

Lawrence Berkeley National Laboratory

LBL Publications

Title

Application of the DNAS framework expansion to occupant population synthesis

Permalink

<https://escholarship.org/uc/item/94b115p8>

Authors

Chandra Putra, Handi

Hong, Tianzhen

Publication Date

2021-09-01

DOI

10.26868/25222708.2021.30207

Peer reviewed



Building Technologies & Urban Systems Division
Energy Technologies Area
Lawrence Berkeley National Laboratory

Application of the DNAS framework expansion to occupant population synthesis

Handi Chandra Putra, Tianzhen Hong

Lawrence Berkeley National Laboratory, Building Technology and Urban Systems Division, Berkeley, CA,

Energy Technologies Area
September 2021



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 Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231.

Disclaimer:

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

Application of the DNAS framework expansion to occupant population synthesis

Handi Chandra Putra¹, Tianzhen Hong¹

¹Lawrence Berkeley National Laboratory, Building Technology and Urban Systems Division, Berkeley, CA, USA

Abstract

Research in occupant behaviour is now using a more elaborate framework of building occupant interaction. Researchers often face challenges in collecting data, particularly for the data to meet the minimum number of required data points and the data interoperability requirements. Researchers address the first issue with the synthetic population and the latter with data ontologies. While synthetic population is commonly used to address the first issue, data ontology development is used to address the latter. The two solutions are complementary to each other. One of the known ontologies in building occupant behaviour research is the Drivers-Needs-Actions-Systems (DNAS) ontology, which has been used by building modelers to describe energy-related occupant behaviour. This paper describes the ontology-based synthetic population generation that can be used in the agent-based modeling (ABM) applications. This paper considers multiple data sources, including ASHRAE Thermal Comfort DB II and IEA Annex 66 data sets. A case study of an office building is used to present the workflow of DNAS framework expansion, synthetic population generation, and agent-based modeling.

Key Innovations

- The expansion of occupant data ontology, namely DNAS framework that includes a more elaborate occupant characteristics and its use in population synthesis.
- Population synthesis framework using Bayesian Networks approach.
- An application of synthetic occupant population generation in building simulation research using a case study of an office building

Practical Implications

The expansion of occupant data ontology, namely DNAS framework, supports occupant data collection effort in the building life cycle. Another use case of the effort is to inform occupant population synthesis, which is largely motivated several reasons. The main motivation is to pave the increasing the increasing need of a more elaborate data on real occupants while maintaining anonymity. Population synthesis is considered as the right approach in other disciplines, such as in transportation and urban planning, yet, still rare in the building simulation research. In the building energy simulation and occupant research, population intends to support modeling work, especially those that use agent-based modeling (ABM) approach, where each agent represents a real individual occupant.

Introduction

Building simulation research has been advancing from using standard occupancy schedules to utilizing the power of agent-based modeling (ABM) approach to represent more realistic occupancy and occupant behaviours in buildings (Berger and Mahdavi (2020)). ABM comes as a preferred technique to represent building occupants for its ability to capture interactions between building environment and occupants, and among building occupants through locus of control. The validity of occupant behavior (OB) models, including ABM-based OB models is usually done through sets of field experiments and analysis on occupancy and behavioral patterns. The better granularity of the occupant data, by including socio- and demographic characteristics as well as typical indoor activities, the better the models in providing behavioral insights to practitioners (Gaetani et al., 2016).

