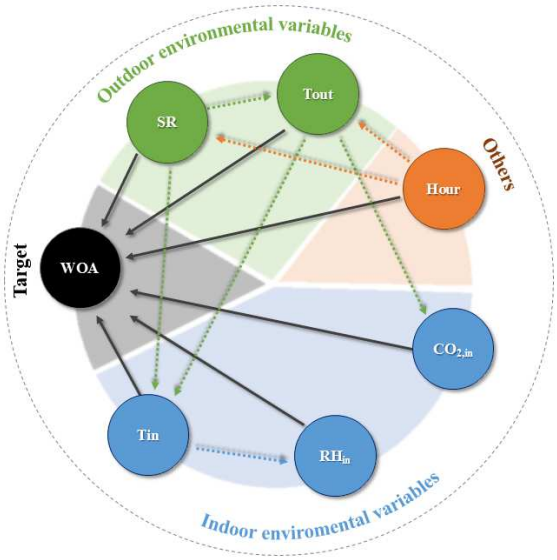


4. BN modelling for predicting window opening behaviour

Figure 9 shows the proposed Bayesian Network for predicting window opening actions developed on the basis of the analysis results in Section 3. As outlined in Section 3.1 (Question A), the key variables that most influence the window control behaviour are the time of the day, indoor CO_2 concentration, solar radiation, indoor air temperature, indoor relative humidity, and outdoor air temperature. We highlight that this study is based on measurements from one residential unit with the four-month of measurements and consequently the proposed model may not include potentially significant drivers that impact window opening actions. such as the season, ventilation type, room type, occupants (e.g., age, gender, smoker/non-smoker), building characteristics, noise level, and security issues. On the basis of the outcomes in Section 3.2 (Question B), the pairs of the variables with stronger correlations are linked by arcs. As the correlation results do not provide causal relationships between the variables, the directions of the arcs are determined based on building physics. Following the findings in Section 3.3 (Question C), the target variable is the window opening action

instead of the window state. As an extension, the window closing action (WCAs) can be included in the same model.

Figure 9. Proposal of a Bayesian Network for window opening behaviour.



With the determined BN structure, parameter learning is carried out to train unknown parameters associated with conditional distributions in the BN against the dataset. Typically, in this process, the usability of the model is evaluated by the BIC score , and the importance of the variables is evaluated by the strengths of the arcs connected between the variables [35][36]. The BIC score is a criterion used to select the best model among a given set of models in terms of the predication accuracy and the model complexity; the lower the BIC score is, the better the model is. The arc strength measures the importance of individual parent nodes on predicting the state of their child node. The strength is measured by the score gain or loss as the result of removing one arc while keeping the rest of the network fixed. Negative strength values indicate decreases in the network score due to the arc's removal, and positive values indicate increases in the network score; the lower the arc strength is, the stronger the relationship between the two variables linked by the arc is. As the proposed BN structure can be applied for both discrete and continuous cases, this paper compares the BN model based on a fully discrete dataset (Models A, C and E) and on a fully continuous dataset (Models B, D and F). Furthermore, the proposed BN structure (Models E and F) is compared against the structure derived only by machine learning (Models A and B) and the Naïve BN where WOA is the only child node and there is no arc between the other variables (Models C and D). For the discrete case, the continuous data is discretised into equal intervals of values based on logical reasoning as shown in Table 4.

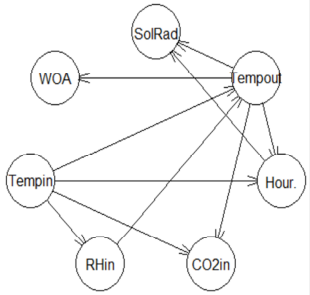
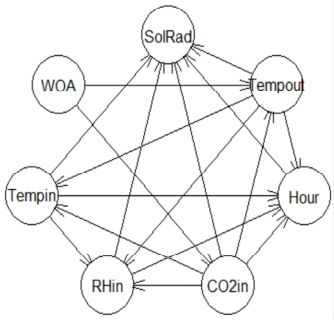
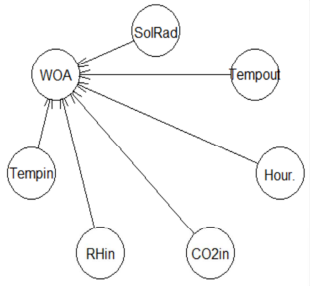
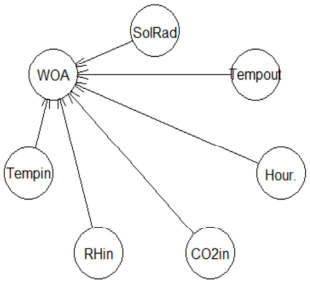
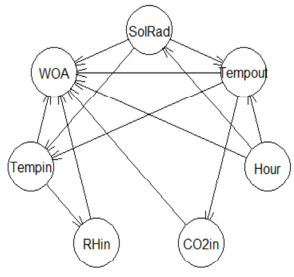
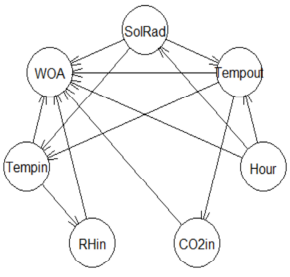
Table 4. Discretization of the continuous VARs.

