Energy Audit Data Scraping for Rapid Building Energy Modeling – A Case Study

Sudipta Debnath[1], Kuljeet Singh Grewal[1]
[1]Future Urban and Energy Lab for Sustainability (FUEL-S), Faculty of Sustainable Design Engineering, University of Prince Edward Island 550 University Ave, Charlottetown, PE, Canada

Abstract

This research presents a novel approach to energy modeling by leveraging data scraping techniques to extract insulation and air tightness values from EnerGuide energy audit reports. The extracted data is integrated into Rhino-Grasshopper for automated energy modeling, enhancing the accuracy and efficiency of simulations. The study focuses on the case study location of Georgetown, Prince Edward Island (PEI), Canada, analyzing 170 buildings to understand the insulation characteristics over different construction years. The results highlight the importance of building-specific characteristics in energy modeling and underscore the need for data scraping to inform energy-efficient building designs and urban energy management.

1. Introduction

In today's world, optimizing energy consumption in buildings has become a vital initiative compelled by both environmental concerns and economic imperatives (Mutani et al., 2018). Building energy modeling is pivotal in the process of addressing contemporary cities and neighborhood's ever-growing energy demands and sustainability necessities (Sayfiddinov et al., 2020). By the generation of precise and insightful models, a more profound comprehension of energy consumption can be attained, therefore contributing to advanced sustainable and efficient urban landscapes (Best et al., 2015). To achieve this, the identification and retrieval of diverse and pertinent data sources are of foremost importance (Hong et al., 2020).

Energy audits play a paramount part in the energy retrofit of buildings, while data collection is crucial for accurate and reliable energy modeling (Congiu et al., 2022). A range of studies draw attention to the significances of data collection and analysis in energy retrofitting. (Ham & Golparvar-Fard, 2012) emphasize the necessity of both data collection and data analysis in energy performance modeling, specifically highlighting the use of digital and thermal imagery for 3D modeling. (Maricar & Jamal, 2005) underline the vital role of energy audit data in industrial energy simulation, accentuating its prospect to drive efficiency advancements and cost savings in complex industrial processes. (Kong et al., 2014) highlight the necessity of energy audit data in existing building energy simulations, depicting its value in identifications of opportunities for energy conservation and performance optimization in the built environment.

Particularly, for building energy modeling, non-geometric data such as envelope characteristics, energy systems parameters, air tightness, etc. are critical to identify and use. Further, problems become more prominent when energy modeling of multiple buildings is involved. Energy audits are one of the useful sources for building-related information extraction. For instance, in Canada EnerGuide (Francoeur et al., 2006) is a popular rating system to evaluate the energy performance of buildings estimating building insulation, energy systems, fenestration characteristics, and air tightness. Furthermore, a rating of annual energy consumption (typically in GJ/y) is assigned based on energy performance estimation. However, since the EnerGuide audit provides annual energy data estimation, therefore, building-related characteristics data can be used to set up a detailed energy model to develop several scenarios such as prediction of hourly energy consumption, evaluating the impact of an improved building envelop, net-zero potential, energy management, etc. The data from energy audits can be extracted and fed to energy modeling software such as EnergyPlus and Grasshopper for building specific detailed energy simulations. For this purpose, advanced techniques such as data scraping can be used to save modeling efforts, avoid manual work and to automate the energy modeling process.

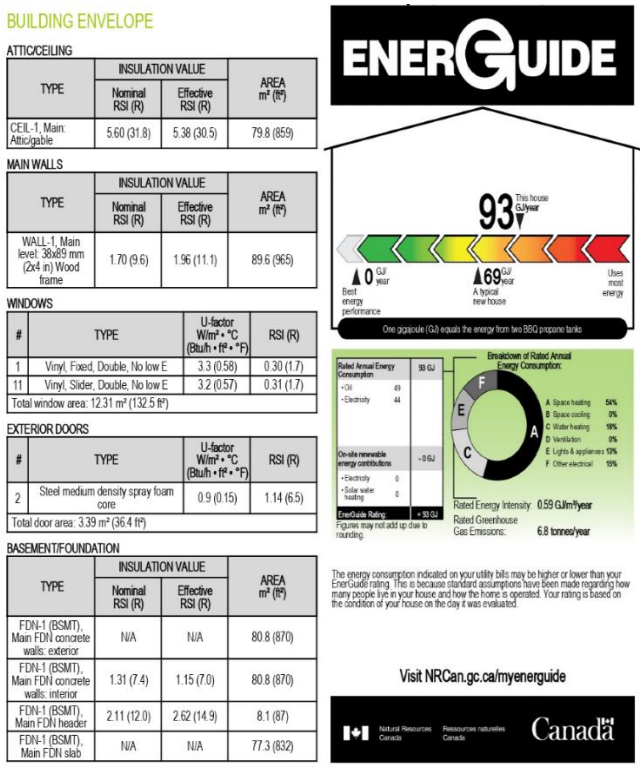
(Grover et al., 2020) developed an Application Programming Interface (API) to scrape data from the BAS (Building Automation System) of a mid-size pilot building, which included environmental conditions and equipment activity. (Nichiforov et al., 2020) applied the Matrix Profile (MP) technique to energy time series data for load modeling in large commercial buildings, enabling trend identification and fault anticipation. (C. Yang et al., 2018) used data mining-based methodology to develop data-