The lack of available data becomes one of the big challenges in developing such models. Previously developed OB models were to answer specific research questions, such as window-opening behavior in specific buildings with unique locations, purposes, and designs (Barr et al., 2011; Widén & Wäckelgård, 2010). The issue of generalizability has motivated researchers to develop larger, multi-building databases that could provide a more robust basis for OB modeling (International Energy Agency, 2017; Licina et al., 2018). Higher-level OB data often includes demographic, economic, and social attributes (Andrews, 2017; Kontokosta & Jain, 2015). The most ambitious effort to date is the development of ASHRAE Thermal Comfort Database 1 and 2 (short name: comfort database) (Licina et al., 2018). The database serves a rather smaller scope and focuses only on a selection of thermal comfort aspects of OB. Needed are larger databases that include the rich dimensionality of OB by documenting the factors that lead to behaviours, the observed behaviours themselves, and the contextual considerations of location, timing, and design that situate the behaviour. OB researchers often rely on available ontology to guide the construction of such a database. (Salimi et al., 2019) review and map existing OB data ontologies to identify use cases and gaps in the applications. The discussion on sensible practice of fit-for-purpose modeling and desired specifications of a universal OB data set is emerging in the building simulation research (Gaetani et al., 2016). This paper contributes with the development of data ontology specification to meet the fit-for-purpose modeling framework. We consider the well-established Drivers, Needs, Actions, and Systems (DNAS) ontology to serve

the goal (Hong, D’Oca, Taylor-Lange, et al., 2015; Hong, D’Oca, Turner, et al., 2015).

The paper also describes a workflow of using the ontology to generate a synthetic occupant population. Population synthesis has not been very popular in building research (Andrews, 2017). Population synthesis has grown in popularity in other disciplines, particularly in transportation (Williamson et al., 1998), public health (Beckman et al., 1996; E. Ramadan & P. Sisiopiku, 2020; Xie & Waller, 2010), and urban planning (Malleon et al., 2010). In occupant behavior research, synthetic population is useful to inform the development of occupant models. For example, researchers often find hard times to develop a complete model of under-represented populations, such as seniors in senior housing and children in a household. Population synthesis aims to fill the gap by considering other relevant data sets, such as the regional census database and National Household Travel Survey (NHTS). In generating a synthetic occupant population, the synthetic version should match in statistical measures with those in a target population without attempting to replicate each individual in the target population. Synthetic population is also useful to maintain privacy and anonymity of the real individuals being surveyed. Also, it is rather less expensive than to collect new raw data. Another use case is for the modeling purposes, a synthetic population may serve as the initial population in the simulation modeling workflow, such as ABM-based models.

In Section 2 of this paper, we introduce the expansion of the DNAS framework which includes four new major components: (1) socio-economic, (2) geographical location, (3) subjective-values, and (4) activities. We also discuss a population synthesis framework using Bayesian Networks in the section. In Section 3, we present a generation of synthetic population and its application on office building occupants. Section 4 discusses the results and Section 5 offers conclusions and recommendations for future research.

Methods

This section presents the framework in which use cases, data ontology, data sets, and Bayesian Networks (BN) are the main components in generating a statistically-fit synthetic occupant population.

The expansion of DNAS framework

Occupant behavior research incorporates an interactive component between the occupants and building systems, equipment, and mechanisms (Hong, D’Oca, Turner, et al., 2015). The epistemological model of occupant behavior uses the concept of comfort, which is driven by socio-demography, location, subjective values, activities, and social influence. Comforts are specific combinations of the attributes of mechanisms and measures taken as consideration by an occupant. The distinction between ontological and epistemological concerns is clear on how occupant’s perception of the environment depends on

their characteristics, adopted values, experience, and particular activities (Grandjean et al., 2012).

Therefore, we use an ontological approach in selecting the datasets to maintain a well-received building occupant data structure. Without the guiding data structure, population synthesis can take a large computing power and result a synthetic population not matched with the original samples. This study follows the existing Drivers, Needs, Actions, and Systems (DNAS) framework to provide a better representation of building occupants (Hong, D’Oca, Taylor-Lange, et al., 2015; Hong, D’Oca, Turner, et al., 2015).

The expanded DNAS framework categorizes the occupant characteristics into four groups, including socio-demographic, location, subjective values, and activities (Chandra-Putra et al., 2021). The framework also develops further the comfort adaptive action component by dividing them into two sub-categories based on individual control or collective decisions (see Figure 1). The sociodemographic 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. They 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”.