VAR	**Discrete values**
T_{in}	<14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25
T_{out}	-5-0, 1-5, 6-10, 11-15, 16-20, 21-25,
SR	0-250, 251-500, 501-750, 751-1000
RH_{in}	<35, 35-40, 41-45, 46-50, 51-55

$CO_{2,in}$	0-500,501-1000, 1001-1500, 1501-2000, 2001-2500

Figure 10 summarises the BIC score and arc strengths of different BN models. The BNs were modelled with the R bnlearn package [35] and the structure of the variables was established by a search-and-score-based algorithm (Hill-Climbing algorithm) [36]. Models A and B show that the learning algorithm alone is not able to derive the BN structure that correctly captures relationships between the physical variables. The arcs automatically created by the learning algorithm do not represent the real physical dynamics beyond correlations between the variables. In addition, comparison between Models D and F highlights that the correlations between the explanatory variables are very high but the effect of modelling correlations between the variables on the model predictive power is very minor as the BIC score of Model F does not change much from that of Model D. The models based on the discrete data (Models C and E) are not able to quantify probabilistic dependencies between the explanatory variables and the WOA, while the continuous data (Models D and F) allows for identifying probabilistic dependencies between them. Indeed, the discretization of the dataset leads to a significant loss of information. Different discretization techniques have been developed to maintain substantial information embedded in the continuous dataset in the discretisation process. Suzuki [58], for instance, proposed a scoring method that incrementally discretises the continuous data at finer resolution and evaluates the predictive power of the resulting models. On the other hand, the continuous data cases hold all information, but they do not appropriately handle categorical variables (e.g., time of the day) and binary variables (e.g., window control actions). Recent studies, such as [59] developed methods for learning BNs from datasets joining continuous and discrete variables, but they are not readily available for the wider research community.

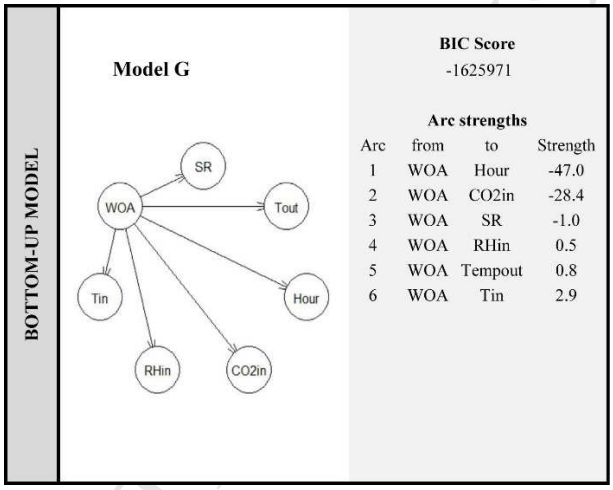
Figure 10. Exploitation of BNs for modelling window opening behaviour.

	DISCRETE DATASET	CONTINUOUS DATASET
ONLY MACHINE LEARNING	**Model A**  **BIC Score** -443,240 **Arc strengths** Arc from to Strength 1 Hour SolRad -20,141 2 Tin Rhin -21,944 3 Tin Tout -17,647 4 Rhin Tout -7,618 5 Tout SolRad -6,550 6 Tout Hour -4,274 7 Tin CO2in -6,379 8 Tout CO2in -4,034 9 Tin Hour -3,675 10 Tout WOA -1,410	**Model B**  **BIC Score** -1,495,630 **Arc strengths** Arc from to Strength 1 Tin RHin -15,880 2 Tout SolRad -8,824 3 Tout Tin -8,745 4 CO2in RHin -2,917 5 Tout RHin -2,674 6 CO2in Hour -1,587 7 Tout Hour -1,402 8 CO2in Tout -1,331 9 RHin SolRad -1,041 10 CO2in Tin -548 11 CO2in SolRad -279 12 RHin Hour -264 13 Tin SolRad -194 14 Tin Hour -149 15 Hour SolRad -26 16 WOA CO2in -20 17 WOA Tout -5
NAIVE BN	**Model C**  **BIC Score** -INF **Arc strengths** Arc from to Strength 1 Tin WOA -Inf 2 Tout WOA -Inf 3 SolRad WOA -Inf 4 Hour WOA -Inf 5 CO2in WOA -Inf 6 RHin WOA -Inf	**Model D**  **BIC Score** -1,539,440 **Arc strengths** Arc from to Strength 1 Tin WOA 5 2 Hour WOA 5 3 RHin WOA 5 4 SolRad WOA 3 5 Tout WOA -5 6 CO2in WOA -25
MACHINE LEARNING + EXPERT KNOWLEDGE	**Model E**  **BIC Score** -INF **Arc strengths** Arc from to Strength 1 Hour SolRad -21,959 2 Tin RHin -21,944 3 Tout Tin -14,487 4 SolRad Tout -6,550 5 Hour Tempout -4,130 6 SolRad Tin -2,423 7 Tout CO2in -1,602 8 Tin WOA -Inf 9 Rhin WOA -Inf 10 CO2in WOA -Inf 11 Hour WOA -Inf 11 SolRad WOA -Inf 13 Tout WOA -Inf	**Model F**  **BIC Score** -1,503,114 **Arc strengths** Arc from to Strength 1 RHin Tin -13,788 2 SolRad Tout -10,524 3 Tout Tin -7,618 4 Tout CO2in -1,325 5 Hour Tout -1,028 6 Hour SolRad -700 7 SolRad Tin -74 8 CO2in WOA -24 9 Tout WOA -3 10 SolRad WOA 3 11 Hour WOA 5 11 Tin WOA 5 12 RHin WOA 5