driven models for estimating energy consumption in building energy management systems (BEMS), achieving promising accuracy. (S. Yang et al., 2018) proposed a web crawler and information gathering algorithm to collect physical properties information of building materials for simulation tools, overcoming limitations of embedded databases. (Mantha et al., 2016) and (Himeur et al., 2020) both deliver taxonomies of data types and collection methods for building energy monitoring and performance simulation. (Mantha et al., 2016) focus on variety of data types and the review of collection methods. (Himeur et al., 2020) survey and compares existing building power consumption datasets. (Ahmad et al., 2018) and (de Wilde et al., 2013) delve into typical strategies for predicting building energy demand and training automated data analysis tools. (Ahmad et al., 2018) classify energy-demand projection models into data-driven, physics-based, large-scale, and hybrid approaches, while (de Wilde et al., 2013) study the fusion of building simulation approaches into energy management practice. Energy audit data scraping facilitates the process of assembling elementary information from diverse sources which include utility bills and sensor data (Capozzoli et al., 2016). This scraping employs technologies like web scraping and machine learning while such automated approaches speed up the collection and analysis of extensive datasets (Hopf et al., 2018). It becomes easier to pinpoint opportunities for energy-saving actions, for instance improvement of heating, ventilation and air conditioning (HVAC) systems or upgrading insulations which ultimately leads to maximizing the return on investment in energy efficiency measures (Zucker et al., 2014).

The present research has spotlighted the critical role of data scraping in the extraction of relevant information concerning energy and environmental performance of buildings. Specifically, Python-based data scraping techniques to collect and analyze RSI-values (insulation) and AIR50P (air leakage at 50 pascals - quantifies air infiltration through any building envelope at 50 Pa pressure difference from the exterior) of different building envelopes from energy audit reports are employed. The method can be employed with the building owner's consent if access to energy audit information is not available from Natural Resources Canada (NRCan). Also, the data available from NRCan for research purposes is anonymized, hiding the building address which makes it challenging to utilize the audit results for detailed specific neighborhood/building energy modeling. Hence, the net-zero assessment project for the case study location of Georgetown, Prince Edward Island (PEI) has raised the need for a data scraping method to extract and use data for the town's building energy modeling. Upon validating the data scraping efficacy, the research further shows the integration of this scraping data into sophisticated energy modeling tools like Rhino-Grasshopper to automate the energy modeling process. Further, this work analyzed the building envelope data for 170 buildings based on the year of construction to evaluate the need to use energy audit data rather than assuming insulation values based on the year of construction. Ultimately, the research forms the prospect to develop tools that are instrumental in the optimization of energy utilization, mitigation of environmental footprints, and streamlining of urban energy management.

2. Method

EnerGuide is a Canadian program aiding building owners in enhancing energy efficiency (Francoeur et al., 2006). By visual evaluations and blower door tests, EnerGuide auditors pinpoint areas of energy loss in insulation, HVAC systems, windows/doors, and appliances. Building owners receive detailed reports with upgrade recommendations, often subsidized by government programs to save money and reduce environmental impact. Therefore, current research values EnerGuide audit information sheets as crucial data sources for energy modeling, including thermophysical properties like RSI-values and U-values, for in-depth analysis.



