Title
Generating synthetic occupants for use in building performance simulation

Journal
Journal of Building Performance Simulation, 14(6)

ISSN
1940-1493

Authors
Putra, Handi Chandra
Andrews, Clinton
Hong, Tianzhen

Publication Date
2021-11-02

DOI
10.1080/19401493.2021.2000029

Peer reviewed

Generating Synthetic Occupants for Use in Building Performance Simulation

Handi Chandra Putra[1], Clinton Andrews[2], Tianzhen Hong[1*]
[1]Building Technology and Urban Systems Division, Lawrence Berkeley National Laboratory
[2]Rutgers Center for Green Building, Rutgers University
*Corresponding author: T. Hong, thong@lbl.gov

Abstract

Occupant behavior simulation frameworks can employ synthetic populations to characterize occupancy and behavioral patterns in buildings based on real demographic data at a certain geographical location. Multiple methods are available to generate a synthetic population, with pros- and cons- for different applications and contexts. For buildings, very few synthetic occupant populations have been generated. This paper uses a Bayesian Networks (BN) structural learning approach to synthesize populations of occupants in a multi-family housing. Two additional cases of office occupants and senior housing residents are considered as a cross-case comparison. We draw upon the extended version of Drivers-Needs-Actions-Systems (DNAS) framework to guide the selection of variables and data imputation. The resulting synthetic occupant data is evaluated by comparing the joint distributions between the actual and synthetic data sets, % difference, and Standardized-Root-Mean-Squared-Error (SRMSE). Our results show that the BN approach is powerful in learning the structure of data sets. The synthetic data sets successfully match the joint distributions of the underlying combined data sets. Experiments on the multi-family housing particularly show the best performance compared to the cases of office and senior housing. Future work includes testing and demonstration of how the synthetic data sets feed to occupant behavior module through co-simulation with building performance simulation tools, such as EnergyPlus.

Keywords: synthetic occupants, occupant behavior, building performance simulation, occupant behavior model, data model

1. Introduction

Research in building and occupant interactions is now using data-driven modeling approaches to represent more realistic occupancy and behaviors in buildings (Berger & Mahdavi, 2021). Whereas conventional models tend to use standard occupancy schedules or simple rule-based behavioral models, the trend is developing towards richer data processing. In connection with the rapid development of machine learning (ML) and agent-based modeling (ABM) that fall under the data-driven modeling category, researchers face the daunting task of collecting and processing a large amount of data. The high cost, the issues of availability and user privacy have

hindered the research progress. One resort to solve this problem is by generating synthetic population based on existing data.

Synthetic occupant population has an advantage because it can directly feed to ABM for modeling and simulation of occupant behavior and their impacts on building performance. Modelers can explore the subset of occupant behavior topics that are most relevant to building performance simulation. For example, modeling occupant behaviors for warmer climates may be different from those for colder climates, for which the synthetic data sets can be drawn from the socio-demographic data sets from U.S. Census (2015) that are useful in addition to the previously collected comfort data sets (Licina et al., 2018). Another advantage is that synthetic occupant population enables the modeling of under-represented occupant population, such as children and seniors. Residential building, school, and senior housing are building types with diverse age, gender, and other socio-demographic characteristics attributed to the occupants. This paper introduces the population synthesis and applications in the OB modeling research. A Bayesian Network (BN) model is proposed to generate the synthetic building occupant population. As an illustration, we consider three case studies of a multi-family housing residents, office occupants, and senior housing residents.

The paper is structured as follows: In Section 2, we describe how population synthesis can close some of the gaps in OB modeling research. Section 3 reviews previous research on population synthesis methodologies. In section 4, we describe the proposed synthetic population generation process. We also describe the Bayesian Network (BN) structure learning as our choice to guide the population synthesis. We use the resulting structures to represent the underlying joint distribution for generating synthetic data across the selected cases of buildings that leads to the next section. Section 5 discusses the applications of synthetic population to three case experiments. In Section 6, we further discuss the results, limitations of the approach and its potential improvements. The paper ends with take away points relevant to the population synthesis and future research in Section 7.

2. Research gaps

This section identifies three major gaps in OB modeling research that synthetic occupant populations can help to close. The first is from the modeling standpoint, that ABM-based modeling emerges to open new research opportunities in exploring patterns of occupancy and behaviors at a time-step level during simulation (Berger & Mahdavi, 2021). Like many other OB models, ABM-based OB models are principally used as predictive tools in which occupant patterns of presence and behavior interact with building systems in using energy or other resources (Gaetani et al., 2016). ABM simulates each agent's behavior and activity patterns over time. Essentially, ABM models require OB data to characterize the initial sets of agents and a group of agents in the model. Population synthesis enable the generation of occupant-agents with characteristics necessary for the modeling purposes. Henceforth, the characteristics of each agent emulate realistic patterns of building occupant composition.

The problem in modeling occupant-agents leads to the second gap, which is the various explanations of building-occupant interaction. Theories behind the the occupant behaviors imply the building as a complex social system within which occupants interact with one another and with multiple building systems (Heydarian et al., 2020). The comfort preferences may be one factor determining the occupant behavior and interaction in buildings, socioeconomic and demographic factors are two other important factors that pertain to specific geographical locations (Heydarian et al., 2020). For example, office occupants in the U.S. may have different behaviors from those in Singapore. Households in multi-family housing in Buffalo, NY, may have different operational patterns from residents in Texas. A simplified representation of an entire population of the specific geographical area of interest helps to explain its complexity. A synthetic occupant population represents real building occupants using a limited set of characteristics selected on the basis of context and interests. One important note is that the synthetic occupants should have a matching statistical measures of the actual population rather than trying to replicate each individual in the actual population.

Finally, methods for generating synthetic populations based on empirical data at certain geographical locations have been developed in fields of demography, urban planning and transportation. Despite various applications, there are, however, no found studies or applications of synthetic occupant population applied to building peeerformance simulation as suggested in (Andrews et al., 2016). There are several comprehensive reviews on various methods for population modeling for applications in other fields (Barthelemy & Toint, 2013; Chapuis & Taillandier, 2019; E. Ramadan & P. Sisiopiku, 2020). Rather than attempting to reinvent these efforts, this paper focuses specifically on the application of one of the known methods to OB modeling research using the established Bayesian Network structure learning. Therefore, the purpose of this study is to introduce a framework that generates synthetic data sets that can feed to occupant behavior models and coupled with building performance simulation tools.

3. Literature review

We have identified three gaps related to the building OB modeling research. While this paper sees the gaps as not mutually exclusive, there have been a number of previous studies that have addressed specific gaps. Previous efforts have done extensive studies on the first two topics related to the advancement of data-driven models requiring amounts of data and the development of occupant data ontology. The third gap on the application of population synthesis in the OB modeling research has not been addressed.

3.1. Data-driven models

Within the context of emerging data-driven models in OB modeling research, the need for more elaborate data has been given particular attention. The ABM approach, for example, is known to require much data on individual agents to characterize the initial number of occupant-agents in the simulation. The agents are expected to evolve throughout the simulation process, showing the behavioral and interaction patterns between the individual occupant-agents, within agent-groups, and with the building systems. Paired with other modeling tools, such as EnergyPlus, ABM is a

powerful tool that is useful to produce stochastic OB models with behavioral patterns and interactions that are close to real building occupants (Chapman et al., 2018). The explorative nature of ABM has distinguished the modeling approach among the existing OB models (Berger, 2020; Bonabeau, 2002; Chandra-Putra et al., 2017).

ML based OB models also fall under the data-driven model category. The models have shown promising results according to (Markovic et al., 2017). ML-based models have recently been employed in experiments on various occupant behaviors that are relevant to building operations, including lighting operation (Nagy et al., 2016), thermostat adjustment (Peng et al., 2019), and appliance use (Ueda et al., 2015; Wang & Ding, 2015). It is expected that there will be more applications of ML-based models to the OB modeling research, which will also be put in comparison with other more traditional models, such as rule-based and stochastic models (Carlucci et al., 2020).

The traditional ML-based OB models require substantial amounts of data that split into training and evaluation data sets. Markovic (2018) points out that data imbalance affects the prediction accuracy and training complexity of using the approach. A more obvious limitation of data-driven models, as discussed in (Berger & Mahdavi, 2021), is the lack of data, which disables higher-resolution models, hence yielding inaccurate depictions of occupant behavioral patterns. One solution to address the data issue is by investing in the data collection process. The rise of Internet-of-Things (IOT) through sensors and installation produces a more granular and high-quality wealth of data (Carlucci et al., 2020).

3.2. Occupant data ontology

The development of an occupant data framework grows in parallel to the inceasing need to collect more occupant data. Researchers develop ontologies to guide their data collection and assembly process. A data ontology is useful in defining the specific applications and use cases within occupant behavior research (Salimi & Hammad, 2019). The development of an occupant data ontology to date seek to incorporate the diversity in traits and attributes of real building occupants (O'Brien et al., 2016). Some of the common features found in previously developed ontologies are those related to occupant comfort with regards to behavioral comfort-driven responses of occupant within the built environment. The Drivers-Needs-Actions-Systems (DNAS) framework environment (Hong, D'Oca, Taylor-Lange, et al., 2015; Hong, D'Oca, Turner, et al., 2015) includes actions such as thermostat adjustment, lighting adjustment, fans and portable heater operation, and moving between spaces. Other frameworks include detailed characteristics of occupants such as socio-conomic (Andrews, 2017; Kontokosta & Jain, 2015; Tsoulou et al., 2020), subjective values (Hong et al., 2020; Ortiz & Bluyssen, 2018), and activities (Hewitt et al., 2016; Malik & Bardhan, 2020). The recent update on the DNAS framework adds more components relevant to static attributes of building occupants (Chandra-Putra et al., 2021). Attributes, such as geographical location, socio-economic, and subjective values are considered as traits that are unique to occupants and determine their perceptions and adaptive behaviors in the built environment.