4.1 Question D: Treatment of mixed data

This section proposes a BN modelling procedure that properly treats mixed data. This capability is crucial especially for the context of window control behaviour in which the main target variable is often binary (open/close) and key response variables are continuous. In particular, the target node "WOA" and time of the day are discrete variables while all the indoor and outdoor environmental variables are continuous. Currently, most available statistical analysis packages, including the bnlearn package, support either discrete or continuous variables. The bnlearn package offers more flexibility as it does not support the dependence of discrete variables on continuous variables but support the other way around. Hence, it is possible to build a bottom-up model in which the arcs are reversely connected from the discrete target variable to the continuous response variables (Figure 11). The semantic representation of this model might seem less intuitive, but since the BN model supports any direction of reasoning, it still can correctly infer the window opening action given the set of variable values.

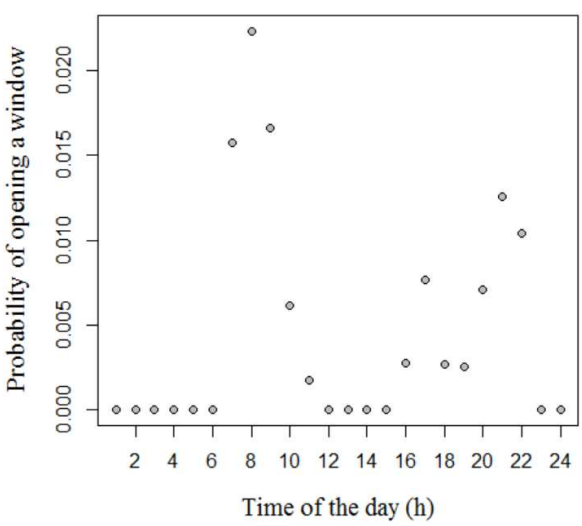
Figure 11. Treatment of mixed data: Bottom-up (BU) model



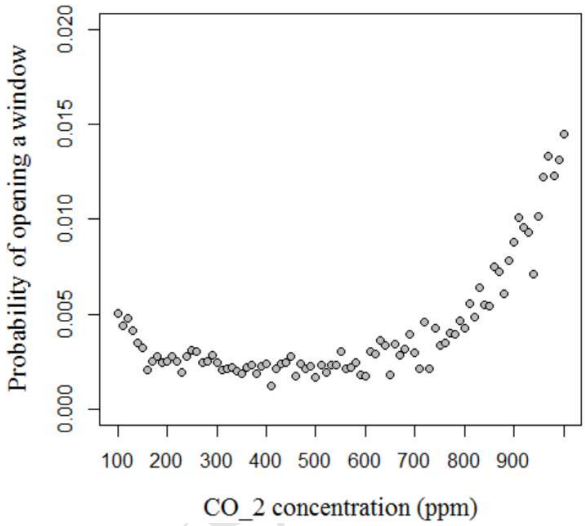
The BIC score of the model suggests that appropriately handling the mixed data improves the predictive power of the model in comparison to Models C and D. Furthermore, Model G yields the ranking of the response variables that well aligns with the outcomes of the K-S test described in section 3.1. In contrast, the continuous case (Model D) results in a much lower arc strength value for the time of the day as it does not correctly treat this variable as a categorical variable and instead expects a consistent trend between this variable and its child node. This comparison clearly illustrates the importance of appropriately treating mixed data to yield a reliable BN model and correctly analyse the effect of different variables on control actions.

Figure 12 depicts the outcomes of the queries related to the probability of a window opening action given the main key response variables (Model G). As regards the main influencing driver, the time of the day (Fig. 12a), the results show that the probability of performing a window opening action is higher during the morning and late afternoon/evening hours. Furthermore, also found in the existing literature [11][14], the results indicate that the probability of opening a window increases in correspondence of a higher CO_2 concentration (Fig. 12b), indoor air temperature (Fig. 12c), and outdoor air temperature (Fig. 12d).

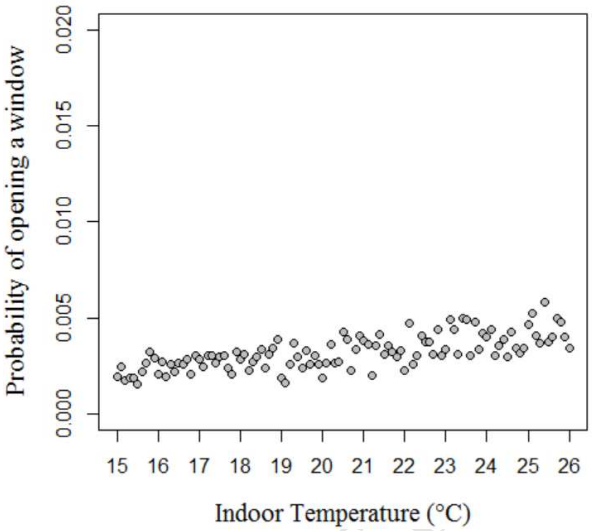
Figure 12. Probability of a window opening action given (a) time of the day (b) CO_2 concentration, (c) indoor temperature, (d) outdoor temperature, (e) solar radiation, and (f) indoor relative humidity



