Development of a Framework for the Analysis of Decarbonisation Paths for Existing Canadian Houses

Rasoul Asaee
Natural Resources Canada, CanmetENERGY-Ottawa, Ottawa, Ontario, Canada

Abstract

Understanding the energy performance of existing buildings is an essential step for developing programs aiming to reduce the energy demand of the building sector. Data-driven approaches gained significant attention in recent years for the analysis of energy use in buildings. This article presents a semi-automated framework for the rapid development of stock models at the national, regional, and municipal levels. It utilizes the strength of both data-driven and energy modeling techniques. The framework lays the groundwork for the impact analysis of energy efficiency programs at provincial and municipal levels. The stock models benchmark the status of existing houses and provide tools for comprehensive analysis of decarbonization scenarios. Major components of the proposed pipeline have been developed, and the work on the remaining pieces is in progress. The results of stock representation for seven Canadian regions are presented in this article. The model was used to assess the current status of existing Canadian houses.

Introduction

Energy efficiency, advanced construction techniques, and improved heating, ventilation and air conditioning (HVAC) systems are considered viable solutions for reducing the greenhouse gas emissions of the building sector. However, during the last few decades, the population growth and lifestyle improvements in developing countries increased the energy demand. The energy efficiency improvements were not sufficient to compete with the growing demand. Previous studies indicated that about 70% of existing houses in Canada are over 30 years old, and the majority of existing houses were built in the absence of building energy efficiency programs (Asaee et al., 2018). Energy retrofit options, including energy efficiency, renewable/alternative energy technologies, and decarbonizing the energy supply, would be necessary to achieve net-zero emissions status for the residential sector. Therefore, regulatory organizations and policymakers focused their efforts on improving energy efficiency and promoting renewable energy technologies in various types of buildings. The Pan-Canadian Framework on Clean Growth and Climate Change (Government of Canada, 2016) recognized the need to develop net-zero energy ready model building codes for new buildings and alterations codes for existing buildings.

Historically, researchers used building performance simulation to evaluate the building envelope energy performance, estimate the energy consumption of HVAC systems, and assess the effectiveness of energy upgrade scenarios. Growing access to building data, cloud infrastructure, and advanced machine learning tools promoted the use of machine learning models in building design in the last decade. Developing a machine-learning model requires data collection, cleaning, processing, model development, training, and validation.

This study presents a semi-automated framework that utilizes energy modeling and machine learning techniques to expedite the development of bottom-up housing stock models to support regulations and incentive programs for the decarbonization of existing Canadian houses at national, regional, and municipal levels. The framework includes two main components: (i) stock representation and (ii) engineering model. Figure 1 depicts the schematic of the workflow for the proposed framework. The main phases of the framework are:

- Extract, transform, and load (ETL) of the housing data: The model consumes data from EnerGuide for housing database (EGHD), Survey of Household Energy Use (SHEU), and municipal data to develop representative archetypes. Since the data is collected from various sources, the attributes and data quality may differ for each use case. The first step in this framework is to clean the data and ensure the format is compatible with the rest of the model.
- Stock representation: Statistical data is used to evaluate the distribution of existing houses by shape, geometry, construction characteristics, and energy sources. The distributions will be used as a guiding principle for sub-sampling the available energy models and creating a database of representative archetypes.
- Market data: The availability of energy conservation measures (ECM), materials cost, and installation rates are provided as input to the model. These data, along with the archetypes, will be used to assess the performance of energy upgrades in the existing houses.

- Engineering model: The engineering model uses an automation and optimization engine to create the input files for each combination of ECMs, run the batch simulations, and collect the output results. A translator automatically creates an equivalent version of the energy models in EnergyPlus for hourly calculations. The hourly results enable a more accurate estimation of the cooling loads and advanced renewable energy technologies analysis.
- Data analysis: Finally, the results of the engineering model will be used to assess the effectiveness of decarbonization scenarios. The surrogate model will be re-trained for each use case to fit the simulation results. The surrogate model will be accessible through an API and provide data to downstream applications such as energy mapping.