3.3. Population synthesis methodology

Finally, this sub-section discusses the largest portion in the literature review section. The role of population synthesis in OB modeling research is to provide a simplified representation of real occupants at a certain geographical location. Sun & Erath (2015) describe three challenges in generating such geographical-based populations. The first challenge is in preserving the dependency structure and matching the aggregate data and avoiding potential biases. At the individual level of occupants, correlations may be stronger between age and income than between age and sex, for example. At the group level, tenant units, for example, are associated with the number of occupants. The second challenge is in associating group-level attributes with individual-level attributes in a unified manner. For example, shared thermostat adjustment (as a group attribute) is strongly related to whether any of the individual occupants have access to a thermostat. The third and final challenge is in reproducing the interdependencies among occupant-agents in the same group. Using the thermostat example, where two individual tenants share the same thermostat interface, such type of structural relationship may not be reported in the occupant survey data.

In general, synthesizing a population involves two steps – fitting and generation (Kirill Müller & Axhausen, 2010). "Fitting" usually refers to estimating the joint distribution of all included attributes from the disaggregated data and marginal distributions of a target dataset. The "Generation" step is to create a new set of data by drawing from the fitted distribution of the earlier step. Being the more important step, fitting techniques continually advance to produce distributions that are ever more realistic and representative of the actual population distribution. The literature has recorded these developments, which can be categorized into three different approaches (Chapuis & Taillandier, 2019; Sun & Erath, 2015). Synthetic Reconstruction (SR) creates a vector of characteristics of an agent in the agent simulation; Combinatorial Optimization (CO) duplicates known individual attributes of real observations; and Statistical Learning (SL) considers complex interactions and synthesizes multivariate real data.

The first two approaches to the fitting problem, SR and CO, are popular in public health to study the health behaviors and spread of virulent outbreaks (Cooley et al., 2008; Smith et al., 2011; Tomintz et al., 2008). Cooley et al. (2008) generates a synthetic population and feeds it into an agent-based model (ABM) to project the spread of influenza in North Carolina. The methods are also popular in transportation (Beckman et al., 1996; E. Ramadan & P. Sisiopiku, 2020; McFadden et al., 1977; Xie & Waller, 2010) , urban water management (Williamson, 2012), disaster evacuation (Jumadi et al., 2018), crime (Malleson et al., 2010) and urban social dynamics (Malleson & Birkin, 2013). To date, there is relatively less application in building simulation research, particularly in the sub-discipline of occupant presence and behaviors (Andrews et al., 2016; Chen et al., 2021).

In the SR domain, Iterative Proportional Fitting (IPF) is one of the most widely-used distribution estimation algorithms that fits individual cells within the distribution using known marginal controls from the sample. It, however, cannot deal with multi-level populations, i.e., households and individuals. Improved versions are called Hierarchical IPF and Iterative Proportional

Updating (IPU) (K (K Müller, 2017; Sun et al., 2018). In contrast, CO-based population synthesis relies on the drawing of individuals from a sample to assess a fitness criterion. Simulated Annealing, Hill Climbing, Genetic Algorithms, and Greedy Heuristics are known optimization algorithms used in population synthesis (Harland et al., 2012). One major issue is that CO relies on replicating individuals in the population rather than random sampling as a requirement for true synthesis (Farooq et al., 2013). Another issue is that they do not consider multi-level attributes, e.g., households and individuals. Modeling the multi-level associations is usually done by first sampling a group-level and then gathering individuals to fill the group-level (Barthelemy & Toint, 2013; Pritchard & Miller, 2012).

Statistical Learning (SL) offers more flexibility in terms of data requirements since it focuses on the joint distribution instead of the samples' replication. The developments can be traced back to (Reiter, 2005), whose work introduces an imputation method using classification and regression tree, CART. Other methods are random forest (RF) (Caiola & Reiter, 2010), Markov Chain Monte Carlo (MCMC) (Farooq et al., 2013), and Hidden Markov model (HMM) (Saadi et al., 2016). These models also do not consider multi-level population synthesis. Sun & Erath (2015) solves this problem by using a Bayesian Networks (BN) method that is also used in this study, by modeling the interdependencies among household and individual attributes (Koller & Friedman, 2013).

Sun & Erath (2015) suggests that their BN model translates complex relationships within the network structure into a simple graphical model that is called Directed Acylic Graph (DAG). Some common BN structure learning algoritms can be grouped into three categories, which are constraint-based, score-based, and hybrid as described in Table 1. While constraint-based algorithms ensure that conditional independence constraints are met using statistical tests, score-based algorithms rely on optimization techniques in which each candidate DAG is assigned a score as the objective function. Hybrid algorithms integrate the two methods by reducing the space of candidate DAGs using a constraint-based strategy and implement a score-based strategy to find the optimal DAG in the constrained space. Constraint-based algorithms seem considerably more accurate than score-based algorithms for small sample sizes and they both are as accurate as hybrid ones.

Table 1: Types of Bayesian Networks (BN) Structural Learning Algorithms.

Categories	**Algorithms**	**Description**	**Literature**
Constraint-based	Prototypical Constraint (PC)	An applicaton of an Inductive Causation (IC) algorithm and independent test for network structure learning.	(Spirtes et al., 2012)
	Hiton-PC	A variant of PC algorithm that removes false-positives in the post-processing.	(Aliferis et al., 2010)
	Grow-Shrink (GS)	A forward selection using Markov blanket detection approach.	(Margaritis, 2003)
	Incremental Association (IAMB)	A two-phase selection scheme using Markov blanket approach.	(Tsamardinos et al., 2003)

	Fast IAMB	A variant of IAMB which uses speculative stepwise forward selection to reduce the number of conditional independence tests	(Yaramakala & Margaritis, 2005)
	Interleaved IAMB	A variant of IAMB which uses forward stepwise selection to avoid false positives in the Markov blanket detection phase	(Tsamardinos et al., 2003)
Score-based	Greedy search algorithms (e.g. Hill-Climbing (HC) and Tabu Search)	A greedy search in the network structure using scores. Tabu Search is an iterative greedy search that uses a memory structure Tabu list to escape local optima.	(Bouckaert, 1995)
	Genetic Algorithms	An iterative selection approach to look for the fittest model.	(Larrañaga et al., 1997)
	Simulated Annealing	A stochastic search approach to find globally-minimum-cost solution iteratively.	(Bouckaert, 1995)
Hybrid	Sparse Candidate (SC)	An algorithm that restricts and maximizes the network structure iteratively until it reaches a stable network score.	(Friedman & Koller, 2003)
	Max-Min Hill Climing (MMHC)	An algorithm to perform restrict and maximise only once.	(Tsamardinos et al., 2006)
	Max-Min Parents and Children (MMPC)	A heuristic application of restrict and maximize algorithms.	(Tsamardinos et al., 2006)
	Aracne and Chow-Liu	Algorithms learn graph structures using pairwise mutual information coefficients.	(Chow & Liu, 1968; Margolin et al., 2006)

Success stories of modeling applications in transportation research have inspired us to use BN models in building occupant population synthesis (Castillo et al., 2008; Zhang & Taylor, 2006). Some of the applications that inspire this study include modal choice behavior using household travel surveys (Xie & Waller, 2010) and agent-based activity simulation and prediction (Janssens et al., 2006). The application of BN in the population synthesis for building occupant behaviors utilizes its ability to learn the structure of the occupant data, particularly when the number of attributes of interest is large using socio-economic and demographic microdata and behavior data. Moreover, the BN approach infers the multivariate probability distribution of both data at the household and individual-level attributes. Similar to the Markov process-based approach, the BN approach requires neither marginals as input not any conditionals to inform the structural learning. The learning model integrates the parameter estimation. Sun & Erath 2015 demonstrate a good performing BN approach in synthesizing the 2010 household interview travel survey of Singapore with low SRMSE values and good heterogeneity.

4. Modeling framework

The modeling framework for the synthetic occupant population generation is quite straightforward. This section discusses each framework component starting from combining the original data sets, followed by variables selection using an existing data ontology and data imputation procedure. The generation of the synthetic version of the data sets uses Bayesian Network (BN) structure learning. Then, the synthetic model is calibrated by generating additional synthetic occupants for two confirmatory cases (e.g., office occupants and senior housing occupants). Finally, the model is validated by looking at the % difference, joint distribution, and Standardized-Root-Mean-Square-Error (SRMSE). Figure 1 illustrates the overall population synthesis workflow used in the study.

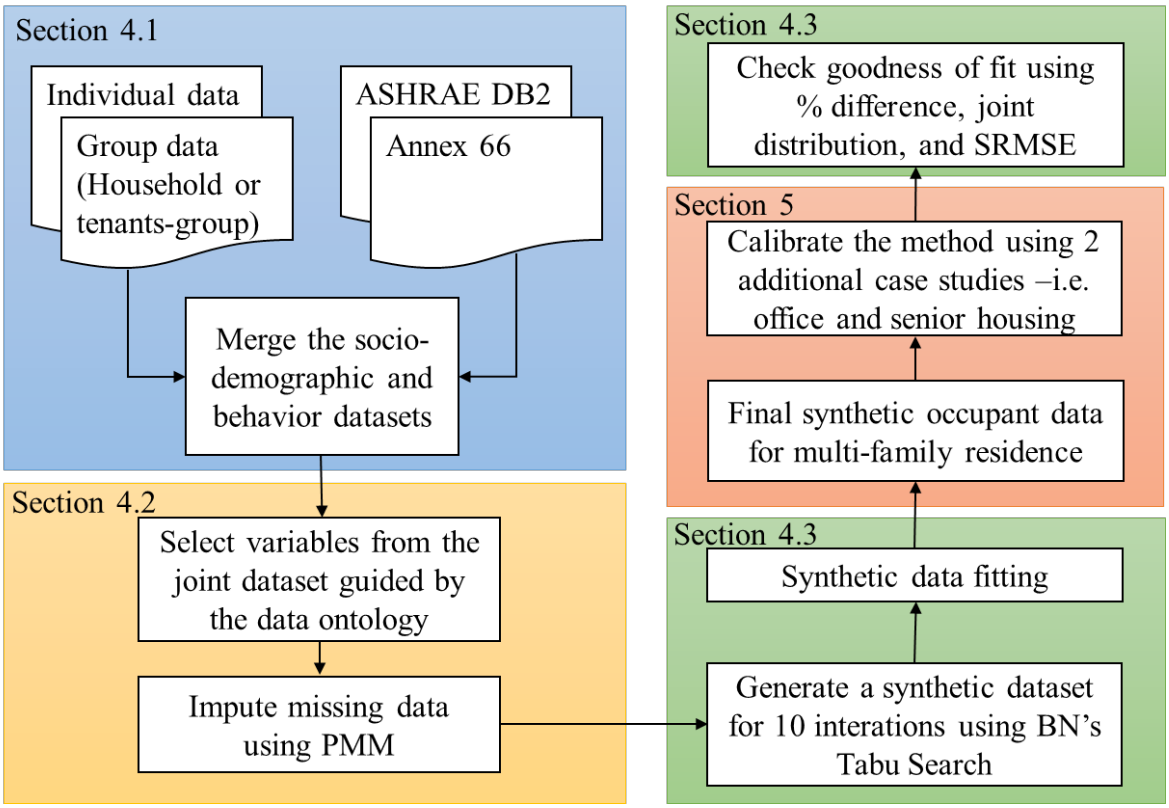


Figure 1: The population synthesis workflow.