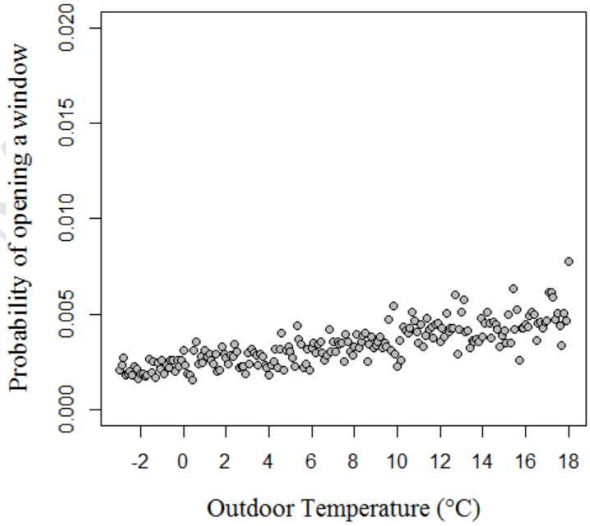
(a)



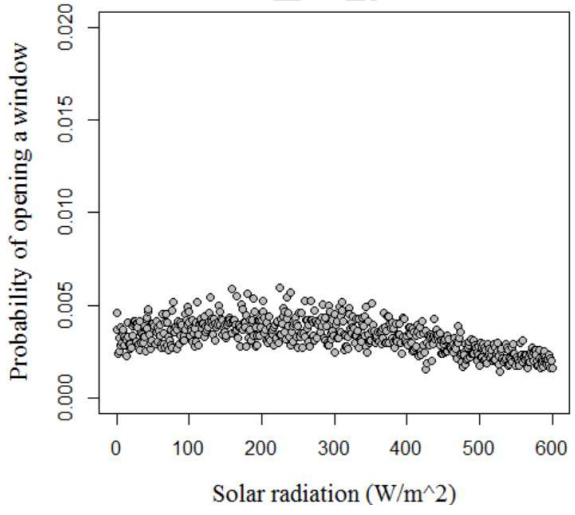
(b)



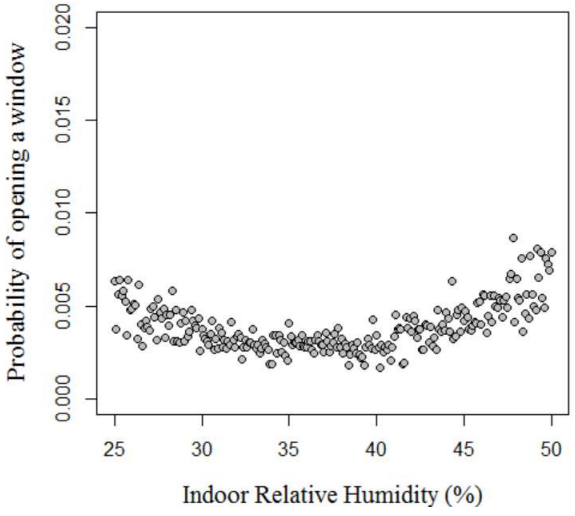
(c)



(d)



(e)



(f)

4.2 Question E: Model validation

This section investigates validation approaches, which is a crucial step to test the predictive power of stochastic models. In particular, this research step validates and tests the predictive power of the final BN model described in Section 4.1. For model validation, cross-validation is a standard way to obtain unbiased estimates of a model's goodness of fit by partitioning the dataset into training and testing subsets. K-fold cross-validation in the bnlearn package is applied to randomly partition the entire dataset into *k* equally sized subsamples. Out of the *k* subsamples, a single subsample is retained as the validation data for testing the accuracy of the trained model, and the remaining *k*-1 subsamples are used as training data. In this case study, the dataset was split into 10 subsets, and the BN model was trained against 9 subsets and tested against 1 subset.

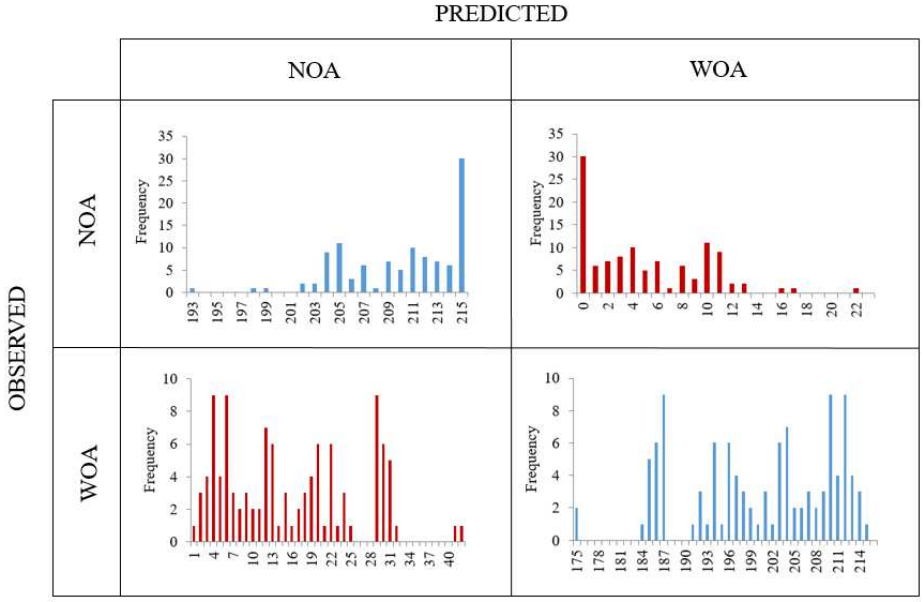
In cross-validation for classification problems similar to the context of predicting binary control actions, the prediction error of a stochastic model is commonly calculated by a loss function that compares the predicted label of the target variable against measurements through the testing dataset. The expected loss value of the final BN model (Model G) is 0.5% (k=10). Although this indicates that predictions are wrong only 5 times out of 1000, it is worth mentioning that as the original dataset is very unbalanced (0.3% of the dataset corresponding to "window opening = TRUE" events and 99.7% of the dataset corresponding to "window opening = FALSE" events), the low classification error does not guarantee that the model reliably predicts the window opening action.