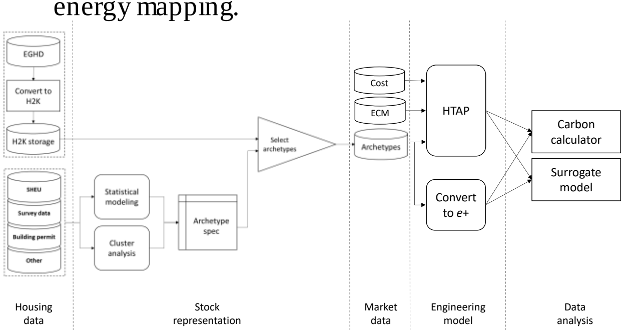

Figure 1. Schematic workflow of the proposed model

This article provides a detailed description and the results of the stock representation process used to develop a housing stock model for seven regions across Canada.

Methodology

This study uses a data-driven approach for stock representation and an engineering model for the impact analysis of retrofit scenarios.

Housing data

The Energuide for Houses Database (EGHD) is maintained by NRCan's Office of Energy Efficiency (OEE). The EGHD contains housing audit data and energy models collected by OEE's various housing programs over the last 20 years. HOT2000 was used for the development of energy models. HOT2000 estimates the energy requirements (i.e., space heating, water heating, appliances, and lighting) of a building using the monthly energy balance technique. The EGHD contains over 2,000,000 audit records (including pre- and post-retrofit audit data), representing approximately 1,000,000 unique dwellings in Canada. Each audit record includes over 200 attributes describing the location, physical measurements, and HOT2000 analysis estimates for the home. Audit data is collected when homeowners and homebuilders participate in one of OEE's voluntary programs:

1. The EnerGuide for Houses (EGH) program provides independent expert advice regarding the energy efficiency level of homes to homeowners. EGH is a voluntary program, and interested participants contact local program representatives to arrange an energy audit (NRCan 2018a, Aydinalp 2001). The EnerGuide for Houses program provides pre- and post-retrofit audits, allowing homeowners who undertake upgrades to quantify energy-related benefits
2. R-2000 (NRCan, 2018b) is a voluntary housing standard designed and maintained by the OEE to promote a high level of energy efficiency in the Canadian housing stock. Typically, an R-2000 certified home is at least 50% more energy efficient than a code-built house. The R-2000 program was initially launched in 1982, and the standard has been updated continuously.
3. The US Environmental Protection Agency originally developed the ENERGY STAR program in 1992. OEE adopted the ENERGY STAR program and promoted it in Canada since 2001. As a part of that program, the ENERGY STAR for new homes was introduced to enhance the energy efficiency of new houses across Canada. An ENERGY STAR certified home is built to use about 20% less energy than a typical home built based on the latest building energy code.

Participation in any of these programs requires physical measurements of the home, HOT2000 analysis, and submission of the data to OEE for inclusion in the EGHD. Since the participation is voluntary, the EGHD is not statistically representative of the Canadian housing stock.

Extract, transform, and load (ETL) workflow

As described in the previous section, EGHD consists of over 2,000,000 records of existing houses collected over 20 years. The audit data was stored in HOT2000 models. The EGHD data pipeline generates a summary of house characteristics upon submission of HOT2000 models to the database. The summary data is stored in a tab-separated format (TSV) and can be conveniently parsed by conventional data analysis software such as Microsoft Office Excel. However, the detailed information about the geometry, size, and construction characteristics of envelope components and a detailed description of heating, ventilation, and air conditioning systems is only available in the original HOT2000 model.

Over the last 20 years, builders, designers, and energy advisors extensively used HOT2000 to design energy-efficient homes and participate in energy efficiency programs. While the overall simulation of the approach of HOT2000 remains consistent, the underlying assumptions for modeling various building components have been updated, and the data structure has evolved to ensure it can fulfill the new requirements over the last two decades.

Generally, each version of HOT2000 should be able to read the models developed with the earlier versions of the software. This feature is available through the graphical user interface (GUI) of HOT2000. However, in some instances, the errors due to the upgraded assumptions and algorithms may not be automatically solved by the GUI. An expert should diagnose the converted model and fix the errors in such circumstances.

Additionally, the HOT2000 data structure has evolved through each update. A significant update occurred during the release of HOT2000 version 11 when it started using extensible markup language (XML). In earlier versions, HOT2000 used a binary format for saving the input and output data. The change in the file format enabled the development of tools for automating processes on the HOT2000 files without accessing the GUI. Such applications generally parse XML files to read the data, manipulate inputs, and read the outputs. NRCan's Housing Technology Assessment Platform (HTAP) is one such HOT2000-compatible application. HTAP was developed to identify optimized, cost-effective approaches for housing design and was extensively used for the impact analysis of building codes and energy incentive programs. Therefore, upgrading the HOT2000 files to version 11 is paramount for automating batch processes.