4.1. Dataset overview

We consider two categories of data sets to inform the generation of our synthetic occupant population. The socio-demographic data is cross-sectional and attributed to the individual and group populations at a particular geographical location. The data set draws from two sources, including the National Household Travel Survey (NHTS) (Administration, 2017) and Public Use Microdata Survey (PUMS) (U.S. Census, 2015). The occupant behavior data set draws on a dataset on occupant behavior that was part of the deliverables of an international project, Annex 66. Annex 66 was established under the International Energy Agency's Energy in Buildings and

Communities Program (short name: IEA Annex 66) with aims to provide resources for occupant behavior research (International Energy Agency, 2017). The dataset includes of 4,324 observations. The larger and more recent ASHRAE Global Thermal Comfort Database II data set has been introduced in (Licina et al., 2018) and it includes approximately 81,846 occupant-specific data points spread across 160 buildings worldwide between 1995 and 2016. Similar to the socio-demographic data set, the two occupant behavior data sets are cross-sectional and collected with a sole purpose to support the building occupant behavior research. This study is interested in subsets of the dataset associated with specific building types.

These datasets are stored separately and need to be merged, fitted, and imputed as necessary. While this study considers data fusion using a simple left-join of 1-3 common variables, data imputation is implemented using Predictive Mean Matching (PMM) (Little & Rubin, 2002) PMM has been around for a long time, but only recently has it been widely used in population synthesis. It was originally used to impute missing data of a single variable, in which the missing data is more monotonic. Compared with standard methods based on linear regression and the normal distribution, PMM calculates the imputed values based on a set of values from the observed dataset, so they are often more realistic. Therefore, it allows one to impute discrete values, which is useful for the survey datasets used in the study.

4.2. Data ontology

We use an ontological approach in selecting the datasets in order to maintain a well-received building occupant data structure. Without the guiding data structure, population synthesis can require large amounts of computing power and result in a synthetic population not matched with the actual population. This study follows the existing Drivers, Needs, Actions, and Systems (DNAS) framework to provide a better representation of building occupants (see Figure 2). The framework has been recently updated with a more elaborate occupant characteristics that was drawn from previous multiple papers (Andrews, 2017; Chandra-Putra et al., 2021; Hewitt et al., 2016; Hong et al., 2020; Senick, 2015). As described in (Chandra-Putra et al., 2021), the extended DNAS framework categorizes the occupant characteristics into four groups, including socio-demographic, location, subjective values, and activities. The framework also develops further the comfort adaptive action components by dividing them into two sub-categories based on individual control or collective decisions. The socio-demographic component includes census-related information, such as “Age”, “Sex”, “Education”, “Income”, “Employment”, and “Marital Status”. Behavior-related data also includes attributes of subjective values that drive one to perform certain building energy-related actions. These are “Past Experience”, “Cost Conscious”, “Environment Awareness”, “Technology Oriented”, and “Social Influence”. With the many behavioral data sets available for population synthesis, a geographical location identifier is also important, which includes information such as “Country”, “Climate Zone”, “Policy”, and “Utility Cost”.

The extended DNAS framework is useful to inform the data structures of both metadata and dataset collection procedures. The researcher writes the metadata in either XML or JSON formats and the datasets in CSV, DAT, or other formats, depending on the data manipulation

methods. The ease of closely following the data structure depends on data availability and interoperability between datasets, and it is very common for population synthesis to also perform data imputation.

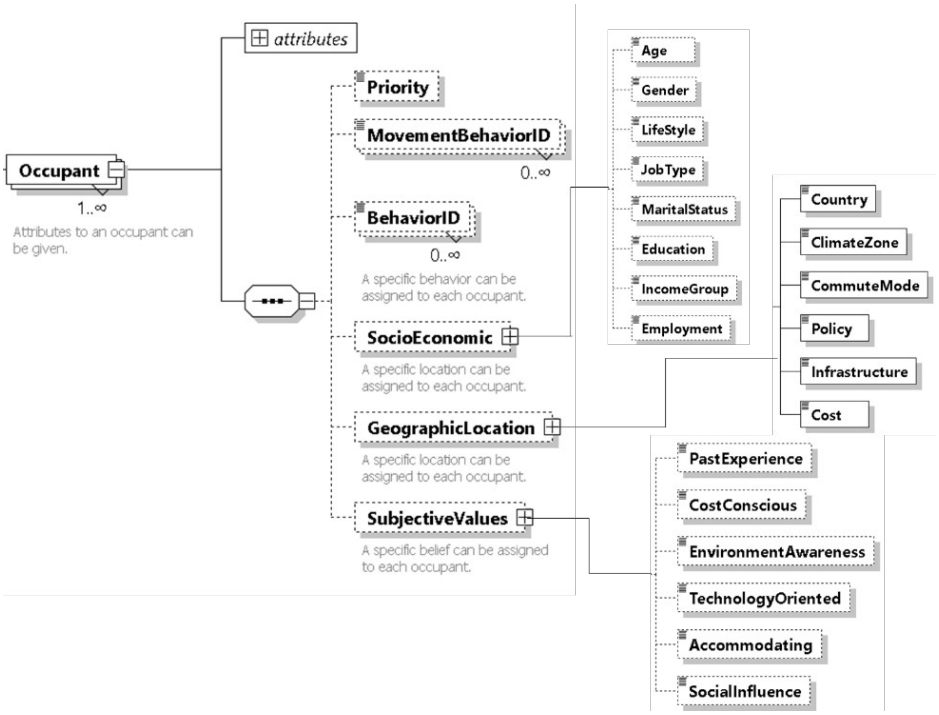


Figure 2: Occupant attributes, DNAS framework (Chandra-Putra et al., 2021).

4.3. Bayesian Networks

The Bayesian network (BN) offers a graphical representation of probability distributions for a set of variables of interest $X=\{X_1,\ldots,X_n\}$ (Heckerman et al., 1995; Koller & Friedman, 2013). In principle, a BN structure consists of two parts: 1) a network structure G in the form of a directed acyclic graph (DAG) (see Figure 2), in which vertices are random variables X and edges characterize the one-to-one mapping between the vertices, and 2) a set of local probability distributions $\Theta=\{P(X_1|\Pi_1),\ldots,P(X_n|\Pi_n)\}$ for each vertex X_i, conditional on its parents Π_i. In a BN, the DAG topology asserts only the conditional dependence of children given parents:

$$P(X)=\prod_{i=1}^{n} P(X_i|\Pi_i) \tag{1}$$

Figure 3 shows three variables including age, thermal perception, and income as the vertices and the directed edges linking vertex *Age* to vertex *ThermalPerception* and vertex *Age* to vertex *Income*. Therefore, the conditional probability distributions of this condition are P(*ThermalPerception*|*Age*) and P(*Income*|*Age*).

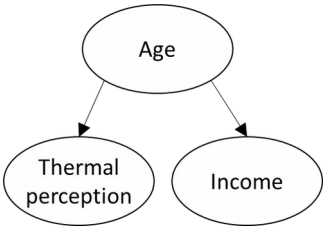


Figure 3: Example of a Directed Acyclic Graph (DAG)

The challenging part in BN is to determine the network structure. A researcher often defines the structure based on his or her domain-expert knowledge. This study has a particular interest to build the network structure directly from data, which is also called structural learning. Structure learning is a flexible feature of the package that identifies the relations and hierarchies of the variables. The R package, `bnlearn`, offers several algorithms to perform structure learning, including Tabu Search and Hill Climbing, which are described in Section 3. Prior to fusing multiple data sets, applying these algorithms results in a more robust estimated model structure.

Tabu Search is one of the widely used score-based algorithms that utilizes an iterative searching procedure to obtain a best solution from complex correlation patterns (Glover et al., 1993; Scutari, 2010). Tabu Search also selects a close solution to optimality in order to minimize the score. Bnlearn is also able to restrict the variables' dependencies by using its features of "whitelist" or "blacklist".

BN scores the candidate of the graphical structure that fits to the targeted data and is useful to produce synthetic population by using several methods, such as maximum likelihood:

$$l\left(G^h \middle| D\right) = \max_{G} \; \underset{\Theta}{¿}\, l\left(G, \Theta | D\right) = \max_{G} l\left(G, \hat{\Theta} \vee D\right) \tag{2}$$

Where $l\left(G, \Theta | D\right) = \log P\left(D \vee G, \Theta\right)$ is the log-likelihood of a provided pair $\left(G, \Theta\right)$ given observation D. (Sun & Erath, 2015) describe that the log-likelihood is not representative due to overfitting, hence, it always builds a fully connected DAG. Bayesian Information Criterion (BIC) (Rissanen, 1978; Schwarz, 1978) and Akaike Information Criterion (AIC) (Akaike, 1974; Rissanen, 1978) are the two most-popular score functions that solve this issue of overfitting:

$$BIC\left(G^h \middle| D\right) = logP\left(D \middle| G^h, \hat{\Theta}\right) - \frac{d}{2} \log m \tag{3}$$

$$AIC\left(G^h \middle| D\right) = logP\left(D \middle| G^h, \hat{\Theta}\right) - d \tag{4}$$

where $\hat{\Theta}$ refers to the maximum likelihood estimates of parameter given hypothetical structure G^h, d is the number of free parameters (degrees of freedom) in Θ, and m is the size of observation D. Both BIC and AIC contain two parts, the optimal likelihood and penalty that balances model fit and model complexity. While BIC has a penalty term of $\frac{d}{2} \log m$ and AIC only has d, the structure of BIC is more preferable for a large sample size.