A tailored approach for model validation is applied to examine the model predictability in detail with considering the imbalance of the dataset. This approach consisted of the following steps:

1. Creation of a testing dataset containing 215 samples of "window opening = TRUE" and 215 samples of "window opening = FALSE";
2. Computation of model predictions by the BN model for given response variable values in the testing dataset;
3. Creation of a confusion matrix of observed and predicted WOAs and NOAs.

Steps 1 and 2 were repeated approximately 100 times to obtain the probabilistic distribution of prediction accuracy. The confusion matrix in Figure 13 indicates the model yields in average correct predictions 410 times out of 430. In detail, the accuracy of the model to predict the window opening action and no opening action is in average 93% and 98%, respectively. The expected loss value obtained with the balanced data is 5%, which also confirms the strong predictive power of the BN model.

Figure 13. Confusion matrix of observed and predicted WOAs and NOAs.



5. Discussion

As the case study in this study is based on measurements from one Danish residential apartment, statistical results from the case study are limited to draw generalizable findings due to the small sample size. Nevertheless, the case study serves as an adequate and useful testbed to investigate the applicability of the BN framework for modelling window control behaviour and demonstrate the statistical methods used for variable selection and model validation in the modelling process. A next step is to use an extensive dataset from a large number of residential buildings to develop a generalizable model. We highlight that the case study in this paper focused on environment- and time-related variables for predicting window control actions. In the further work, it is necessary to investigate other building-related factors that may yield different patterns of window control behaviour, such as different ventilation strategies (i.e., presence of controlled mechanical ventilation), room type, and building design characteristics. More importantly, as substantial variation is observed in the window control behaviour due to individual users [60], further work is needed to include contextual information such as occupant types (e.g. age, gender, smokers/non-smokers), social factors (energy-related knowledge and attitudes), and psychological and physiological factors.

The case study demonstrated that the proposed BN approach yields a probabilistic prediction model with higher confidence and better interpretability by fully exploiting information from the mixed dataset in comparison to the typical BN approaches. In all BN models, however, the joint distribution of all variables (global distribution) is factorised into local probability distributions, which reduces computational requirements for complex networks and increases power for parameter learning [61]. On the other hand, this also means that the local probability distribution between two nodes only explain the effect of the parent node on the child node, but does not take into account parameter interaction effects [61]. Although the case study in this paper showed the high predictive power of the BN model without accounting for parameter

interactions, it is not sufficient to conclude the effect of parameter interactions on the model predictive power. Further investigation is necessary to test the importance of parameter interactions in the context of window control behaviour modelling with Bayesian Networks.

The BN-based approach, in principle, allows for modelling complex hierarchical relationships between a large number of continuous and discrete variables through a clear semantic graphical representation. Moreover, the graphical representation is a valuable conceptual benefit since the structure and its underlying probabilistic dimension are easily interpretable for modellers in the building simulation community. However, owing to the limitation of the existing statistical packages, BN approaches used in existing studies with mixed data are based on discretized data of continuous and discrete variables, which may likely result in a significant loss of information [59]. As the first step to overcome this limitation, this paper proposed the bottom-up modelling approach that handles mixed data when the target node is discrete and depends on continuous and/or discrete explanatory variables. However, the proposed approach is not extendible to model a hierarchical complex structure that links continuous and discrete explanatory variables in multiple layers. In fact, occupants take a specific action or combination of actions among many control actions, such as thermostat settings, light dimming, blind control, to maintain their thermal and visual comfort level, and modelling a series of control actions has been identified as one of the future needs for occupant behaviour modelling [5]. A Bayesian hierarchical network model can provide a mathematical framework for holistically modelling such adaptive actions in relation to environmental and contextual variables.

This study mainly validated the technical performance of the BN model for predicting occupant control actions through the case study of a single residential unit. In addition, comparison of the BN approach against the existing statistical methods, such as logistic regression and Markov chain processes, is essential to test whether the BN approach offers improvement in prediction accuracy. Further work is underway to compare the BN model against the above-mentioned existing statistical methods through the same case study. Future research steps for improving and validating the model include further developing the model with the extensive dataset and evaluating the applicability of the model to other residential buildings.