BUILDING ENVELOPE

ATTIC/CEILING

TYPE	INSULATION VALUE Nominal RSI (R)	INSULATION VALUE Effective RSI (R)	AREA m² (ft²)
CEIL-1, Main: Attic/gable	5.60 (31.8)	5.38 (30.5)	79.8 (859)

MAIN WALLS

TYPE	INSULATION VALUE Nominal RSI (R)	INSULATION VALUE Effective RSI (R)	AREA m² (ft²)
WALL-1, Main level: 38x89 mm (2x4 in) Wood frame	1.70 (9.6)	1.96 (11.1)	89.6 (965)

WINDOWS

#	TYPE	U-factor W/m² • °C (Btu/h • ft² • °F)	RSI (R)
1	Vinyl, Fixed, Double, No low E	3.3 (0.58)	0.30 (1.7)
11	Vinyl, Slider, Double, No low E	3.2 (0.57)	0.31 (1.7)
Total window area: 12.31 m² (132.5 ft²)			

EXTERIOR DOORS

#	TYPE	U-factor W/m² • °C (Btu/h • ft² • °F)	RSI (R)
2	Steel medium density spray foam core	0.9 (0.15)	1.14 (6.5)
Total door area: 3.39 m² (36.4 ft²)			

BASEMENT/FOUNDATION

TYPE	INSULATION VALUE Nominal RSI (R)	INSULATION VALUE Effective RSI (R)	AREA m² (ft²)
FDN-1 (BSMT), Main FDN concrete walls: exterior	N/A	N/A	80.8 (870)
FDN-1 (BSMT), Main FDN concrete walls: interior	1.31 (7.4)	1.15 (7.0)	80.8 (870)
FDN-1 (BSMT), Main FDN header	2.11 (12.0)	2.62 (14.9)	8.1 (87)
FDN-1 (BSMT), Main FDN slab	N/A	N/A	77.3 (832)

Figure 1: EnerGuide audit information sheets.

A Python script is created to import the energy audit sheets, resembling Figure 1, and extract RSI (R) and

AIR50P values from the EnerGuide audit information Sheets typically stored in PDF format. As shown in Figure 2, the algorithm pulls textual contents implanted within PDF documents throughout "pdfminer.high_level.extract_text" function supplied by the pdfminer library in Python. This process is vital as it enables access to the underlying data for further analysis. The text is segmented into individual lines upon extraction which facilitates granular analysis and processing. This action enables algorithm to parse the content efficiently and line by line.

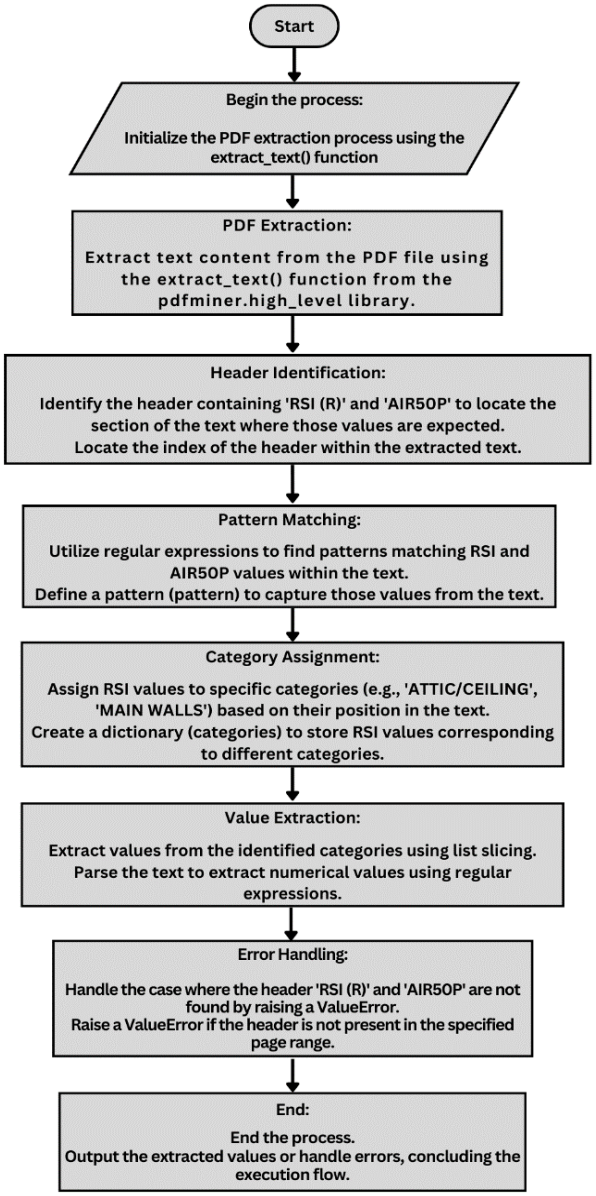


Figure 2: Flowchart of Python Script for Scraping RSI-values and AIR50P from EnerGuide PDFs