To validate the resulting synthetic population generated by using BN, we adopt a metric measuring goodness-of-fit, which is the standardized root mean square errors (SRMSE) (Müller, 2017):

$$SRMSE = \frac{\sqrt{\frac{i}{I}\sum_{i}\left(¿\,\widehat{N}_i - N_i\right)^2}}{\frac{1}{I}\sum_{i} N_i}¿ \tag{5}$$

where, I is the number of categories of all attributes, and $\widehat{N}_i$ and N_i are totals of the population synthesized and actual population, respectively, for category $i \in [1, \ldots, k]$. The metric captures the relative frequency of each category of all attributes. A perfect fit between actual and synthetic population is indicated by a SRMSE value of zero, while a high value represents a bad fit.

5. Applications

This section demonstrates the performance of the proposed BN approach by conducting experiments on synthetic occupant population. Model inference and synthetic population generation are implemented in R.

5.1. Socio-demographic and behavior data

We select a multi-family housing as a case study. As described in section 4.1, this study draws on multiple data sources on population characteristics that are based on building type, location of the building, and socio-demographic as well as comfort behaviors of typical occupants. The case study attempts to demonstrate how geographically-specific data sets can be used and show how a population make-up of a certain location is different from a population of another area and that determines the way the building occupants behave differently in a built environment. Multi-family housing residents vary in age and activities. A typical household consists of adults and children. Occupants' adaptive actions towards their changing environments also vary by building type. Like in most multi-tenanted buildings, residents of the multi-family housing have hierarchical decisions in adjusting their environmental conditions based on the household composition. The case study considers 75 households in a multi-family housing with a varied number of members of each household with a total of 300 individual residents by synthesizing 2.8% of a total of 10,764 joint observations.

In addition to generating a synthetic occupant population for multi-family housing residents, we conduct additional experiments as confirmatory cases on occupants in an office building and a senior housing building. Similarly, we select the case of office buildings to show the diverse comfort-related behaviors since the topic has been extensively studied within the occupant behavior research. The final case of senior housing residents covers understudied population including seniors and children. We synthesize 27% of a total 1,858 joint observations (=500 occupants) for the case of office occupants and 15.48% of a total 646 joint observations (=100 occupants) for senior housing occupants (see Table 2).

The combined data sets consider all the variables from the original data sets and some of them may turn out unnecessary. We filter these variables with a guidance data ontology as described in Section 4.2. The resulting data sets only include 23 variables that are relevant for the population synthesis. Table 3 describes the variables in all three case studies with the variable names, definition, description of values for each ordinal variable, and data sources.

Table 2. Size of data sets for the population synthesis.

Building type	**Synthetic**	**Observed**	**Data sources**
Multi-family housing	300 (= 2.8%)	10,764	NHTS, ASHRAE, Annex66
Office building	500 (= 27%)	1,858	NHTS, ASHRAE, Annex66
Senior housing	100 (=15.48%)	646	PUMS, ASHRAE, Annex66

Table 3. Variable names and descriptions

Variable	**Definition**	**Values**	**Data sources**
sex	Gender of an occupant	Male; Female	NHTS, PUMS, ASHRAE, Annex66
age	Age of an occupant	multifamily housing <18,18-29,30-44,45-59,60-69,>70 office <29,29-39,40-50,51-61,>62 senior housing <65,65-74,75-85,>8	NHTS, PUMS, ASHRAE, Annex66
pmv	Predicted Mean Vote (PMV)	-3,-2,-1,0,1,2,3	ASHRAE, Annex66
marital_status	Marital status	1 = married, 2 = widowed, 3 = divorced, 4 = separated, 5 = never married or under 15 years old	NHTS
race	Ethnicity	1 = White, 2 = Black or African American, 3 = Asian, 4 = American Indian or Alaska Native, 5 = Native Hawaiian or other Pacific Islander	NHTS
relate	Relationship	1 = self, 2 = spouse/unmarried partner, 3 = child, 4 = parent, 5 = brother/sister, 6 = other relative, 7 = non-relative	NHTS
wkfpt	Work full time or part time	1 = full time, 2 = part time	NHTS
wrk_home	Working from home	1 = yes, 2 = no	NHTS
activity	Indoor activity	1 = working, 2 = temporarily absent from a job, 3 = unemployed, 4 = a homemaker, 5 = going to school, 6 = retired, 7 = other	NHTS

income	Household income	-7 = not answer/don't know, 1 = <$10,000, 2 = $10,000-$24,999, 3 = $25,000-$49,999, 4 = $50,000-$74,999, 5 = $75,000-$99,999, 6 = $100,000 - $124,999, 7 = > $125,000	NHTS, PUMS
hhsize	Household Size.	1 = 1, 2 = 2, 3 = 3, 4 = 4, 5 = 5, 6 = 6, 7 = 7, 8 = >8	NHTS
met	Metabolic activity (MET)	0.6-1,1-1.4,1.4-1.8,1.8-2.1	ASHRAE
clo	CLO	0.2-0.9,0.9-1.5,1.5-2.2,2.2-2.8	ASHRAE
airtemp	Air temperature in °C	16-21,21-25,25-30,30-34	ASHRAE
relhum	Relative humidity in %	25-36,36-47,47-58,58-69	ASHRAE
tpref	Preferred temperature	cooler, no change, warmer	ASHRAE
fan	Fan use	yes, no, don't have	ASHRAE
window	Window operation	yes, no, don't have	ASHRAE
heater	Heater use	yes, no, don't have	ASHRAE
educ	Educational attainment	highschool, assoc. degree, university graduate	Annex66
workspace	Workspace	enclosed space, open space, cubicle	Annex66
wrk_hrs	Working hours	in hours	Annex66
satisf_temp satisf_iaq satisf_light satisf_daylt satisf_artlt satisfy_sound	Satisfaction over indoor environmental quality (e.g. temperature, IAQ, natural lighting, daylighting, artificial lighting, acoustics)	unsatisfied, neutral, satisfied	Annex66
control_light control_windows control_blinds control_thermostat	Have control on fixture (e.g. lighting, windows, blinds, thermostat)	yes, no, I don't know	Annex66
group_light group_windows group_blinds group_thermostat	Group control on fixture (windows, blinds, lighting, thermostat)	only me, no control, with others	Annex66

5.2. Model selection

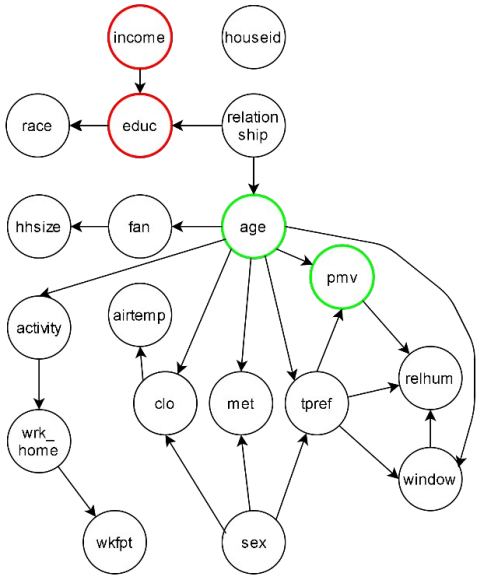
The R package, `bnlearn` has a feature to score BN models using a repeated two-fold cross-validation. In this study, we select 11 algorithms to help identify the structure, each of which is described in section 3.3. The scoring method repeats the two-fold cross-validation across 11 selected models.As a result of this approach, Tabu Search and Hill Climbing have the lowest mean of the average loss compared to the rest of the models, indicating the two would construct the most-representative BN structure (see Table 4). We use Tabu Search to run 10 iterations using a Tabu list to avoid focusing on a local optima, which is an improvement that Hill-Climbing does not have. In order to keep the network structure as simple as possible, we choose BIC as the score function in searching the best structure (see Equation 4). The third most

important procedure is the data imputation. In the experiment, we impute multivariate datasets before and after the estimation procedure using the PMM approach as described in 4.1.

The resulting BN models show strong dependencies among variables for each dataset. As an illustration, here we only discuss the model for the case of multi-family housing residents. Figure 4 shows a model graph, *G*, structures from the datasets that are adjusted via truncation and imputation. The nodes represent the variables in the model and the arrows represent the dependency between two or more variables. We also show the conditional probability table of selected variables that depend on each other. For example, “Education” and “Household Income” have a positive relationship where, households with higher income are able to afford higher education. In “PMV” given “Age”, older people tend to feel either too cold of too warm in the room.

Table 4. Bayesian network models by the choice of score function.

algorithm	**mean of the average loss**
PC	-608338.6
GS	-618044.6
IAMB	-615815.9
Fast.IAMB	-606254.0
Interleaved IAMB	-615815.9
MMPC	-611819.7
Hiton-PC	-607416.1
Hill Climbing	**-598921.4**
Tabu Search	**-598921.4**
MMHC	-602897.1
Aracne	-609735.9
Chow-Liu	-610179.6



educ	income							
	Not answer	<$10K	$10K-$25K	$25K-$50K	$50K-$75K	$75K-$100K	$100K-$125K	>$125K
High school	0.383	0.400	0.418	0.428	0.354	0.272	0.221	0.183
Assoc. degree	0.344	0.392	0.404	0.355	0.385	0.418	0.400	0.365
Bachelor's	0.176	0.147	0.126	0.153	0.167	0.202	0.255	0.276
Graduate degree	0.097	0.060	0.051	0.064	0.095	0.108	0.125	0.176

pmv	age					
	<18	18-29	30-44	45-59	60-69	>70
-3	0.396	0.000	0.000	1.000	0.479	0.000
-2	0.000	0.479	0.000	0.000	0.521	0.000
-1	0.000	0.521	0.000	0.000	0.000	0.000
0	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000
2	0.604	0.000	1.000	0.000	0.000	0.564
3	0.000	0.000	0.000	0.000	0.000	0.436

Figure 4: Model structures ***G*** for each dataset used in the case of multi-family housing residents. See Table 2 for a detailed description of the variables.

5.3. Synthetic population generation

After learning the model structure, *G*, population synthesis generates synthetic values of *X* from *P(X),* a factorized joint probability distribution defined by the BN. The values are independent and its probabilities can be computed using Equation 1. By synthesizing from the obtained network structure, we generate a large list of individuals, as introduced in Section 5.1, to form the population pool.