6. Conclusions

This paper proposed a Bayesian Network modelling as a methodology to model window opening behaviour of occupants in residential buildings. The case study on the basis of measured data in a residential apartment located in Copenhagen, Denmark demonstrated the potential benefits of using the Bayesian network framework for modelling stochastic processes of energy-related behaviour with consideration of various factors that drive final control actions. The key research questions related to modelling stochastic window control behaviour were addressed through the case study and key findings are summarised below:

(A) The Kolmogorov-Smirnov (K-S) two sample test allows for identifying key variables that impact window control actions regardless the data type (i.e., continuous, categorical) and underlying trend between

variables. The K-S two sample test ranked the following variables with respect to their influence on triggering window opening actions: time of the day, CO_2 concentration, indoor and outdoor temperature, and indoor relative humidity.

(B) Correlation analysis was performed to identify strong correlations between dominant variables that impact window opening behaviour. Adding correlations between the variables in the BN model by linking them with arcs did not increase the BIC score as it increases the model complexity but does not substantially increase the predictive power of the BN model.

(C) This study showed that the window opening action is more suitable as a target variable to model window control behaviour than the window open/close state. Indoor environment variables such as indoor CO_2 concentration level and indoor temperature were identified as key variables that change the window state, but at the same time, the indoor environment conditions are directly influenced immediately after a window control action takes place. Hence, when the window state was used as a target variable, the statistical model with using indoor environment variables as predictors did not correctly represent relationships between the indoor variables and window control behaviour.

(D) The study demonstrated the most BN models used for only discrete or continuous datasets are not suited to fully exploiting information embedded in the mixed dataset. A reversed BN model was proposed to appropriately handle mixed data in the bnlearn environment. The proposed model was structured to predict the probability of a window opening action given the identified key environmental and time variables. In line with existing studies and the K-S two sample test results, arc strengths in the BN model also indicated that the time of the day, CO_2 concentration and indoor/outdoor temperature are the most important variables.

(E) The BN model was validated in terms of the expected loss value and the confusion matrix through the classical cross-validation procedure. As the data points with WOAs are much smaller than those with NOAs, a tailored validation approach was applied to select the same number of data points for each case and compute the confusion matrix. The validation measures confirmed the high predictive power of the model and its successful application for modelling window control behaviour.

In summary, Bayesian network modelling well represents the stochastic nature of window control behaviour in relation to a variety of explanatory variables and consequently provides predictions with high confidence. However, steps involved in the modelling process, specifically variable selection and validation, need to be carefully set up to correctly reflect the stochastic nature in the analysis process.

7. Acknowledgements

The authors would like to thank Dr. Rune Korsholm Andersen, Prof. Jørn Toftum, and Prof. Bjarne Olesen from the Denmark Technical University (ICIEE-International Centre for Indoor Environment and Energy) for sharing the measurements of the residential apartment located in Copenhagen, Denmark. The authors are also very grateful for all the precious comments and suggestions given by Dr. Marco Scutari from Oxford

University (Department of Statistics) as regards the implementation of the BN models in the R software package bnlearn.

8. References

[1] O. T. Masoso and L. J. Grobler, "The dark side of occupants' behaviour on building energy use," *Energy Build.*, vol. 42, no. 2, pp. 173–177, 2010.

[2] K. B. Janda, "Buildings don't use energy: people do," *Archit. Sci. Rev.*, vol. 54, no. 1, pp. 15–22, Feb. 2011.

[3] C. M. Clevenger, J. R. Haymaker, and M. Jalili, "Demonstrating the Impact of the Occupant on Building Performance," *J. Comput. Civ. Eng.*, vol. 28, no. 1, pp. 99–102, 2014.

[4] V. Fabi, R. Andersen, and S. Corgnati, "Description of occupant behaviour in building energy simulation: state-of-art and concepts for improvements," *Build. Simul. 2011 12th Conf. Int. Build. Perform. Simul. Assoc. Sydney, 14-16 Novemb.*, pp. 2882–2889, 2011.

[5] D. Yan *et al.*, "Occupant behavior modeling for building performance simulation: Current state and future challenges," *Energy Build.*, vol. 107, pp. 264–278, 2015.

[6] M. Schweiker, "Understanding Occupants' Behaviour for Energy Efficiency in Buildings," *Curr. Sustain. Energy Reports*, pp. 1–7, 2017.

[7] A. Mahdavi, "The human dimension of building performance simulation," in *Proceedings of Building Simulation 2011: 12th Conference of International Building Performance Simulation Association*, 2011, pp. K16–K33.

[8] T. Hong, S. C. Taylor-Lange, S. D'Oca, D. Yan, and S. P. Corgnati, "Advances in research and applications of energy-related occupant behavior in buildings," *Energy Build.*, vol. 116, pp. 694–702, 2016.

[9] I. Gaetani, P. J. Hoes, and J. L. M. Hensen, "Occupant behavior in building energy simulation: Towards a fit-for-purpose modeling strategy," *Energy Build.*, vol. 121, pp. 188–204, 2016.

[10] J. F. Nicol, "Characterising occupant behavior in buildings: Towards a stochastic model of occupant use of windows, lights, blinds heaters and fans," *Seventh Int. IBPSA Conf.*, pp. 1073–1078, 2001.

[11] R. Andersen, V. Fabi, J. Toftum, S. P. Corgnati, and B. W. Olesen, "Window opening behaviour modelled from measurements in Danish dwellings," *Build. Environ.*, vol. 69, pp. 101–113, 2013.