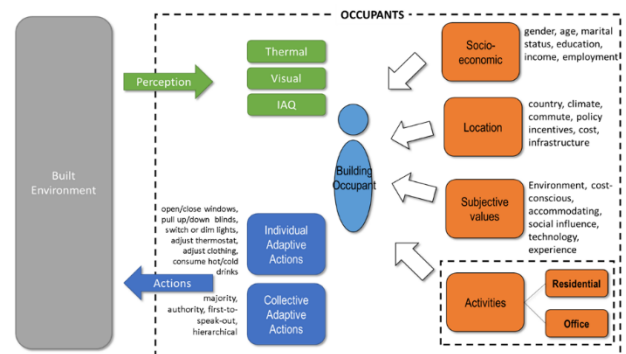


Figure 1: Conceptual interaction of components of the building occupant (Chandra-Putra et al., 2021).

The 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. We utilize the same schema format as the original DNAS framework, which is in obXML (short name for occupant behavior XML) (Hong, D’Oca, Taylor-Lange, et al., 2015). Figure 2 shows a number of child elements are added under the Occupant element to detail the occupant characteristics. They are *SocioEconomic*, *GeographicLocation*, and *SubjectiveValues*. Each child element also has its own children as described in Figure 1. The ease of closely

following the data structure depends on data availability and interoperability between data sets, and it is very common for population synthesis to also perform data imputation.

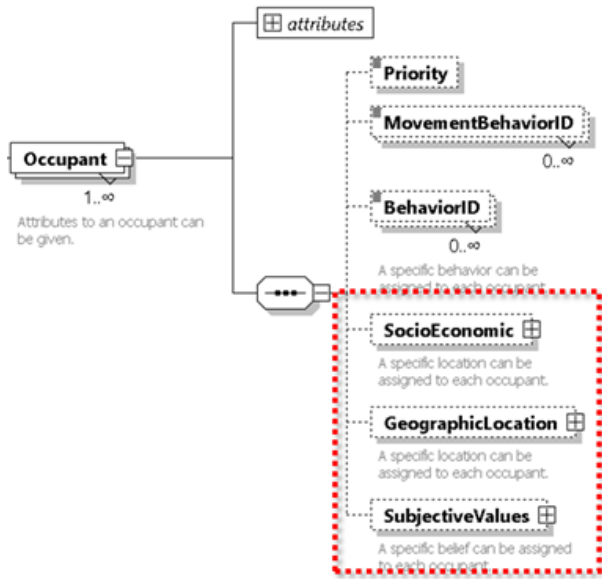


Figure 2: Occupant's attributes of social, economic, and demographic characteristics (Chandra-Putra et al., 2021)

Datasets overview

We consider two types of datasets to inform the generation of synthetic occupant population. The first is socio-demographic data that is attributed to the occupants. The second is behavior data that is specified by type of buildings. In Figure 3, the socio-demographic data considers two sources, including the National Household Travel Survey (NHTS) (Administration, 2017) and Public Use Microdata Survey (PUMS) (U.S. Census, 2015). The occupant behavior dataset, also shown on Figure 3, 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 comprises of 4,324 observations. The larger and more recent comfort database 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. 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, 1988; Rubin, 1986). 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 realistic. Therefore, it allows to impute discrete values, which is useful for the survey datasets used in the study. Figure 3 shows the framework model of data imputation, fusing, and synthesis.