In the experiment, the estimation using Tabu Search runs successfully without errors on any of the variables. We fuse the two datasets (i.e. the comfort database and NHTS dataset) by two common variables of “Age” and “Sex”. The fusion proceeds using Left-Join and with a condition of the two variables to meet. We find that the more variables to include, the stricter the join process is, and the more representative the joint data is. It requires a larger dataset, a joint dataset with NaN / NA columns would result, otherwise. Next, we run the estimation procedure, which preserves a set of relationships between the two datasets (i.e., the comfort database, and NHTS dataset), such as the relationship/age dependency, which means the arrow “Relationship" → “Age” is “whitelist”-ed. The dependency structure for occupants in the multi-family housing has its own complexity at the point of assigning the individual dataset with the household dataset.

5.4. Results

We first compare the percentage difference in the distributions between the observed and synthetic populations by variables to quantify the accuracy/fitness of the population synthesis as described in Figures 5-7. For illustrations, we select several variables for each case: 15 variables for multi-family housing residents, 16 variables for office occupants, and seven for senior

housing occupants. Figure 5 shows the relatively small difference in percentages, which range between -1% and 1% for the case of multi-family housing. Office occupants have a larger range between -5% and 5%, as described in Figure 6. A larger number of samples for smaller populations may contribute to a less accurate synthetic version for the office occupants. Another possible explanation is that the office case considers more variables than those in the multi-family housing case. A slightly better result is shown for the case of senior housing residents, for which it has percent difference ranges between -2% and 2% across all seven variables.

To further test the goodness of fit, we next map the synthesis results as two-dimensional distributions. As an illustration, here, we only show two sets of joint distributions for each case. The results come from one randomly selected case from the ten total runs for each case. Figures 8-10 show the joint distributions of the observed and synthetic data on the most left and middle panels. The most-right panel of the same figure shows the probability-probability-plot (pp-plot of the joint cumulative distribution function (CDF) of the selected variables. Figure 8 shows the joint distribution of Household Income and Education in the BN for the case of multi-family housing occupants. As can be seen, BN satisfies the joint distribution between the two variables. The network structure shown in Figure 4 confirms that the pairs are approximately dependent on each other, and the dependency structures between the variables are preserved. In the case of office occupants, the resulting joint distribution of the synthetic dataset for the two selected variables Age and Met, also matches with the joint distribution of the same-set of variables from the observed data set as described in Figure 9. Being the least representative case in the field, senior housing occupants carry unique characteristics, such as homogeneity in terms of age and more sensitive perceptions of environmental conditions. Figure 10 shows that the joint distribution of Age and Met shows less variance yet matches the observed data set. The pp-plot for office and senior housing cases (Figure 9-10) shows fewer data points than for the multi-family housing case (Figure 8) for the selected variables. Therefore, a slight deviation from the $Y=X$ line can be observed in the pp-plot, yet the joint CDF is close to 1. It is then fair to conclude that despite variance from sampling, BN results satisfy the goodness of fit, even though only 2.8% (multi-family housing), 27% (office), 15.48% (senior housing) of samples are used in the learning process.

Finally, the SRMSE values are 0.485, 1.108, and 1.03 for multi-family housing, office, and senior housing, respectively. The lowest SRMSE value for the multi-family housing case demonstrates the best performance of the BN approach in generating a synthetic version of the observed data. Note that multi-family housing has a significantly large number of populations (10,764) to be sampled from. Although the population for the senior housing case (646) is three times smaller than the senior housing has (1858), it exhibits slightly better SRMSE. This is mainly because the senior housing case has a slightly smaller sampling, 15.48% (100), than the office case, 27% (500).

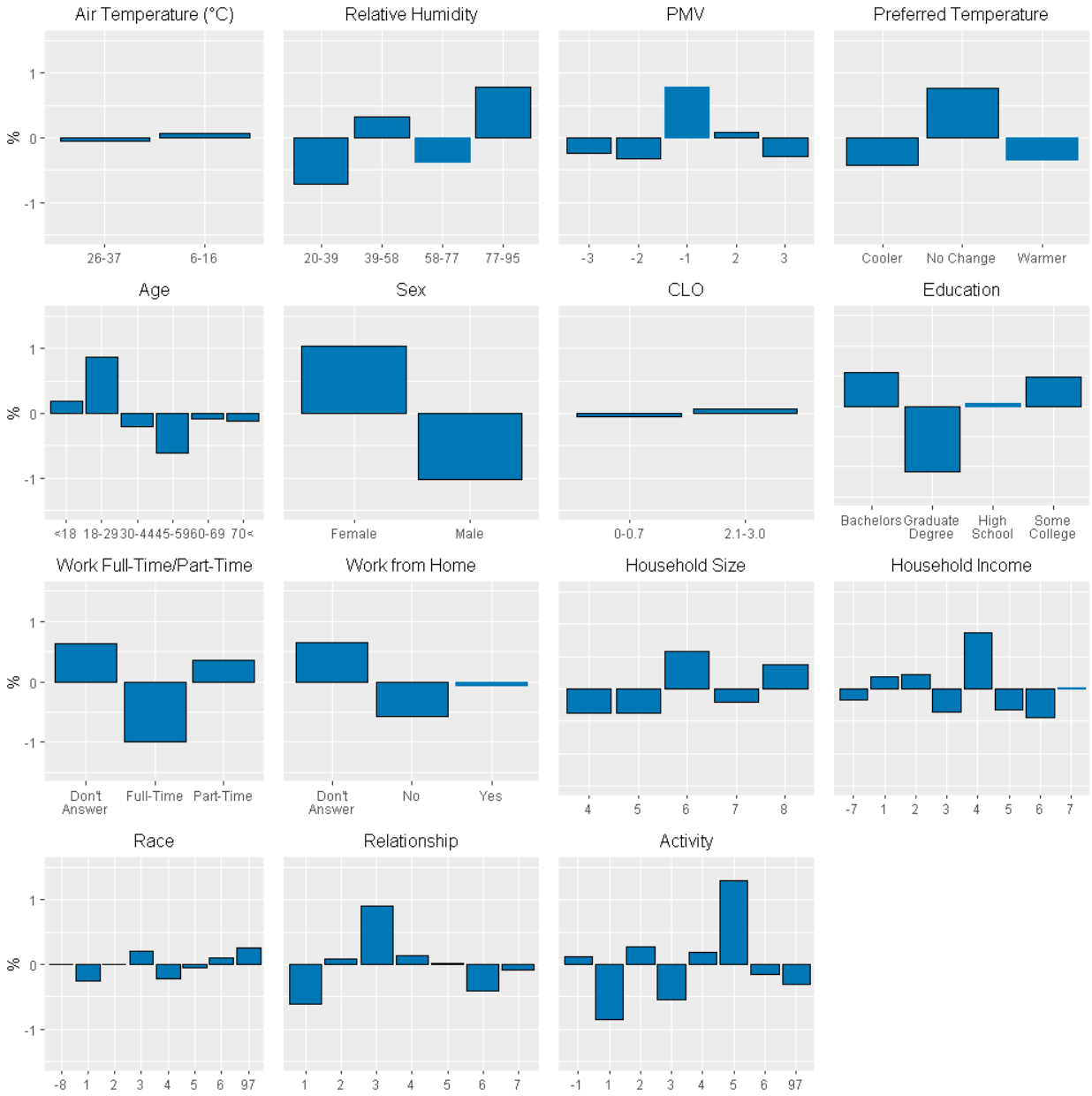


Figure 5. Comparison of observed and synthetic occupants in a multi-family housing case. (Observed – Synthetic)/Observed in Percent. See Table 2 for a detailed description of the variables.

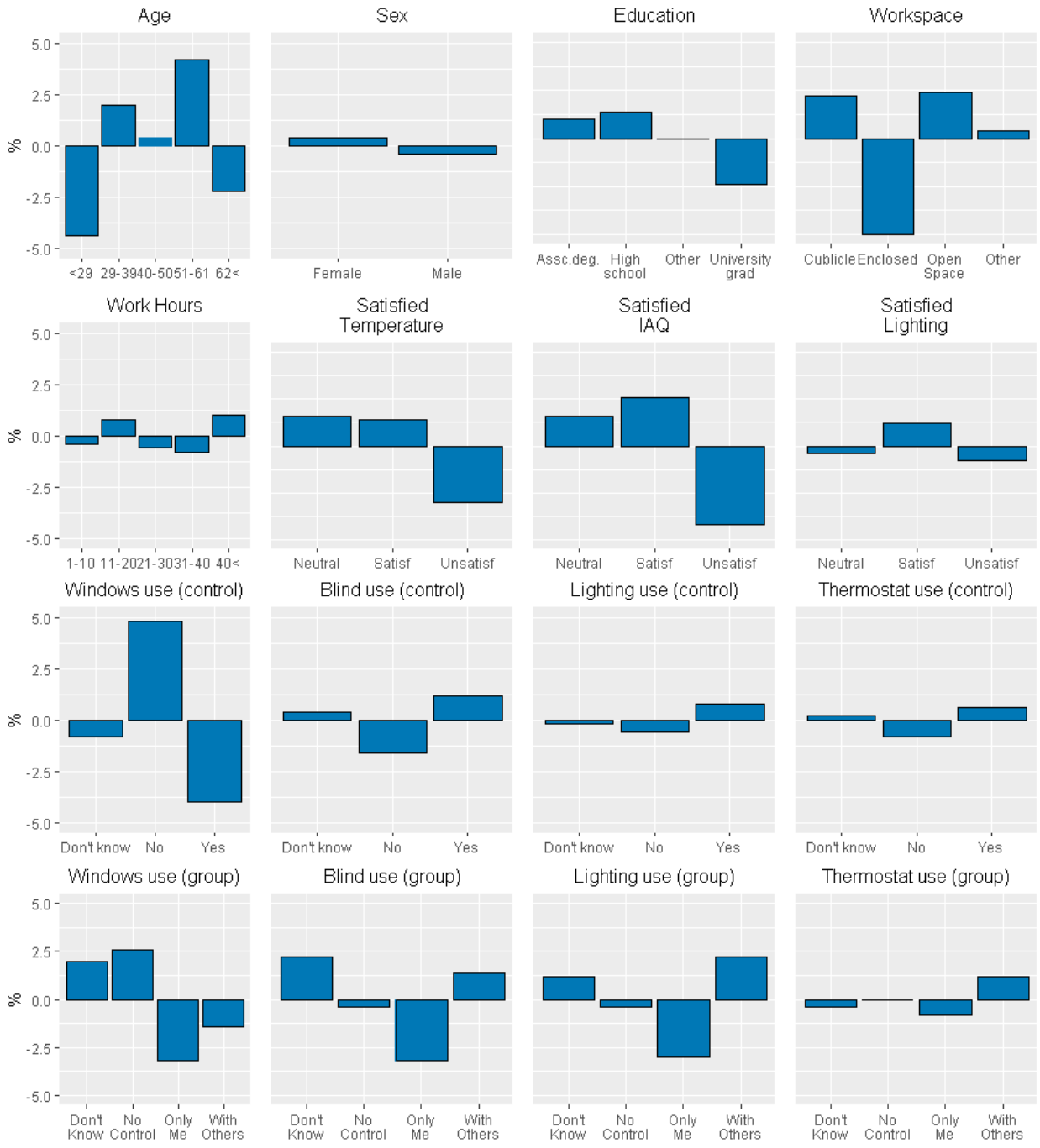


Figure 6. Comparison of observed and synthetic occupants in an office building. (Observed – Synthetic)/Observed in Percent. See Table 2 for a detailed description of the variables.