[12] H. B. Rijal, P. G. Tuohy, J. F. Nicol, M. A. Humphreys, A. a. a. Samuel, and J. a. Clarke, "Development of an adaptive window-opening algorithm to predict the thermal comfort, energy use and overheating in buildings," *J. Build. Eng.*, vol. 1, no. 1, pp. 17–30, 2008.

[13] S. D'Oca, V. Fabi, S. P. Corgnati, and R. K. Andersen, "Effect of thermostat and window opening occupant behavior models on energy use in homes," *Build. Simul.*, vol. 7, no. 6, pp. 683–694, 2014.

[14] G. Y. Yun and K. Steemers, "Time-dependent occupant behaviour models of window control in summer," *Build. Environ.*, vol. 43, no. 9, pp. 1471–1482, 2008.

[15] Y. Zhang and P. Barrett, "Factors influencing the occupants' window opening behaviour in a naturally ventilated office building," *Build. Environ.*, vol. 50, pp. 125–134, 2012.

[16] F. Haldi and D. Robinson, "Interactions with window openings by office occupants," *Build. Environ.*, vol. 44, no. 12, pp. 2378–2395, 2009.

[17] F. Haldi and D. Robinson, "A comparison of alternative approaches for the modelling of window opening and closing behaviour," *Air Cond. Low Carbon Cool. Chall.*, no. July, pp. 27–29, 2008.

[18] R. Fritsch, A. Kohler, M. Nygård-Ferguson, and J.-L. Scartezzini, “A stochastic model of user behaviour regarding ventilation,” *Build. Environ.*, vol. 25, no. 2, pp. 173–181, 1990.

[19] M. Schweiker, F. Haldi, M. Shukuya, and D. Robinson, “Verification of stochastic models of window opening behaviour for residential buildings,” *J. Build. Perform. Simul.*, vol. 5, no. January 2015, pp. 55–74, 2012.

[20] D. Heckerman, “A Tutorial on Learning with Bayesian Networks,” in *Innovations in Bayesian Networks*, vol. 1995, no. November, 2008, pp. 33–82.

[21] D. Margaritis, S. Thrun, C. Faloutsos, A. W. Moore, and G. F. Cooper, “Learning Bayesian Network Model Structure from Data,” *Learning*, no. May, 2003.

[22] D. Zhao and C. Weng, “Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction,” *J. Biomed. Inform.*, vol. 44, no. 5, pp. 859–868, 2011.

[23] D. B. Matellini, A. D. Wall, I. D. Jenkinson, J. Wang, and R. Pritchard, “Modelling dwelling fire development and occupancy escape using Bayesian network,” *Reliab. Eng. Syst. Saf.*, vol. 114, no. 1, pp. 75–91, 2013.

[24] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, “Dynamic Bayesian Networks for Audio-Visual Speech Recognition,” *EURASIP J. Adv. Signal Process.*, vol. 2002, no. 11, pp. 1274–1288, 2002.

[25] K. L. Jensen, J. Toftum, and P. Friis-Hansen, “A Bayesian Network approach to the evaluation of building design and its consequences for employee performance and operational costs,” *Build. Environ.*, vol. 44, no. 3, pp. 456–462, 2009.

[26] S. Wu and D. Clements-croome, “Estimating the relationship among occupant behaviours and indoor environmental parameters using Bayesian networks,” in *Proceedings of Clima 2007 WellBeing Indoors*, 2007.

[27] R. H. Dodier, G. P. Henze, D. K. Tiller, and X. Guo, “Building occupancy detection through sensor belief networks,” *Energy Build.*, vol. 38, no. 9, pp. 1033–1043, 2006.

[28] J. Petzold, A. Pietzowski, F. Bagci, W. Trumler, and T. Ungerer, “Prediction of indoor movements using bayesian networks,” *Locat. Context.*, pp. 211–222, 2005.

[29] C. Harris and V. Cahill, “Exploiting user behaviour for context-aware power management,” *2005 IEEE Int. Conf. Wirel. Mob. Comput. Netw. Commun. WiMob’2005*, vol. 4, pp. 122–130, 2005.

[30] L. Hawarah, S. Ploix, and M. Jacomino, “User Behavior Prediction in Energy Consumption in Housing Using Bayesian Networks,” *Artif. Intell. Soft Comput.*, pp. 372–379, 2010.

[31] K. Tijani, Q. D. Ngo, S. Ploix, B. Haas, and J. Dugdale, “Towards a general framework for an observation and knowledge based model of occupant behaviour in office buildings,” *Energy Procedia*, vol. 78, pp. 609–614, 2015.

[32] K. B. Korb and A. E. Nicholson, *Bayesian Artificial Intelligence, Second Edition*. CRC Press, 2010.

[33] P. Mylly Aki, T. Silander, H. Tirri, and P. Uronen, “B-COURSE: A WEB-BASED TOOL FOR BAYESIAN AND CAUSAL DATA ANALYSIS,” *Int. J. Artif. Intell. Tools*, vol. 11, no. 3, pp. 369–387, 2002.

[34] S. L. Lauritzen, “The EM algorithm for graphical association models with missing data,” *Comput. Stat. Data Anal.*, vol. 19, no. 2, pp. 191–201, 1995.

[35] M. Scutari, “Learning Bayesian Networks with the bnlearn R Package,” *J. Stat. Softw.*, vol. 35, no. 3, pp. 1–22, 2010.