As shown in Figure 2, about 6 percent of the EGH records were created with HOT2000 version 11. The remainder of the EGH records are the HOT2000 models generated with earlier versions and are not available for use in automated data pipelines in their current format.

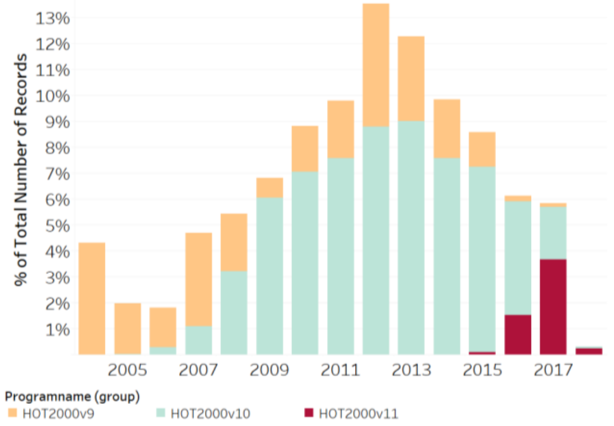

Figure 2. Distribution of EGH records by the version of HOT2000 software used for the model development

HOT2000's file conversion functions are designed to port house files from older binary formats to newer XML formats. However, this process is time-consuming, and the transformation of thousands of files by a human expert is not feasible. This study might have expedited archetype generation by focussing on XML-formatted files generated by recent versions of HOT2000 (v11+). However, this approach would have excluded the vast majority of EGHD records and was deemed unacceptable. Therefore, a process was designed for automating the conversion of models using HOT2000 GUI. For this purpose, a python module, "pywinauto" (Mahon, 2022), was used to automate the Microsoft Windows GUI. The pywinauto provides tools to automatically sends mouse and keyboard actions to windows dialogs and controls.

HOT2000 only supports incremental conversions between file formats — Version 9 files (known as .hdf format) must first be converted into version 10 formats (known as .hse) before they can be switched into version 11 (known as .h2k). The subsequent upgrade from HOT2000 version 11.3 to the latest versions is relatively less complicated and can be accomplished using the command-line interface (CLI) of HOT2000 version 11.

The schematic of the extract, transform, and load (ETL) pipeline to upgrade the EGHD records is shown in Figure 3. An algorithm was developed in Python to extract the files, make the transformation, and save the updated files. The algorithm includes multiple steps to handle errors and exceptions.

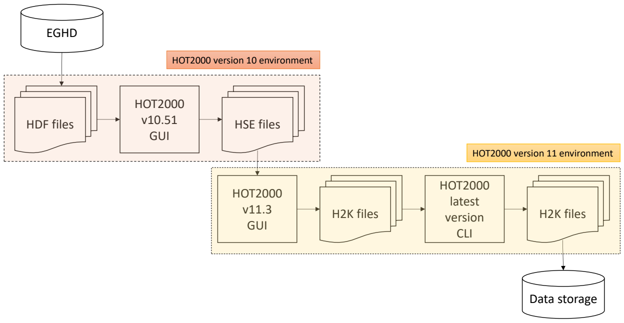

Figure 3. The ETL pipeline for the conversion of legacy HOT2000 files to the latest version

Stock representation

Archetypes are not a new tool and have proven their capabilities for evidence-based energy policy development through many studies. Identification of archetypes can be based on experience or data-driven (Ghiassi & Mahdavi, 2017; De Jaeger et al., 2020). Experience-based (also known as engineering-based) archetype identification relies on human experts who possess practical and theoretical knowledge to select buildings they believe represent the stock. This method is preferential when data is poor or unavailable. Data-based methods are used to identify archetypes when extensive information regarding the housing stock's characteristics is available. Traditional data-based methods utilize filtering techniques on categorical variables like end-use and construction period to define building classes before statistical analysis, where an archetype is identified that represents the samples' mean. In experience-based identification, subjectivity is introduced from the use of human judgment, as it is common to analyze single building characteristics one at a time. More importance is placed on particular characteristics, and others are disregarded, resulting in biased archetypes. Hence alternative approaches that can analyze multiple factors in parallel are preferable (Tardioli et al., 2018).

Clustering algorithm

The clustering process in this work includes database preparation, variable selection, data pre-processing, clustering, and interpretation and validation of the resulting clusters.