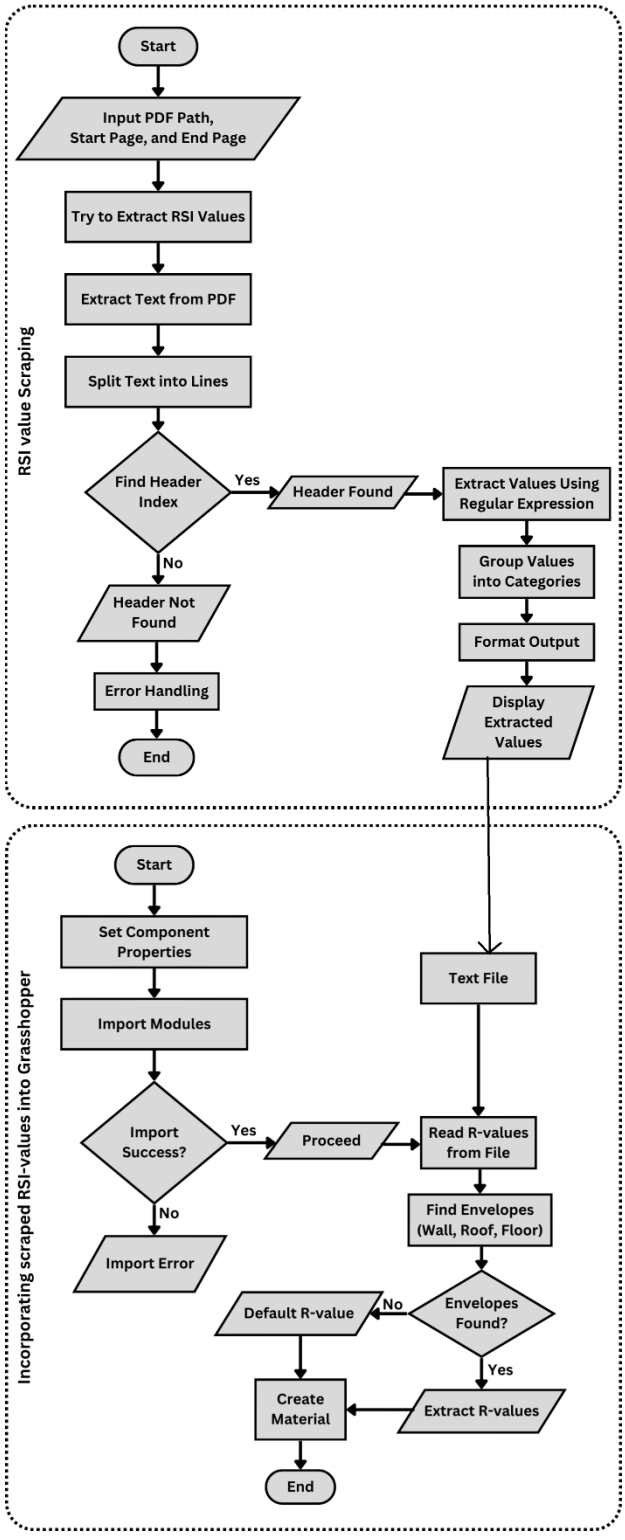


Figure 3: Integration of scraped RSI-values from EnerGuide PDFs into Grasshopper using the Honeybee plugin for energy modeling, including parsing text files and informing Grasshopper components.

Further, it uses regular expressions to search for a specific header pattern, 'RSI (R)' and 'AIR50P', which provides a powerful mechanism for pattern matching within textual data, aiding in precisely identifying relevant sections within the document. Likewise, the code specifies numerical values corresponding to RSI (R) ratings via regular expression pattern matching.

Moreover, the script categorizes the scraped insulation values based on predefined sections corresponding to distinct building areas like attic/ceiling, main walls, etc. This categorization enhances data organization and facilitates subsequent analysis. The extracted RSI (R) values and their respective categories are formatted into a structured output format, which ensures clarity, and readability and makes them easily interpretable for further processing or visualization. Robust error handling mechanisms are incorporated to address potential issues, such as the absence of an 'RSI (R)' header within the specified page range, enhancing the script's reliability and resilience and ensuring smooth execution even in adverse scenarios.