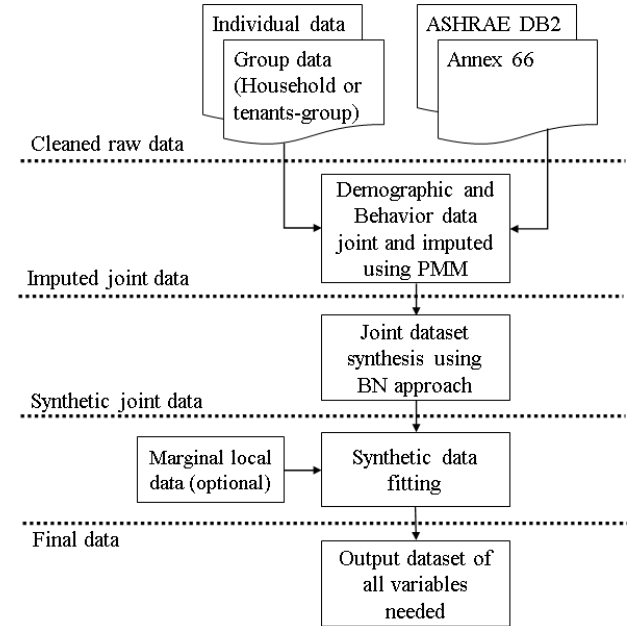


Figure 3: Population synthesis framework.

The selection of relevant variables from the data sets is guided by the DNAS data ontology. Table 1 describes the variables of interest of the case study of office occupants.

Table 1: Variable names and descriptions

Variable	Description
Sex	Gender of an occupant
Age	Age of an occupant
Education	Educational attainment
MET	Metabolic activity
CLO	CLO
Air temperature	Air temperature in °C
Relative Humidity	Relative humidity
PMV	Predicted Mean Vote (PMV)
Preferred temperature	Preferred temperature {cooler, no change, warmer}
Fan	Fan use
Window	Window operation
Heater	Heater use
Workspace	Workspace {enclosed space, open space, cubicle}
Work hours	Working hours
Satisfaction (temperature, IAQ, natural lighting, daylighting, artificial lighting, acoustics)	Indoor environment quality {unsatisfied, neutral, satisfied}

Control (lighting, windows, blinds, thermostat)	Have control on fixture {yes, no, I don't know}
Group control (windows, blinds, lighting, thermostat)	Group control on fixture {only me, no control, with others}

Bayesian Networks

The Bayesian network (BN) offers a graphical representation of probability distributions for a set of variables of interest $X = X_1, \dots, X_n$ (Koller & Friedman, 2013; Spirtes et al., 2012). In principle, a BN structure consists of two parts: 1) a network structure G in the form of a directed acyclic graph (DAG), 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 $\Theta = \{P(X_1|\Pi_1), \dots, P(X_n|\Pi_n)\}$ for each vertex X_i , conditional on its parents Π_i . The structure G follows the conditional independence assumption, by which each variable X_i is independent of its nondescendants given its parents in the graph G . In BN, the DAG topology asserts only the conditional dependence of children given parents:

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

Figure 4 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 $P(\text{ThermalPerception}|\text{Age})$ and $P(\text{Income}|\text{Age})$.

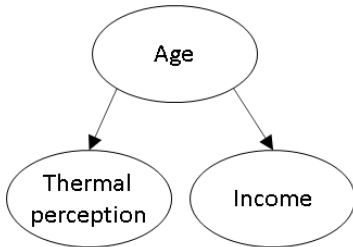


Figure 4: Example of a directed acyclic graph (DAG).

One of the main challenges in using BN is in defining the network structure. A researcher often defines the structure based on the researcher's domain-expert knowledge. This study has a particular interest to build the network structure directly from data, which is often called structural learning. Structure learning is a flexible feature of the package that identifies the relations and hierarchies of the variables. We use an R package, `bnlearn` that offers several algorithms to perform structure learning, including i.e., Tabu Search and Hill Climbing (Scutari, 2010). We consider eleven algorithms in order to find the most robust estimated model structure that is useful in the synthetic population generation. Common structure learning algorithms can be grouped into three categories, constraint-based, score-based, and hybrid. While constraint-based algorithms ensure conditional independence constraints using statistical tests, score-based algorithms rely on optimization techniques with each candidate DAG is assigned a score as the objective function. Hybrid algorithms integrate the two methods by reducing space of candidate DAGs using a constraint-

based strategy and implement a score-based strategy to find the optimal DAG in the constrained space. It was known that constraint-based algorithms are considerably more accurate than score-based algorithms for small sample sizes and they both are as accurate as hybrid ones.