The objective of clustering analysis is to divide the houses into groups of similar characteristics. In this study, the K-means clustering algorithm is applied to achieve this, as this algorithm is considered the best for residential archetype development (Ali et al., 2019; Tardioli et al., 2018; De Jaeger et al., 2020; Li et al., 2018; Ghiassi & Mahdavi, 2017).

K-means is one of the most popular algorithms for clustering tasks due to its simplicity and computational efficiency (linear time complexity). The algorithm partitions n objects (buildings) into K non-overlapping clusters by minimizing the sum of distances to the points of their respective centroids. Euclidean distance is the most common proximity measure. Unsupervised machine learning is meant to uncover natural patterns within unlabelled data, so the number of clusters is not normally known beforehand. In most cases, this information is part of the desired result. The optimal number of clusters should then be determined through cluster validation. This process is helpful in cluster determination and the verification of the whole clustering process, as there is no universal 'best' clustering process. Configuration of different methods in each step of the clustering process can result in different partitions.

1. Database preparation: The present model enables archetype development using a variety of databases such as building permits and municipal records. In case such data is not available, the model utilizes the EGHD and attempts to uncover trends in the records. Those datasets generally have different schemas. Data wrangling and normalization are the mandatory processes to ensure the model receives a workable database for clustering.
2. Variable selection and data pre-processing: the optimal variables are selected, and then the data is pre-processed in preparation for clustering. The choice of variables is crucial to ensure the resulting archetypes identified are representative. Including irrelevant variables does not improve the results; instead, it can result in noisier data and undermine the algorithm's search to partition the data into similar clusters (Ghiassi & Mahdavi, 2017). Furthermore, more variables mean longer computational time. The most straightforward approach is to conduct a literature review to find the variables most used in defining archetypes. This list of variables can be used directly in clustering. The frequency of variables used for cluster analysis in the archetype development studies found in the literature is shown in Figure 4. The blue striped column considers all studies and the red solid columns only consider studies related to residential buildings. Results indicate that the geometry parameters are the most common variables for archetype development.

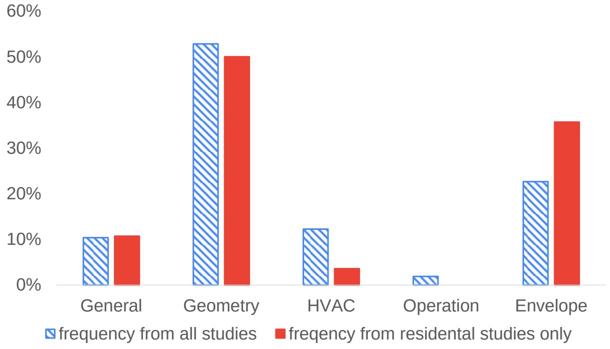

Figure 4. Frequency of features used in previous studies for clustering analysis

3. Clustering: The number of clusters (K) is a frequently noted problem when using K-means, as it is a required input for the algorithm. The optimal value for K should be balanced between being large enough to meaningfully represent the data but not too large that it will represent unnecessary associations and small enough to optimize computational efficiency. This is typically overcome by running the clustering algorithm for various values of K, and the resulting validity measures are plotted against their k values. The optimal K is then represented as either a maximum, minimum, or 'elbow' depending on the validation used (Everitt et al., 2011). The initialization of centroids in the first step of the k-means process significantly influences the results. There is no guarantee that the algorithm will converge to a global minimum if the initial points are not selected carefully. This can be mitigated by excluding outliers from being chosen as the initial centroids, testing multiple starting points, or using other methods such as hierarchical clustering to determine what objects should be the initial centroids. In addition, intermediate steps can be taken to create a more robust list. A common approach in clustering for archetype identification studies is regression analysis. The goal of regression analysis is to define a set of variables that correlate to energy consumption (Gao & Malkawi, 2014; Famuyibo et al., 2012; De Jaeger et al., 2020)
4. Interpretation and validation: The model parses EGHD to identify the closest match with the centroids. Each identified model will be selected as an archetype representing the entire homes in the cluster.