[36] M. Scutari, “Bayesian Networks in R an example in system biology,” *Springer*, p. 157, 2013.

[37] K. B. Korb and Ann E. Nicholson, “Introducing Bayesian Networks,” *Bayesian Artif. Intell.*, pp. 29–54, 2003.

[38] M. Vlachopoulou, G. Chin, J. C. Fuller, S. Lu, and K. Kalsi, “Model for Aggregated Water Heater Load Using Dynamic Bayesian Networks,” *World Congr. Comput. Sci. Comput. Eng. Appl. Comput.*, 2012.

[39] R. Nanda, S. Saguna, K. Mitra, and C. Åhlund, “BayesForSG: A Bayesian Model for Forecasting Thermal Load in Smart Grids,” *Proc. 31st Annu. ACM Symp. Appl. Comput.*, pp. 2135–2141, 2016.

[40] D. Shipworth, “Modelling home internal temperatures using Bayesian networks,” *Model. Soc. Sci. Interdiscip. Comp.*, pp. 1–14, 2010.

[41] D. Margaritis, S. Thrun, C. Faloutsos, A. W. Moore, and G. F. Cooper, “Learning Bayesian Network Model Structure from Data,” *Learning*, no. May, 2003.

[42] J. Pearl and T. S. Verma, “A theory of inferred causation,” *Stud. Log. Found. Math.*, vol. 134, no. C, pp. 789–811, 1995.

[43] A. L. Madsen, M. Lang, U. B. Kjærulff, and F. Jensen, “The Hugin Tool for Learning Bayesian Networks,” *Learning*, pp. 594–605, 2003.

[44] S. Conrady and L. Jouffe, “Introduction to Bayesian Networks & BayesiaLab,” *BayesiaLab*, p. 30, 2013.

[45] “Norsys - Netica Application.” [Online]. Available: https://www.norsys.com/netica.html. [Accessed: 22-Jul-2017].

[46] K. P. Murphy, “The Bayes Net Toolbox for Matlab,” *Comput. Sci. Stat.*, vol. 33, no. 2, pp. 1024–1034, 2001.

[47] W. H. Hsu, B. B. Perry, and J. A. Thornton, “Bayesian Network Tools in Java (BNJ) What is BNJ ?,” *Network*, pp. 1–11.

[48] “BNfinder 2.1.1 : Python Package Index.” [Online]. Available: https://pypi.python.org/pypi/BNfinder/2.1.1. [Accessed: 12-Aug-2017].

[49] F. V. Jensen and T. D. Nielsen, *Bayesian Networks and Decision Graphs*, vol. 44, no. 8. 2007.

[50] R. E. Neapolitan, *Learning Bayesian Networks*, vol. 6, no. 2. 2003.

[51] D. Barber, “Bayesian Reasoning and Machine Learning,” *Mach. Learn.*, p. 646, 2011.

[52] S. Borgeson and G. Brager, “Occupant Control of Windows: Account for Human Behavior in Building Simulation,” 2008.

[53] H. B. Rijal, P. G. Tuohy, J. F. Nicol, M. a. Humphreys, A. a. a. Samuel, and J. a. Clarke, “Development of adaptive algorithms for the operation of windows, fans and doors to predict thermal comfort and energy use in Pakistani buildings,” *ASHRAE Trans.*, vol. 114, no. 2, pp. 555–573, 2008.

[54] P. R. Warren and L. M. Parkins, “Window-opening behaviour in office buildings,” *Build. Serv. Eng. Res. Technol.*, vol. 5, no. 3, pp. 89–101, 1984.

[55] J. Pfafferott and S. Herkel, “Statistical simulation of user behaviour in low-energy office buildings,” *Sol. Energy*, vol. 81, no. 5, pp. 676–682, 2007.

[56] W. J. Conover, “Practical Nonparametric Statistics.,” *The Statistician*, vol. 22. pp. 309–314, 1971.

[57] H. Abdi, “The Kendall Rank Correlation Coefficient,” *Encycl. Meas. Stat.*, pp. 508–510, 2007.

[58] J. Suzuki, “Learning Bayesian Network Structures When Discrete and Continuous Variables Are Present,” in *Probabilistic Graphical Models*, 2014, no. 5, pp. 471–486.

[59] N. Dojer, “Learning Bayesian networks from datasets joining continuous and discrete variables,” *Int.*

J. Approx. Reason., vol. 78, pp. 116–124, 2016.

[60] V. Fabi, R. V. Andersen, S. Corgnati, and B. W. Olesen, "Occupants' window opening behaviour: A literature review of factors influencing occupant behaviour and models," *Build. Environ.*, vol. 58, pp. 188–198, 2012.

[61] M. Scutari and D. Jean-Baptiste, *Bayesian Networks: With Examples in R - CRC Press Book*. 2014.

Highlights:

- The applicability of the Bayesian Network Framework to model window control behaviour is demonstrated
- A procedure for developing a BN model with full exploitation of mixed data is presented
- Key variables that mostly impact window control actions are highlighted
- The window opening action is more suitable as a target variable to model window control behaviour

Department of Energy
Polytechnic of Turin
Corso Duca degli Abruzzi, 24
10129 Torino, Italy
Tel +39 346 72 90 686
verena.barthelmes@polito.it
www.polito.it