We select Tabu Search because it outperforms other algorithms, showing that it produces the most representative BN structure. Tabu Search, essentially, utilizes an iterative searching procedure to obtain a best solution from complex correlation patterns (Glover et al., 1993). The algorithm selects a close solution to optimality in order to minimize the score. Another advantage of using `bnlearn` as the choice of tool for conducting structural learning procedure is its ability 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:

$$l(G^h|D) = \underset{G}{\text{max}} \sup_{\theta} l(G, \theta|D) = \underset{G}{\text{max}} l(G, \hat{\theta}|D) \quad (2)$$

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

$$BIC(G^h|D) = \log P(D|G^h, \hat{\Theta}) - \frac{d}{2} \log m \quad (3)$$

$$AIC(G^h|D) = \log P(D|G^h, \hat{\Theta}) - d \quad (4)$$

Where $\hat{\Theta}$ refers to the maximum likelihood estimates $\hat{\theta}$ 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 contains 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.

Results

The application of the synthetic population framework and procedures show consistent network structures and statistically-fit synthetic occupant populations. We present illustrative results for the calibration scenarios in an office building in this section.

We consider a joint sub- dataset from the Annex66 dataset and comfort database using three joint variables, i.e., "Age", "Sex", and "PMV". We sample 27% of total 1,858 observations (=500 observations) in learning BN models. The experiment illustrates the dynamic in real office buildings, where the locus of control varies among occupants when operating building features. Occupants may not have direct control over overhead lights. Negotiation on controls between an individual tenant and a tenant representative, and between a tenant

representative and a building manager are expected to occur.

We run Tabu Search for ten iterations to construct the network structure on the joint data sets and impute using PMM approach. The resulting BN model structure shows dependencies among variables. A variable "Blind use", as for example, has two edges coming in from nodes "Sex" and "Fan". Figure 5 shows the final model structure for office occupants.

After learning the model structure, G , population synthesis generates sample values of X from $P(X)$, factorized joint probability distribution define by the BN. The samples are independent and its probabilities can be computed using Equation 1. We fuse the two data sets (i.e. the comfort database and NHTS data set) based on two common variables of "Age" and "Sex". The fusing 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.

To quantify the accuracy/fitness of the resulting synthetic population, we compare the percentage difference in the distributions between variables in the observed and synthetic populations. Figure 6 shows relatively small differences in percentages, which range between -5% and 5% (see Figure 6). To further test the goodness of fit, we map the synthesis results in two-dimensional distributions. As an illustration, Figure 7 shows the joint distributions of variables "Age" and "Met" in the BN. The left and middle figures show the joint distributions in the observed and synthetic data, respectively. The right figure, which is a probability-probability-plot (pp-plot) of the cumulative distribution function (cdf) of the two variables, shows the fit of joint distribution of both target and BN. It shows that BN satisfies the joint distribution between the "Age" variable and "Met" variable. The pairs are approximately dependent to each other, and the dependency structures between the variables are preserved as seen on the network structure in Figure 5.

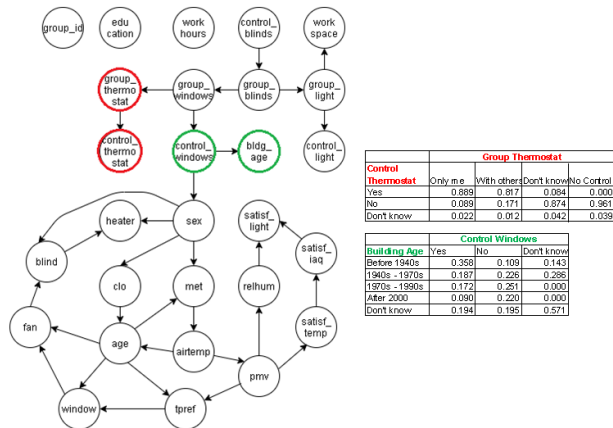


Figure 5: Model structure, G , for each dataset used in an office building.