Statistical modeling

The objective is to develop a baseline model to benchmark the characteristics of existing Canadian houses at the regional level. The model will represent the collective performance of multiple house types, vintages, and construction characteristics and could be used to evaluate the impact of scenarios when implemented at a large scale. For this purpose, the statistical analysis approach developed by Swan et al. (2010) was used to create a subset of EGHD that statistically represents the Canadian housing stock. To address the inherent bias in the EGHD, Swan et al. used the Survey of Household Energy Use data (SHEU) 2005. SHEU is an extensive survey of Canadian household energy use conducted by Statistics Canada for NRCan since 1993. SHEU surveys the dwelling characteristics, usage of appliances and other energy-consuming products, energy efficiency characteristics, and energy consumption of households. Statistics Canada randomly selects houses based on the regional population. While Statistics Canada's selection methods are unbiased, the level of detail within the SHEU database is insufficient to generate energy models.

Swan et al. (2009) developed a methodology to selectively extract data from the EGHD based on the statistical representation of key parameters of the Canadian housing stock that was obtained from SHEU. At that time, SHEU divided the Canadian housing stock into five regions (i.e., Atlantic, Quebec, Ontario, Prairies, and British Columbia). They selected the samples using the following distribution parameters: house type, region, vintage, storeys, living space floor area, space heating energy source, and hot water energy source.

The present study builds a framework to randomly select a subset of EGHD according to the statistical distribution of housing stock characteristics from statistically valid samples such as SHEU. The latest version of SHEU released the data collected in 2015 was used as the basis for this study (NRCan, 2020).

An algorithm was developed in Python to process, compare, and select/discard HOT2000 files for the baseline model. The main steps of the algorithm include:

1. Define the total number of archetypes in the baseline model
2. Establish the housing characteristics distribution: The model uses the regional distribution of SHEU-2015 to define the type of archetypes that should be sampled from the EGHD. The distribution of archetypes characteristics, including house type, storeys, vintage, floor area, and energy sources for space and hot water heating, should match the SHEU-2015.
3. Data import: At the beginning of the process, the algorithm imports the file ID, region, and the main characteristics of the records stored in the database of upgraded HOT2000 files. The records were randomly shuffled to eliminate any underlying bias that could affect the results.
4. Archetype sampling: The algorithm uses a loop to parse the imported data and compare the characteristics of the HOT2000 models with the desired characteristics identified in step 2. For each file, if the house characteristics match the desired parameters, the file was added to the baseline model. A Python dictionary was used to track the attributes of the selected model to ensure the number of archetypes would not exceed the distribution targets defined in step 2. If the characteristics HOT2000 model do not match the desired characteristics identified in step 2, the file is discarded, and the algorithm proceeds to the next record in the database.

The algorithm was repeated with multiple seeds for the random shuffle to ensure the algorithm was able to reach a sample that contained all forms of desired homes. As Swan et al. (2010) described, the resulting dataset would statistically represent the housing stock.

Engineering model

The engineering model aims to use the archetype data and predict the impact of various scenarios at the housing stock level.

- Energy modeling: The goal of the energy modeling in a bottom-up stock model is to evaluate the new technologies by accounting for the specific heat transfer and thermodynamic relationships. The present model uses HOT2000 as its simulation engine. The data pipeline described in the previous section enables the automatic upgrade of the models to the latest version of HOT2000. As discussed earlier, NRCan developed HTAP to study various design scenarios and identify optimal solutions using a grid search.
- Hourly calculations: In the next step, an algorithm will be developed to automatically create an equivalent version of the house model in an hourly energy modeling software (e.g., EnergyPlus). The hourly analysis would enable a more precise calculation of the cooling loads, assessment of advanced HVAC systems, and grid impact analysis based on the time of use for electricity.
- Surrogate modeling: While energy modeling is the most common method for designing and optimizing buildings, surrogate modeling has gained growing attention for building optimization problems. A surrogate model uses measured data to develop a mathematical approximation of the actual system. A surrogate model can significantly cut the computational cost of the engineering model and can be a viable option where a detailed description of building characteristics is not available (Evins 2013, Bamdad 2020).
- Carbon calculation: The carbon footprint of buildings includes operational and embodied. The operational part includes carbon emissions associated with the energy use for the day-to-day operation. Embodied carbon includes carbon emissions associated with the production, transportation, installation, and demolition of building materials. Once the energy use of an

archetype is determined, the calculation of operational carbon emissions is relatively straightforward. For each retrofit scenario, the embodied carbon of all retrofit materials can be aggregated and compared against alternative scenarios. This would ensure the decarbonization scenarios would address both aspects of building carbon footprint.

Discussion

A set of modules were developed in Python to create a stock model based on the statistical analysis approach and cluster analysis to represent the housing stock at the national and municipal levels. The model enables a comprehensive analysis of the current state of the housing stock and evaluates the impact of decarbonization scenarios. The following section provides a summary of statistics derived from the stock model.