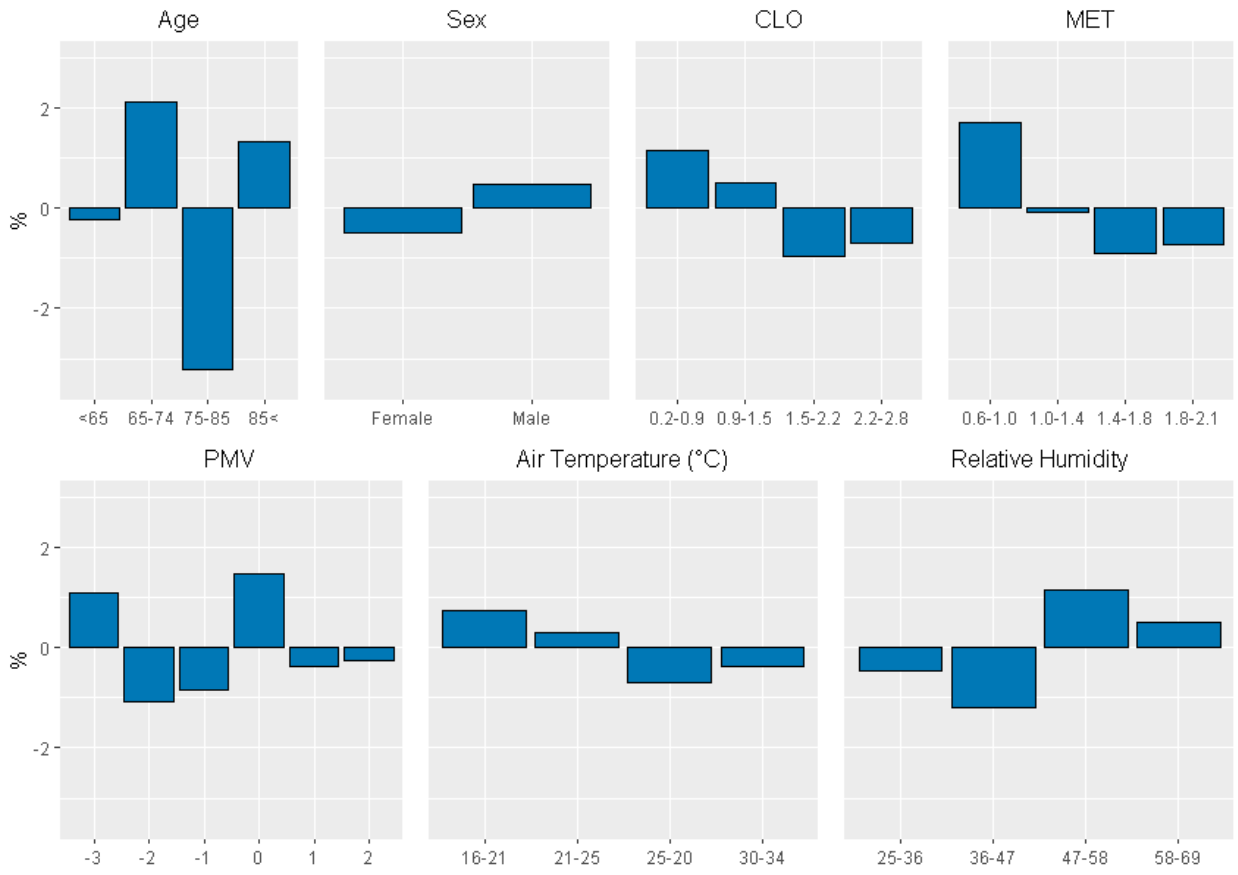


Figure 7. Comparison of observed and synthetic occupants in a senior housing. (Observed – Synthetic)/Observed in Percent. See Table 2 for a detailed description of the variables.

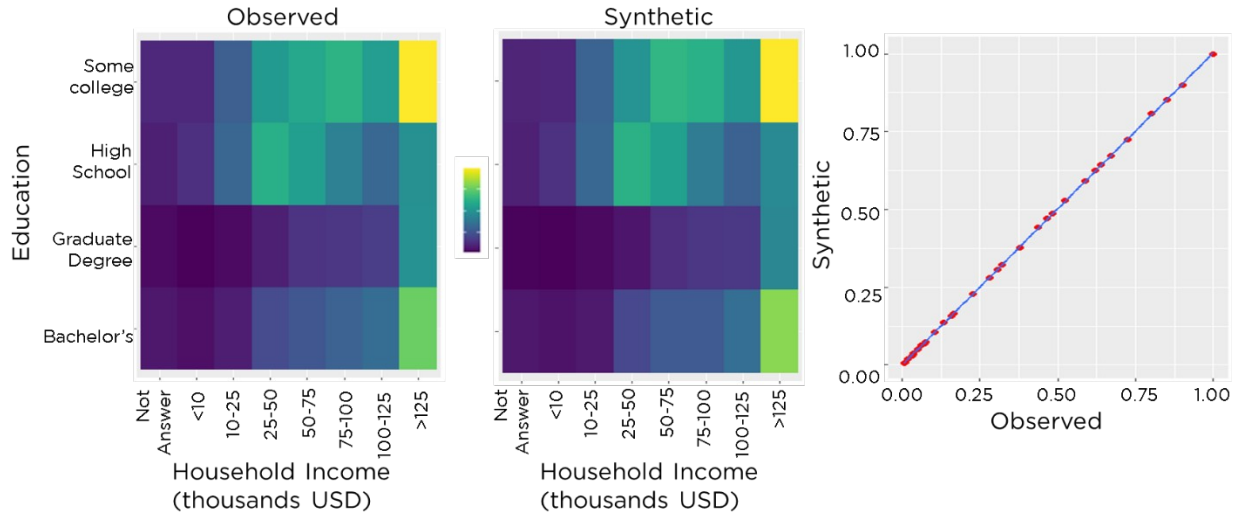


Figure 8. Joint distribution of the household income and education for multi-family housing.

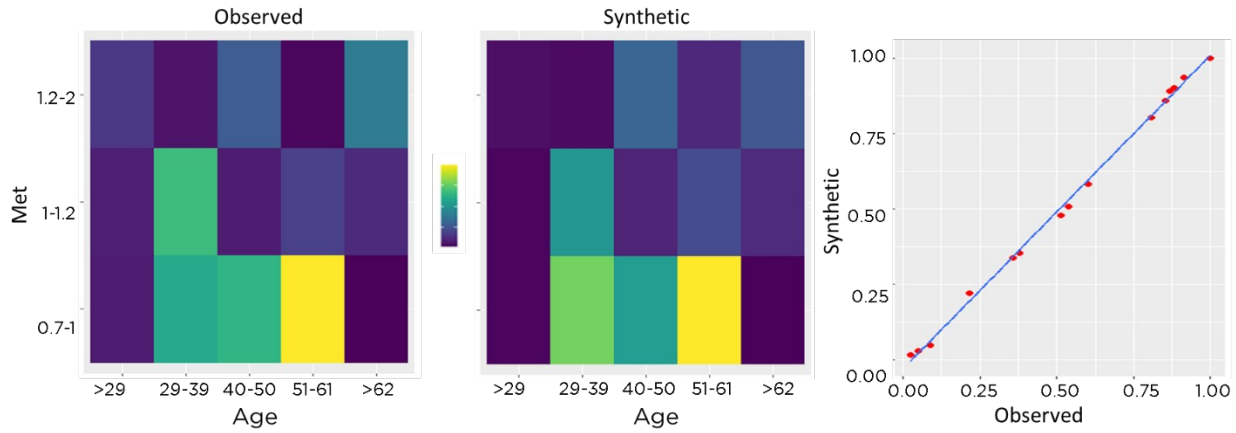


Figure 9. Joint distributions of Age and Met for office occupants

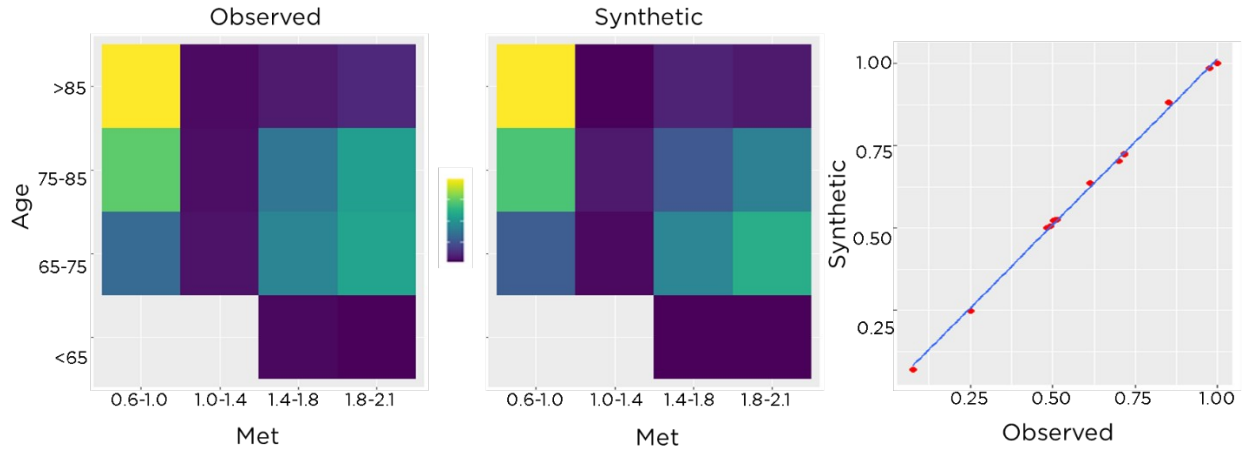


Figure 10. Joint distributions of Age and Met for occupants in a senior housing.