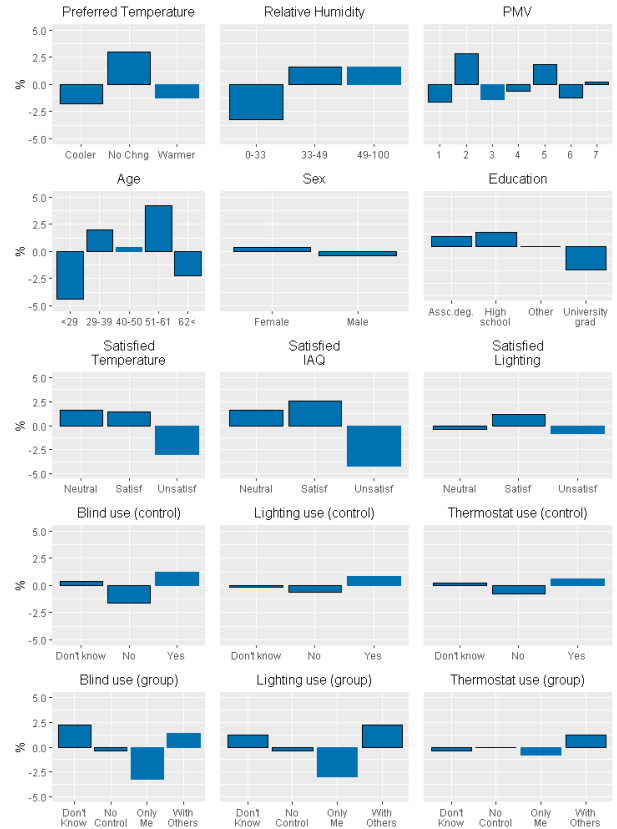


Figure 6: Comparison of observed and synthetic occupants in an office building. See Table 2 for a detailed description of the variables.

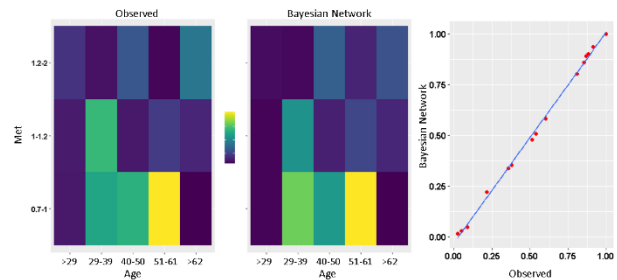


Figure 7: Joint distribution of Age and Met for office occupants.

Discussion

The expansion of DNAS framework is useful in generating synthetic occupant population. This paper does not attempt to develop a more exhaustive ontology that is potentially applicable to limited applications. We propose the expansion of the framework to include a more detailed occupant characteristics such as socio-demographic, climate, and subjective values that have been covered in previous research on building occupant interactions. Two subjective values that are relevant to OB research are, as for example, preference on comfort and willingness to pay for utility.

Then, we use the proposed expanded ontology to inform data collection and variables selection to generate the synthetic occupant population in this paper. Our focus is to develop a BN model that could capture joint

distributions for variables in the structure that are defined by the proposed expansion of the DNAS framework.

From the experiments, we learned that the size of a target population is an important determinant to the quality of a synthetic population. For example, given the target population is relatively small, the network structure may result in independencies among the nodes. Therefore, the synthetic population may have unmatched distributions to the target populations result. We handle this issue by reducing the number of variables in the model. Another important factor is the choice of the estimation algorithm. We score eleven algorithms, among which Tabu Search outperforms the best. We run Tabu Search in ten iterations to achieve optimal network structure. We also ensure the interoperability of these datasets by preserving the distributions and performing Predictive Mean Matching (PMM) for multivariate imputation.

Conclusions

The proposed expansion of DNAS framework, an ontology to represent occupant behaviors, and the generation of synthetic occupant population using Bayesian Network (BN), are intended to serve the fit-for-purpose occupant modeling effort.