Case study

The stock representation model developed in this study produced 5970 archetypes to represent the housing stock in seven regions: British Columbia (BC), Alberta (AB), Manitoba & Saskatchewan (PR), Ontario (ON), Quebec (QC), Atlantic Provinces (AT), and Northern Territories. These regions matched the regional definition of SHEU-2015. The distribution of house types, vintage, storeys, floor area, and energy sources for space and hot water heating was adopted from statistical data according to the following guiding principles:

1. Provinces: The model created a subset of EGHD according to the statistical distribution of 2015 (NRCan, 2020).
   a. The user selects the total number of desired archetypes in the baseline model as an input parameter.
   b. The distribution of archetypes by house type (detached and attached), number of stories (one, two, three), and floor area (<56 m^2, 56 to 93 m^2, 93 to 139 m^2, 139 to 186 m^2, 186 to 232 m^2, and > 232 m^2), as well as energy sources for space heating (Electricity, Natural gas, Heating oil, Wood, Propane) and hot water heating (Electricity, Natural gas, Oil, Propane), was extracted from SHEU.
   c. The number of archetypes was determined by multiplying the total number of archetypes and the percentage of each parameter defined in the previous step.
2. Territories: SHEU has no information regarding the households in Territories. Due to the harsh climate, northern houses have unique features not common in other parts of the country. It is necessary to include archetypes from the actual houses in the Territories. The following rules were used to guide the archetype selection algorithm for the northern houses:
   a. The ratio of the number of northern households to the total number of Canadian houses is relatively small. Therefore, using this ratio to decide the number of archetypes results in a small number of models for Territories. To avoid such a problem, a hundred archetypes were selected as the target for the Territories.
   b. The distribution of households by house type in Yukon, Northwest Territories, and Nunavut was extracted from the Census 2016 (Statistics Canada, 2017).
   c. The distribution of households by vintage and energy sources for space heating was extracted from the National Energy Use database (NEUD) – Comprehensive Energy Use Database (NRCan, 2018c).
   d. The remaining selection parameters (i.e., number of storeys, floor area, and energy source for hot water heating) were not prescribed for the northern archetypes and were free to float.

Since the archetypes are sampled from energy audit records, they contain various information about the size and geometry, construction characteristics, and HVAC systems. This information is not available in any other survey, and the archetypes will be a valuable source for benchmarking the current state of the housing stock. The summary distribution of some parameters is provided in Table 1.

Table 1. Summary of the characteristics of the archetype by region

	Region					
	BC	AB	PR	QC	ON	AT
Basement Presence (%)	48	95	95	95	92	52
Crawl Presence (%)	33	6	5	5	12	41
Slab Presence (%)	29	4	3	3	3	7
Average Heated Floor Area (m^2)	206	221	198	205	214	156
Average Window Area (m^2)	29	22	16	24	21	17
Rural (%)	13	4	29	18	13	31
Urban (%)	87	96	71	82	87	69
Glazing Ratio (%)	16	14	12	17	13	13
Average ACH @50pa	7.5	4.2	4.8	5.7	7.9	8.3

The first three rows indicate the presence of different forms of foundations in the archetypes by region. For example, a basement is the most popular option for houses across Canada except in BC and AT, where crawl space and slab on grade are other common forms of foundation construction. The average floor area of existing homes in AT is about 25 percent lower than in other regions. The homes in BC and QC have a larger window-to-wall ratio than in other regions.

The distribution of air-tightness of the archetypes is shown in Figure 5. This data was collected using a blower-door test during the energy audits. The results show a positively skewed distribution for the air-tightness of the archetypes. While a relatively low air change rate is more frequent in the model, the model captured less frequent leaky envelopes. The representation of those constructions is essential for a comprehensive impact analysis of energy efficiency scenarios.