For detailed energy modeling, it is crucial to integrate the scraped insulation properties into the Honeybee plugin of Grasshopper, a visual algorithm editor tightly incorporated with Rhino's 3D modeling tool. As such, the GhPython scripts is designed to utilize RSI (R) values extracted from EnerGuide PDFs, provided in a text file, to inform the Grasshopper component's inputs. It parses the text file to locate RSI-values associated with specific building envelopes, converting them to float data for processing. Subsequently, material properties are defined, and an EnergyMaterialNoMass object is instantiated based on the extracted RSI values. Figure 3 depicts the detailed methodology highlighting the seamless integration of scraped RSI values from EnerGuide PDFs into Rhino-Grasshopper, enabling efficient energy modeling tasks.

3. Results and Discussion

Implementation of the proposed method rendered promising outcomes in terms of extracting and integrating RSI values from EnerGuide energy audit PDFs into the energy modeling framework. The Python script efficaciously scraped relevant information from the building audit information sheets, extracting RSI values associated with building envelope types along with attic/ceiling, main walls, windows, exterior doors, and basement/foundation. Table 1 gives the extracted RSI values categorized with the aid of building envelope kinds. For instance, for the first audit sample, attic/ceiling insulation values are listed as 5.60 $m^2 \cdot K/W$ and 5.37 $m^2 \cdot K/W$, main walls have values of 1.70 $m^2 \cdot K/W$ and 1.95 $m^2 \cdot K/W$, windows have a value of 1.7 $m^2 \cdot K/W$, and so on. The algorithm is successfully validated to yield error-free extraction of energy audit data. Furthermore, the integration of scraped RSI values into Grasshopper-Honeybee validated seamless interoperability between data extraction and energy modeling duties. Figure 4 visually articulates the incorporation of extracted RSI-values into Grasshopper's Honeybee environment, offering perception into the building's materials, surfaces, and creation details. This integration upon inputting geometric data allows the simulation tool to make use of the scraping data for computing the building's energy use and performance metrics under diverse working situations.

Successful integration of scraped RSI-values into the modeling framework complements the accuracy and efficiency of energy simulations, allowing designers, engineers, and urban planners to make informed decisions concerning building-specific retrofitting strategies, and urban power management. By leveraging scraped attributes from audit sheets, the energy modeling approach becomes extra robust and easy to implement, main to optimized power utilization and decreased environmental impacts in city environments.

Table 1: Scraping data from four different building audits

Audit Samples	Building Envelope Types	Nominal (RSI) $m^2 \cdot K/W$	Effective (RSI) $m^2 \cdot K/W$
1	Attic/ceiling	5.6	5.38
	Main walls	1.7	1.96
	Windows		0.3
	Exterior doors		1.14
	Basement/ foundation	1.31, 2.11	1.15, 2.62
2	Attic/ceiling	7.22	6.76
	Main walls	3.34	3
	Windows		0.3
	Exterior doors		1.14
	Basement/ foundation	1.06, 3.34	1.14, 3.68
3	Attic/ceiling	5.6	5.31
	Main walls	1.7	1.95
	Windows		0.3
	Exterior doors		1.14
	Basement/ foundation		0.73
4	Attic/ceiling	7.22	6.71
	Main walls	3.34	2.98
	Windows	0.3	0.31
	Exterior doors		1.14
	Basement/	1.06, 1.41	1.14, 2.01

foundation

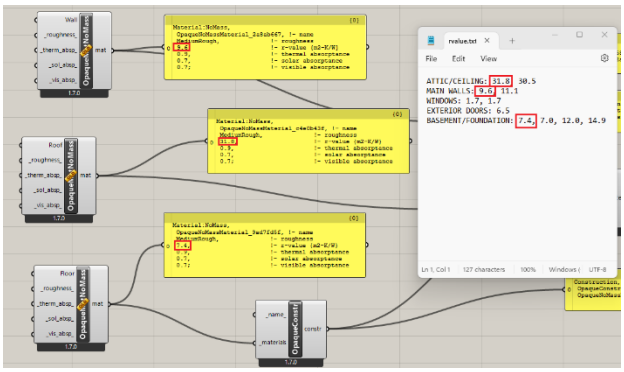


Figure 4: Incorporation of extracted RSI-values into Grasshopper's Honeybee environment