6. Discussion

The previous section demonstrates the process of combining several data sets, selecting relevant variables informed by the established data ontology, and generating synthetic occupant data sets. We combine socio-demographic datasets and behavior datasets to characterize the target population in the population synthesis. Three separate population synthesis procedures are determined for case studies on a multi-family housing, an office building, and a senior housing. In this setting, Bayesian Network structure learning is useful to capture the joint distribution of variables that are defined by the data ontology, DNAS framework.

A lesson learned from the applications on three building-type cases pertains to the quality of synthetic populations. The size of the target populations is one factor that determines the quality of the synthetic version. For example, if the target population is relatively small, the network structure may result in interdependencies among the nodes. Therefore, unmatched distributions between the synthetic and target populations result. We handle this issue by reducing the number of variables for senior housing residents with a relatively smaller number of observations

compared to those in the other two cases. Another important factor is the choice of the estimation algorithm. We score twelve algorithms, select Tabu Search, and run it with ten iterations to achieve optimal network structures. We also ensure the interoperability of these datasets by preserving the distributions and performing Predictive Mean Matching (PMM) for multivariate imputation. The experiments result in a good fit for all three cases, comparing the percentage difference in the distribution between observed and synthetic populations.

There are several limitations of the current implementation of the synthetic occupant population generation for occupant behavior modeling and building performance simulation. First, it considers only cross-sectional data sets to characterize the occupants and their behaviors, while longitudinal data sets (e.g. Time Use Survey data set and meter data) are important and widely used among building modelers. Future versions of the model can consider these longitudinal data sets. Another limitation is the simplistic approach to fuse several data sets together using only the intersecting variables (e.g. Age and Sex). Exploring other data fusion approaches becomes necessary as more data sets come into consideration.

Another area to improve is the grouping of individual occupants, which is commonly found in any built environment. Individual occupants may be grouped into a group-unit such as a household in a multi-family housing and a tenant-group in a multi-tenant office building. The current implementation illustrates a simplistic grouping mechanism by restricting a group variable in the learning routine and comparing the resulting distributions of individual occupants but occupant-groups. This becomes important when discussing the occupants' locus-of-control. For example, individual tenant-occupants may have to contact their tenant-representative when it comes to dimming the overhead light for their floor. Other field of studies have explored this individual and group to generate a more representative synthetic population of certain geographical areas.

7. Conclusions and future research

This paper focuses on the population synthesis using BN approach in building occupant behavior research, particularly for generating synthetic occupants that are more representative by attributing socio-demographic characteristics. Our approach, however, applies BN only once to find the best network structure for the variables. Needed are integrated population synthesis methods, i.e., Iterative Proportional Updating (IPU); utilize powerful machine learning algorithms, i.e., Generative Adversarial Networks (GANs), and a synthetic population with a greater mix of data types. Transportation research has advanced these methods up to developing a dynamic synthetic population. A dynamic synthetic population discussed in (Namazi-Rad et al., 2014) involves the age-ing of individuals in the population that is drawn upon age-dependent life-event probabilities (e.g., birth, death, marriage, and divorce). Similarly, future research on population synthesis of building occupants can include updates on the occupants' comfort preferences and choice of adaptive actions to ensure the evolution of their behaviors and determining characteristics.

Population synthesis generates a representative occupant behavior data set replacing the simplified and static schedules for building performance simulation. The resulting synthetic data sets can be useful in the cosimulation procedure linking occupant behavior and building energy models. Today's cosimulation procedures are getting more robust with fast data communication between building performance simulation tools (e.g. EnergyPlus and Modelica) and other models. Occupant behavior modules, on the other hand, are getting more complex in representing real occupants. Highly complex models may appear as an expense, particularly in the simulation time, rather than a quality improvement to the building performance simulation tools. Synthetic occupant generation comes to solve this dilemma that building occupnat modelers are facing. For example, synthetic occupant data sets could serve as an input to occupant behavior co-simulation module, such as an ABM-based OB model. Each occupant-agent in the model will be attributed by the information from the data set initially, then the attributes of preferences and actions evolve and are updated over simulation time based on actions taken at the previous time steps. Therefore, future implementation will include testing and demonstrating the synthetic occupant data sets within a cosimulation framework.

Finally, synthetic populations of building occupants are a useful tool to accommodate activities across the building lifecycle at various degrees. For example, consideration of HVAC system and thermal zoning as part of the building design phase would benefit from some relevant occupant behavior factors, including gender, age, geographical location, and thermal preferences. Other factors like income and environmental attitudes may be helpful to inform building operational retrofit activity, such as installing occupancy sensors to improve energy efficiency. Therefore, the application of population synthesis for a sensible practice of fit-for-purpose modeling is worth exploring given relevant occupant data specification.

Acknowledgments

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Building Technologies of the United States Department of Energy, under Contract No. DE-AC02-05CH11231. Authors benefited from participation and discussion in the project (2018-2023) Annex 79, Occupant-centric building design and operation, under the International Energy Agency's Energy in Buildings and Communities Programme.

References

Administration, F. H. (2017). *National Household Travel Survey*. https://nhts.ornl.gov/

Aliferis, C. F., Org, C. A., Statnikov, A., Gr, T. F., Mani, S., & Koutsoukos, X. D. (2010). Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. *Journal of Machine Learning Research*, *11*(1), 171–234.

Andrews, C. J. (2017). The Changing Socioeconomic Context of Buildings. *Journal of Solar Energy Engineering, Transactions of the ASME*, *139*(1, SI). https://doi.org/10.1115/1.4034911

Andrews, C. J., Allacci, M. S., Senick, J., Chandra-Putra, H., & Tsoulou, I. (2016). Using synthetic population data for prospective modeling of occupant behavior during design. *Energy and Buildings*, *126*, 415–423. https://doi.org/10.1016/j.enbuild.2016.05.049

Barthelemy, J., & Toint, P. L. (2013). Synthetic population generation without a sample. *Transportation Science*, *47*(2), 266–279. https://doi.org/10.1287/trsc.1120.0408

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, *30*(6 PART A), 415–429. https://doi.org/10.1016/0965-8564(96)00004-3

Berger, C. (2020). *Agent-based Modeling in Building Simulation: Where do we stand? ausgeführt*.

Berger, C., & Mahdavi, A. (2021). Exploring Cross-Modal Influences on the Evaluation of Indoor-Environmental Conditions. *Frontiers in Built Environment*, *7*. https://doi.org/10.3389/fbuil.2021.676607

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*(suppl 3), 7280–7287. https://doi.org/10.1073/PNAS.082080899

Bouckaert, R. R. (1995). *Bayesian Belief Networks: From Construction to Inference*.

Caiola, G., & Reiter, J. P. (2010). Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy*, *3*(1), 27–42.

Carlucci, S., de Simone, M., Firth, S. K., Kjærgaard, M. B., Markovic, R., Rahaman, M. S., Annaqeeb, M. K., Biandrate, S., Das, A., Dziedzic, J. W., Fajilla, G., Favero, M., Ferrando, M., Hahn, J., Han, M., Peng, Y., Salim, F., Schlüter, A., & van Treeck, C. (2020). Modeling occupant behavior in buildings. *Building and Environment*, *174*(December 2019), 106768. https://doi.org/10.1016/j.buildenv.2020.106768

Chandra-Putra, H., Andrews, C. J., & Senick, J. A. (2017). An agent-based model of building occupant behavior during load shedding. *Building Simulation*, *10*(6), 845–859. https://doi.org/10.1007/s12273-017-0384-x

Chandra-Putra, H., Hong, T., & Andrews, C. J. (2021). An ontology to represent synthetic building occupant characteristics and behavior. *Automation in Construction*, *125*(103621). https://doi.org/ https://doi.org/10.1016/j.autcon.2021.103621

Chapman, J., Siebers, O., & Robinson, D. (2018). On the multi-agent stochastic simulation of occupants in buildings. *Journal of Building Performance Simulation*, *11*(5), 604–621. https://doi.org/10.1080/19401493.2017.1417483

Chapuis, K., & Taillandier, P. (2019). *A brief review of synthetic population generation practices in agent-based social simulation*.

Chen, S., Zhang, G., Xia, X., Chen, Y., Setunge, S., & Shi, L. (2021). The impacts of occupant behavior on building energy consumption: A review. *Sustainable Energy Technologies and Assessments*, *45*, 101212. https://doi.org/10.1016/j.seta.2021.101212

Chow, C. K., & Liu, C. N. (1968). Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Transactions on Information Theory*, *14*(3), 462–467. https://doi.org/10.1109/TIT.1968.1054142

Cooley, P., Ganapathi, L., Ghneim, G., Holmberg, S., Wheaton, W., & Hollingsworth, C. R. (2008). Using influenza-like illness data to reconstruct an influenza outbreak. *Mathematical and Computer Modelling*, *48*(5–6), 929–939. https://doi.org/10.1016/j.mcm.2007.11.016

E. Ramadan, O., & P. Sisiopiku, V. (2020). A Critical Review on Population Synthesis for Activity- and Agent-Based Transportation Models. In *Transportation Systems Analysis and Assessment*. IntechOpen. https://doi.org/10.5772/intechopen.86307

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, 243–263. https://doi.org/10.1016/j.trb.2013.09.012

Friedman, N., & Koller, D. (2003). Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, *50*(1–2), 95–125. https://doi.org/10.1023/A:1020249912095

Gaetani, I., Hoes, P. J., & Hensen, J. L. M. (2016). Occupant behavior in building energy simulation: Towards a fit-for-purpose modeling strategy. *Energy and Buildings*, *121*, 188–204. https://doi.org/10.1016/j.enbuild.2016.03.038