We build the confidence in both the proposed framework and population synthesis procedure by calibrating against the real data on building occupants, linking the previously collected behaviour data and socio-demographic data. We acknowledge the challenging aspect of the overall procedure is present at the data fusion phase. Researchers often collect their data that is recent, small, follow a specific ontology and mix-types between cross-sectional and longitudinal data depending on the purpose of the study. Therefore, the framework helps in integrating the data sets and ensuring their interoperability.

While population synthesis using the BN approach is not new, this paper enriches existing literature on its practical applications on the building occupant behaviour research, particularly about generating synthetic occupants that are more representative by attributing socio-demographic characteristics. Our efforts are, however, not without challenges. While our approach is tested using BN, our future research regarding the topic of developing integrated population synthesis methods will consider other existing approaches as for comparison, including, 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. Dynamic discussed in (Namazi-Rad et al. (2014)) involves the ageing of individuals in the population that is drawn upon age-dependent life-event probabilities (e.g., birth, death, marriage, and divorce). Similarly, population synthesis of building occupants may include updates on the occupants' comfort preferences and choice of adaptive actions to ensure the evolution of their behaviours and determining characteristics. Finally, we are looking forward to applying under different

building types, such as multi-family residences and senior housing. The use case of building occupant population synthesis under these two buildings is even greater since children in a multi-family household and senior population in senior housing are often overlooked in the occupant behaviour research.

Acknowledgment

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. DEAC02-05CH11231. 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 Program.

References

- Administration, F. H. (2017). *National Household Travel Survey*. <https://nhts.ornl.gov/>
- Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- 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>
- Barr, S., Shaw, G., & Coles, T. (2011). Sustainable lifestyles: Sites, practices, and policy. *Environment and Planning A*, 43(12), 3011–3029. <https://doi.org/10.1068/a43529>
- 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](https://doi.org/10.1016/0965-8564(96)00004-3)
- 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>
- 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>
- 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>
- Glover, F., Taillard, E., & Taillard, E. (1993). A user's guide to tabu search. *Annals of Operations*

- Research*, 41(1), 1–28.
<https://doi.org/10.1007/BF02078647>
- Grandjean, A., Adnot, J., & Binet, G. (2012). A review and an analysis of the residential electric load curve models. *Renewable and Sustainable Energy Reviews*, 16(9), 6539–6565.
<https://doi.org/10.1016/j.rser.2012.08.013>
- 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>
- International Energy Agency. (2017). *Energy in buildings and communities program. Annex 66: definition and simulation of occupant behavior in buildings*. <http://www.annex66.org/>
- 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>
- 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. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, 6(3), 287–296.
<https://doi.org/10.1080/07350015.1988.10509663>
- 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>
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5), 465–471.
[https://doi.org/10.1016/0005-1098\(78\)90005-5](https://doi.org/10.1016/0005-1098(78)90005-5)
- Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1), 87–94.
<https://doi.org/10.1080/07350015.1986.10509497>
- Salimi, S., Nik-Bakht, M., & Hammad, A. (2019). Towards an ontology for holistic building occupant information modelling. *IOP Conference Series: Materials Science and Engineering*, 609(62025).
<https://doi.org/10.1088/1757-899x/609/6/062025>
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
<https://doi.org/10.1214/aos/1176344136>
- Scutari, M. (2010). Learning Bayesian networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3), 1–22. <https://doi.org/10.18637/jss.v035.i03>
- 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>
- U.S. Census. (2015). *United states census bureau / American FactFinder*. <http://factfinder2.census.gov>
- Widén, J., & Wäckelgård, E. (2010). A high-resolution stochastic model of domestic activity patterns and electricity demand. *Applied Energy*, 87(6), 1880–1892.
<https://doi.org/10.1016/j.apenergy.2009.11.006>
- Williamson, P., Birkin, M., & Rees, P. H. (1998). The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A*, 30(5), 785–816. <https://doi.org/10.1068/a300785>
- 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>