Further, this research analyzed the building insulation characteristics identified from EnerGuide audit data for the case study location of Georgetown, Prince Edward Island (PEI), Canada. A total of 170 buildings are analyzed in the work. Figure 5 depicts records on the RSI-values of ceilings, foundations, walls, and air infiltration (AIR50P) for different buildings, alongside their year of construction. RSI-value, which is a measure of thermal resistance used in construction and production, indicates extra insulating capability with higher values. The RSI values of the buildings primarily involving houses are examinable based on their construction years:

Buildings constructed before 1930:

These buildings have the lowest RSI-values, typically around 2-6 m²·K/W for ceilings/roof, 2 m²·K/W for foundations, and 2 m²·K/W for walls. Air infiltration (AIR50P) is maximum (4-30), indicating insufficient air sealing. These low RSI values propose that buildings from this period have insufficient insulation and air sealing that can lead to big heat loss and more power intake. Furthermore, some older buildings may exhibit especially excessive RSI values despite their age, indicating renovations or enhancements to improve insulation. For instance, there is probably some buildings with ceiling RSI-values (7 m²·K/W) greater than 6 m²·K/W, foundations with RSI-values (2.5 and 3 m²·K/W) exceeding 2 m²·K/W, or walls with RSI-values (2.5, 3 and 3.5 m²·K/W) more than 2 m²·K/W. Conversely, a few buildings can also have been poorly insulated, leading to lower RSI values. Furthermore, some houses have significantly better airtightness (AIR50P <10) indicating proper sealing of gaps and cracks.

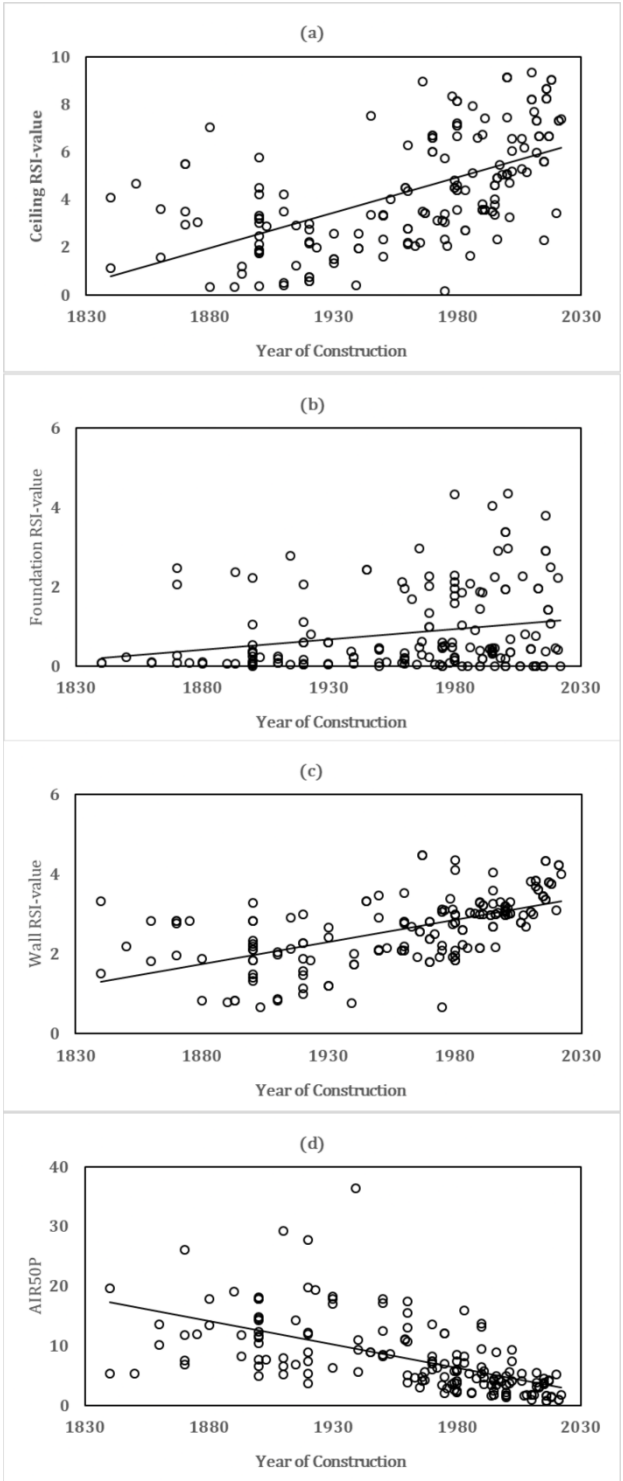


Figure 5: Comparison of Building RSI-Values (a,b,c) and Air Sealing (d) by construction year