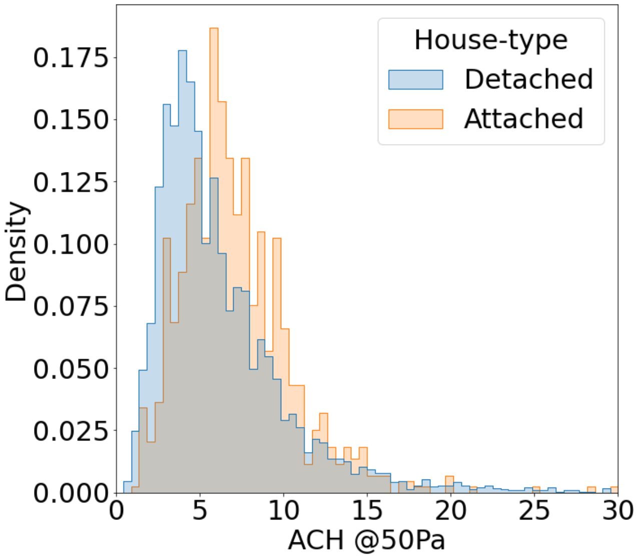

Figure 5. Distribution of air-tightness of the archetypes

The distribution of the window-to-wall (WWR) ratio of the archetypes is shown in Figure 6. The results show that the WWR has a normal distribution. The mean WWR is about 12 percent, and the archetypes represent a variety of forms, such as highly glazed and highly opaque envelopes.

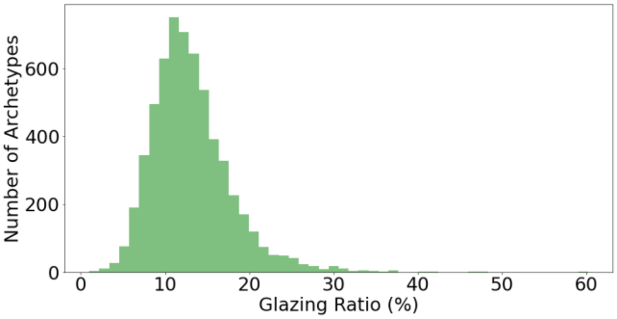

Figure 6. Distribution of window to wall ratio of the archetypes

The above-grade wall insulation values by house type are shown in Figure 7. The insulation values of the above-grade wall for both house types are in the same range, while detached homes tend to have more outliers.

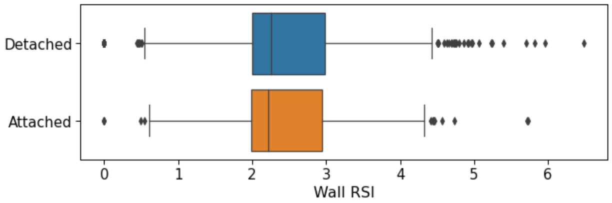

Figure 7. Distribution of above-grade wall insulation

HTAP was used to estimate the current energy performance of the archetypes. The results of annual energy use are shown in Figure 8. The blue histogram represents the detached, and the orange histogram represents the attached (i.e., semi-attached and row) house types. As expected, the attached house types tend to use less energy than the detached forms. The annual energy use data was extrapolated to the stock level and compared with other estimations of the energy use of the Canadian housing stock. The results show the estimation of the current model is in a similar range to the household energy consumption estimation of Statistics Canada.

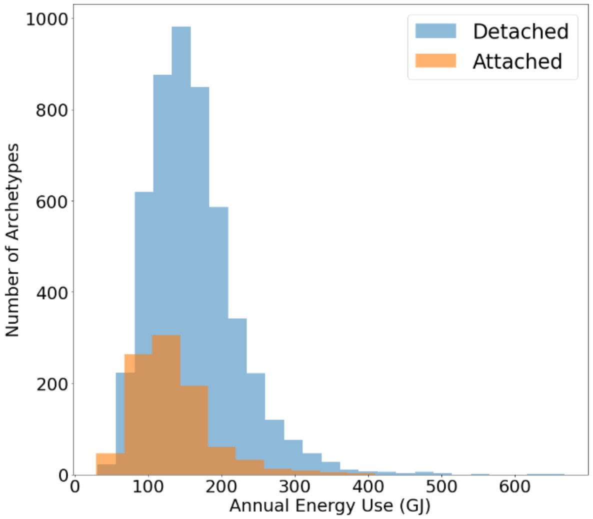

Figure 8. Distribution of annual energy use of the archetypes

Conclusion

This article presents a semi-automated pipeline for the rapid development of stock models at provincial and municipal levels. The proposed framework utilizes a hybrid approach.

- Stock representation: For each region or municipality, the model extracts the existing energy models for the actual houses, upgrade them to the latest version of HOT2000 to enable XML parsing, and creates archetypes using a statistical method and cluster analysis.