Harland, K., Heppenstall, A., Smith, D., & Birkin, M. (2012). Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *JASSS*, *15*(1). https://doi.org/10.18564/jasss.1909

Hewitt, E. L., Andrews, C. J., Senick, J. A., Wener, R. E., Krogmann, U., & Sorensen Allacci, M. (2016). Distinguishing between green building occupants' reasoned and unplanned behaviours. *Building Research & Information*, *44*(2), 119–134. https://doi.org/10.1080/09613218.2015.1015854

Heydarian, A., McIlvennie, C., Arpan, L., Yousefi, S., Syndicus, M., Schweiker, M., Jazizadeh, F., Rissetto, R., Pisello, A. L., Piselli, C., Berger, C., Yan, Z., & Mahdavi, A. (2020). What drives our behaviors in buildings? A review on occupant interactions with building systems from the lens of behavioral theories. *Building and Environment*, *179*(April), 106928. https://doi.org/10.1016/j.buildenv.2020.106928

Hong, T., Chen, C. fei, Wang, Z., & Xu, X. (2020). Linking human-building interactions in shared offices with personality traits. *Building and Environment*, *170*(106602). https://doi.org/10.1016/j.buildenv.2019.106602

Hong, T., D'Oca, S., Taylor-Lange, S. C., Turner, W. J. N., Chen, Y., & Corgnati, S. P. (2015). An ontology to represent energy-related occupant behavior in buildings. Part II: Implementation of the DNAS framework using an XML schema. *Building and Environment*, *94*, 196–205. https://doi.org/10.1016/j.buildenv.2015.08.006

Hong, T., D'Oca, S., Turner, W. J. N., & Taylor-Lange, S. C. (2015). An ontology to represent energy-related occupant behavior in buildings. Part I: Introduction to the DNAS framework. *Building and Environment*, *92*, 764–777. https://doi.org/10.1016/j.buildenv.2015.02.019

Ilahi, A., & Axhausen, K. W. (2019). Integrating Bayesian network and generalized raking for population synthesis in Greater Jakarta. *Regional Studies, Regional Science*, *6*(1), 623–636. https://doi.org/10.1080/21681376.2019.1687011

International Energy Agency. (2017). *Energy in buildings and communities program. Annex 66: definition and simulation of occupant behavior in buildings*. http://www.annex66.org/

Jumadi, Heppenstall, A. J., Malleson, N. S., Carver, S. J., Quincey, D. J., & Manville, V. R. (2018). Modelling individual evacuation decisions during natural disasters: A case study of volcanic crisis in Merapi, Indonesia. *Geosciences (Switzerland)*, *8*(6). https://doi.org/10.3390/geosciences8060196

Koller, D., & Friedman, N. (2013). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Kontokosta, C. E., & Jain, R. K. (2015). Modeling the determinants of large-scale building water use: Implications for data-driven urban sustainability policy. *Sustainable Cities and Society*, *18*, 44–55. https://doi.org/10.1016/j.scs.2015.05.007

Larrañaga, P., Sierra, B., Gallego, M. J., Michelena, M. J., & Picaza, J. M. (1997). Learning bayesian networks by genetic algorithms: A case study in the prediction of survival in malignant skin melanoma. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1211*, 261–272. https://doi.org/10.1007/BFb0029459

Licina, V. F., Cheung, T., Zhang, H., de Dear, R., Parkinson, T., Arens, E., Chun, C., Schiavon, S., Luo, M., Brager, G., Li, P., Kaam, S., Adebamowo, M. A., Andamon, M. M., Babich, F., Bouden, C., Bukovianska, H., Candido, C., Cao, B., … Zhou, X. (2018). Development of the ASHRAE Global Thermal Comfort Database II. *Building and Environment*, *142*, 502–512. https://doi.org/10.1016/j.buildenv.2018.06.022

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.

Malik, J., & Bardhan, R. (2020). Energy target pinch analysis for optimising thermal comfort in low-income dwellings. *Journal of Building Engineering*, *28*. https://doi.org/10.1016/j.jobe.2019.101045

Malleson, N., & Birkin, M. (2013). Generating individual behavioural routines from massive social data for the simulation of urban dynamics. In *Springer Proceedings in Complexity* (pp. 849–855). Springer. https://doi.org/10.1007/978-3-319-00395-5_103

Malleson, N., Heppenstall, A., & See, L. (2010). Crime reduction through simulation: An agent-based model of burglary. *Computers, Environment and Urban Systems*, *34*(3), 236–250. https://doi.org/10.1016/j.compenvurbsys.2009.10.005

Margaritis, D. (2003). *Learning Bayesian Network Model Structure from Data*.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, *7*(SUPPL.1), 1–15. https://doi.org/10.1186/1471-2105-7-S1-S7

Markovic, R., Wölki, D., Jérôme, F., & and van Treeck, C. A. (2017). Transaction from stochastic modelling to supervised learning of occupant's behavior. *Bauphysiktage Kaiserslautern*, *4*(October), 81–83. https://publications.rwth-aachen.de/record/707348

McFadden, D., Cosslett, S., Duguay, G., & Jung, W. (1977). *Demographic Data for Policy Analysis. Urban Travel Demand Forecasting Project, Final Report Series, Vol VIII*.

Müller, K. (2017). *A generalized approach to population synthesis*. https://doi.org/10.3929/ethz-b-000171586

Müller, Kirill, & Axhausen, K. W. (2010). Population synthesis for microsimulation State of the art. *Research-Collection.Ethz.Ch*. https://doi.org/10.3929/ethz-a-006127782

Nagy, Z., Yong, F. Y., & Schlueter, A. (2016). Occupant centered lighting control: A user study on balancing comfort, acceptance, and energy consumption. *Energy and Buildings*, *126*, 310–322. https://doi.org/10.1016/j.enbuild.2016.05.075

O'Brien, W., Gunay, H. B., Tahmasebi, F., & Mahdavi, A. (2016). A preliminary study of representing the inter-occupant diversity in occupant modelling. *Https://Doi.Org/10.1080/19401493.2016.1261943*, *10*(5–6), 509–526. https://doi.org/10.1080/19401493.2016.1261943

Ortiz, M. A., & Bluyssen, P. M. (2018). Proof-of-concept of a questionnaire to understand occupants' comfort and energy behaviours: First results on home occupant archetypes. *Building and Environment*, *134*, 47–58. https://doi.org/10.1016/j.buildenv.2018.02.030

Peng, Y., Nagy, Z., & Schlüter, A. (2019). Temperature-preference learning with neural networks for occupant-centric building indoor climate controls. *Building and Environment*, *154*(January), 296–308. https://doi.org/10.1016/j.buildenv.2019.01.036

Pritchard, D. R., & Miller, E. J. (2012). Advances in population synthesis: Fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, *39*(3), 685–704. https://doi.org/10.1007/s11116-011-9367-4

Saadi, I., Mustafa, A., Teller, J., Farooq, B., & Cools, M. (2016). Hidden Markov Model-based population synthesis. *Transportation Research Part B: Methodological*, *90*, 1–21. https://doi.org/10.1016/j.trb.2016.04.007

Salimi, S., & Hammad, A. (2019). Critical review and research roadmap of office building energy management based on occupancy monitoring. In *Energy and Buildings* (Vol. 182, pp. 214–241). Elsevier Ltd. https://doi.org/10.1016/j.enbuild.2018.10.007

Smith, D. M., Pearce, J. R., & Harland, K. (2011). Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviours? An example of smoking prevalence in New Zealand. *Health and Place*, *17*(2), 618–624. https://doi.org/10.1016/j.healthplace.2011.01.001

Spirtes, P., Glymour, C., & Scheines, R. (2012). *Causation, Prediction, and Search* (Volume 81). Springer Science & Business Media.

Sun, L., & Erath, A. (2015). A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, *61*, 49–62. https://doi.org/10.1016/j.trc.2015.10.010

Sun, L., Erath, A., & Cai, M. (2018). A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological*, *114*, 199–212. https://doi.org/10.1016/j.trb.2018.06.002

Tomintz, M. N., Clarke, G. P., & Rigby, J. (2008). The geography of smoking in Leeds: estimating individual smoking rates and the implications for the location of stop smoking services. *Area*, 341–353. https://doi.org/https://doi.org/10.1111/j.1475-4762.2008.00837.x

Tsamardinos, I., Aliferis, C. F., & Statnikov, A. (2003). Algorithms for Large Scale Markov Blanket Discovery. *American Association for Artificial Intelligence*, 376–381. www.aaai.org

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The Max-Min Hill-Climbing Bayesian network structure learning algorithm. *Machine Learning*, *65*(1), 31–78. https://doi.org/10.1007/s10994-006-6889-7

Tsoulou, I., Andrews, C. J., He, R., Mainelis, G., & Senick, J. (2020). Summertime thermal conditions and senior resident behaviors in public housing: A case study in Elizabeth, NJ, USA. *Building and Environment*, *168*(106411). https://doi.org/10.1016/j.buildenv.2019.106411

Ueda, K., Tamai, M., & Yasumoto, K. (2015). A method for recognizing living activities in homes using positioning sensor and power meters. *2015 IEEE International Conference on Pervasive Computing and Communication Workshops, PerCom Workshops 2015*, 354–359. https://doi.org/10.1109/PERCOMW.2015.7134062

U.S. Census. (2015). *United states census bureau / American FactFinder*. http://factfinder2.census.gov

Wang, Z., & Ding, Y. (2015). An occupant-based energy consumption prediction model for office equipment. *Energy and Buildings*, *109*, 12–22. https://doi.org/10.1016/j.enbuild.2015.10.002

Williamson, P. (2012). An Evaluation of Two Synthetic Small-Area Microdata Simulation Methodologies: Synthetic Reconstruction and Combinatorial Optimisation. In *Spatial Microsimulation: A Reference Guide for Users* (pp. 19–47). Springer Netherlands. https://doi.org/10.1007/978-94-007-4623-7_3

Xie, C., & Waller, S. T. (2010). Estimation and application of a Bayesian network model for discrete travel choice analysis. *Transportation Letters*, *2*(2), 125–144. https://doi.org/10.3328/TL.2010.02.02.125-144

Yaramakala, S., & Margaritis, D. (2005). Speculative Markov blanket discovery for optimal feature selection. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 809–812. https://doi.org/10.1109/ICDM.2005.134