Buildings built between 1930-1980:

These buildings display a moderate improvement in RSI-values, but they are nevertheless rather low. For example,

ceilings have RSI-values of around 0.5-8 m²·K/W, foundations have RSI-values of 0.4-3 m²·K/W, and walls have RSI-values of 1-4 m²·K/W. Air tightness (AIR50P) levels slightly decrease, implying better air sealing. The insulation and air sealing in those buildings are nonetheless inadequate, even though they may be higher than the ones constructed earlier than 1930. Despite being constructed during this period, some buildings have exceptionally low RSI values, indicating poor insulation or lack of upgrades. Such outliers might be attributed to neglect or deferred maintenance.

Buildings constructed after 1980:

These buildings have significantly higher RSI values than the sooner ones. Ceilings have RSI-values of 2-10 m²·K/W, foundations have RSI-values of 0.2-5 m²·K/W, and walls have RSI-values of 2-5 m²·K/W. AIR50P is below 10 to near 0, indicating appropriate air sealing. These homes have a great deal in terms of better insulation and air sealing, essential to decreased heat loss and lower energy consumption.

Nevertheless, it is worth noting there is a range of insulation and air tightness values even if the buildings are constructed in the same year. It can be due to non-uniformity in construction practices as PEI adopted the National Building Code around 2020. Therefore, air tightness and wall insulation values are converging to a narrow range in the recently constructed buildings. Hence, for accurate energy modeling of the neighborhood in the case study location, a building-specific characteristic must be defined avoiding the use of generic insulation values based on the year of construction. For this reason, data scraping can be a useful technique for engaging community members in sharing energy audits ensuring privacy protections. Further, these audits can be processed rapidly utilizing developed data scaping and integration algorithms saving energy modeling time and cost.

4. Conclusion

This study outlines a unique methodology for energy modeling that involves implications of the scraping of energy audit data and their integration for additional energy modeling. A text-based scraping strategy is developed to extract insulation and air tightness values from the energy audits. The scraped data is further supplied for energy modeling in Rhino-Grasshopper to facilitate automated data implementation. The proposed method is successfully validated by extracting various building envelope values for different buildings. The work also analyzed the energy audit data of 170 buildings in Georgetown PEI, Canada. The study reveals the importance of building-specific characteristics in energy modeling for Georgetown, PEI, noting that while insulation and air tightness have improved over time, there remains variability even among buildings of the same age due to non-uniform construction practices. This variability makes it challenging to assume building envelope characteristics based on the year of construction while setting up the neighborhood energy model. While the current focus lies on individual buildings, future endeavors aim to extend this approach to multiple buildings. This research focuses on the practical prospect of data scraping in promoting energy-efficient building designs and urban energy management processes.

5. Future Work

In future work, data scraping and integration with Grasshopper for energy modeling will be undertaken for neighborhood energy modeling involving multiple buildings. A high-fidelity illustration of community topography by way of leveraging numerous data sources, which includes Light Detection and Ranging (LiDAR) and drone imagery will be utilized in integration with data scraping automation. A nuanced understanding of 3-dimensional building structures and their surroundings is achieved by combining these datasets. Researchers and urban planners can examine and optimize energy consumption, examine renewable energy integration, and layout sustainable neighborhoods via meticulous facts fusion. The future workflow process information to generate accurate representations using Arc GIS Pro, Rhino-Grasshopper, and Python facilitating energy modeling, sustainability assessments, and renewable energy implementation. The intended use of these developed workflows is to rapidly create energy modeling of neighborhoods to evaluate their potential to meet net-zero targets.

Acknowledgment

The authors duly acknowledge PEI Energy Corporation for funding support, which has been instrumental in driving forward this ongoing project towards its completion.