- Engineering model: An automation and optimization model uses the archetypes to analyze the impact of retrofit scenarios for existing houses. HOT2000 is used for batch simulations. A translator creates an equivalent version of the model in EnergyPlus. The hourly model is used to estimate the cooling loads and peak demand. The energy modeling results are used to train a surrogate model. The surrogate model aims to provide a quick estimation of energy use through an API.

Stock representation and the engineering model have been developed. Results of the stock representation model were presented and discussed. The model provides information regarding the construction characteristics of existing houses. Future work includes the integration of the hourly model into the framework.

Acknowledgment

The author is grateful for the funding provided by Natural Resources Canada through the Office of Energy Research and Development (OERD).

References

Asaee, S. R., Sharafian, A., Herrera, O. E., Blomerus, P., & Mérida, W. (2018). Housing stock in cold-climate countries: Conversion challenges for net zero emission buildings. *Applied Energy*, 217, 88-100.

Ali, U., Shamsi, M. H., Hoare, C., Mangina, E., & O'Donnell, J. (2019). A data-driven approach for multi-scale building archetypes development. *Energy and Buildings*, 202.

Lara, R. A., Pernigotto, G., Cappelletti, F., Romagnoni, P., & Gasparella, A. (2015). Energy audit of schools by means of cluster analysis. *Energy and Buildings*, 95, 160-171.

Aydinalp, M., Ferguson, A., Fung, A., Ugursal, V.I. (2001). *Energuide for houses database analysis*, Halifax, Canada.

Bamdad, K., Cholette, M. E., & Bell, J. (2020). Building energy optimization using surrogate model and active sampling. *Journal of Building Performance Simulation*, 13(6), 760-776.

De Jaeger, I., Reynders, G., Callebaut, C., & Saelens, D. (2020). A building clustering approach for urban energy simulations. *Energy and Buildings*, 208.

Deb, C., & Lee, S. E. (2018). Determining key variables influencing energy consumption in office buildings through cluster analysis of pre- and post-retrofit building data. *Energy and Buildings*, 159, 228-245.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*. John Wiley & Sons, Incorporated.

Evins, R. (2013). A review of computational optimisation methods applied to sustainable building design. *Renewable & sustainable energy reviews*, 22, 230-245.

Famuyibo, A. A., Duffy, A., & Strachan, P. (2012). Developing archetypes for domestic dwellings—An Irish case study. *Energy and Buildings*, 50, 150-157.

Gangolells, M., Casals, M., Ferré-Bigorra, J., Forcada, N., Macarulla, M., Gaspar, K., & Tejedor, B. (2020, January). Office representatives for cost-optimal energy retrofitting analysis: A novel approach using cluster analysis of energy performance certificate databases. *Energy and Buildings*, 206.

Gao, X., & Malkawi, A. (2014). A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm. *Energy and Buildings*, 84, 607-616.

Ghiassi, N., & Mahdavi, A. (2017). Reductive bottom-up urban energy computing supported by multivariate cluster analysis. *Energy and Buildings*, 144, 372-386.

Government of Canada. (2016). Pan-Canadian framework on clean growth and climate change.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999, September). Data Clustering: A Review. *ACM Computing Surveys*, 31(3).

Li, X., Yao, R., Liu, M., Costanzo, V., Yu, W., Wang, W., Li, B. (2018). Developing urban residential reference buildings using clustering analysis of satellite images. *Energy and Buildings*, 169, 417-429.

Mahon, M. (2022). What is pywinauto. https://pywinauto.readthedocs.io/en/latest/

NRCan, EnerGuide in Canada, (2018a). http://www.nrcan.gc.ca/energy/products/energuide/12523.

NRCan, R-2000: environmentally friendly homes, (2018b). https://www.nrcan.gc.ca/energy/efficiency/homes/20575 (accessed March 23, 2018).

NRCan, (2020). Survey of Household Energy Use (SHEU-2015), Ottawa, Canada.

Santamouris, M., Mihalakakou, G., Patargias, P., Gaitani, N., Sfakianaki, K., Papaglastra, M., ... & Zerefos, S. (2007). Using intelligent clustering techniques to classify the energy performance of school buildings. *Energy and buildings*, 39(1), 45-51.

Swan, L. G., Ugursal, V. I., & Beausoleil-Morrison, I. (2009). A database of house descriptions representative of the Canadian housing stock for coupling to building energy performance simulation. *Journal of Building Performance Simulation*, 2(2), 75-84.

Tardioli, G., Kerrigan, R., Oates, M., O'Donnell, J., & Finn, D. (2018). Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach. *Building and Environment*, 140, 90-106.