References

Ahmad, T., Chen, H., Guo, Y., & Wang, J. (2018). A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review. *Energy and Buildings*, *165*, 301–320. https://doi.org/10.1016/j.enbuild.2018.01.017

Best, R. E., Flager, F., & Lepech, M. D. (2015). Modeling and optimization of building mix and energy supply technology for urban districts. *Applied Energy*, *159*, 161–177. https://doi.org/10.1016/j.apenergy.2015.08.076

Canada, G. of C. N. R. C. (2024, February 10). *National Energy Code of Canada for Buildings: 2020 - NRC*

Publications Archive. https://nrc-publications.canada.ca/eng/view/object/?id=af36747e-3eee-4024-a1b4-73833555c7fa

Capozzoli, A., Cerquitelli, T., & Piscitelli, M. S. (2016). Enhancing energy efficiency in buildings through innovative data analytics technologiesa. In C. Dobre & F. Xhafa (Eds.), *Pervasive Computing* (pp. 353–389). Academic Press. https://doi.org/10.1016/B978-0-12-803663-1.00011-5

Congiu, E., Desogus, G., Quaquero, E., Argiolas, R., Di Pilla, L., & Frau, V. (2022). *BIM-based workflow for energy audits*. *3*, 0–0. https://doi.org/10.35490/EC3.2022.199

de Wilde, P., Martinez-Ortiz, C., Pearson, D., Beynon, I., Beck, M., & Barlow, N. (2013). Building simulation approaches for the training of automated data analysis tools in building energy management. *Advanced Engineering Informatics*, *27*(4), 457–465. https://doi.org/10.1016/j.aei.2013.05.001

Francoeur, D., Roux, L., & Parekh, A. (2006). *EnerGuide for Houses Retrofit Incentive Program: A Winning Strategy to Improve Energy Efficiency of Canadian Housing*.

Grover, D., Fallah, Y. P., Zhou, Q., & Ian LaHiff, P. E. (2020). Data-Driven Modeling and Optimization of Building Energy Consumption: A Case Study. *2020 IEEE Power & Energy Society General Meeting (PESGM)*, 1–5. https://doi.org/10.1109/PESGM41954.2020.9281663

Ham, Y., & Golparvar-Fard, M. (2012). *Rapid 3D Energy Performance Modeling of Existing Buildings Using Thermal and Digital Imagery*. 991–1000. https://doi.org/10.1061/9780784412329.100

Himeur, Y., Alsalemi, A., Bensaali, F., & Amira, A. (2020). Building power consumption datasets: Survey, taxonomy and future directions. *Energy and Buildings*, *227*, 110404. https://doi.org/10.1016/j.enbuild.2020.110404

Hong, T., Chen, Y., Luo, X., Luo, N., & Lee, S. H. (2020). Ten questions on urban building energy modeling. *Building and Environment*, *168*, 106508. https://doi.org/10.1016/j.buildenv.2019.106508

Hopf, K., Sodenkamp, M., & Staake, T. (2018). Enhancing energy efficiency in the residential sector with smart meter data analytics. *Electronic Markets*, *28*(4), 453–473. https://doi.org/10.1007/s12525-018-0290-9

Kong, D.-S., Kim, D.-H., Chang, Y.-S., & Huh, J.-H. (2014). Existing Building Energy Simulation Method Using Calibrated Model by Energy Audit Data. *Korean Journal of Air-Conditioning and Refrigeration Engineering*, *26*(5), 231–239. https://doi.org/10.6110/KJACR.2014.26.5.231

Mantha, B. R. K., Menassa, C. C., & Kamat, V. R. (2016). A taxonomy of data types and data collection methods for building energy monitoring and performance simulation. *Advances in Building Energy Research*, *10*(2), 263–293. https://doi.org/10.1080/17512549.2015.1103665

Maricar, N. M., & Jamal, Md. N. (2005). Industrial Energy Audit using Data Mining Model Web Application. *TENCON 2005 - 2005 IEEE Region 10 Conference*, 1–6. https://doi.org/10.1109/TENCON.2005.301195

Mutani, G., Todeschi, V., Kämpf, J., Coors, V., & Fitzky, M. (2018). Building energy consumption modeling at urban scale: Three case studies in Europe for residential buildings. *2018 IEEE International Telecommunications Energy Conference (INTELEC)*, 1–8. https://doi.org/10.1109/INTLEC.2018.8612382

Nichiforov, C., Stancu, I., Stamatescu, I., & Stamatescu, G. (2020). Information Extraction Approach for Energy Time Series Modelling. *2020 24th International Conference on System Theory, Control and Computing (ICSTCC)*, 886–891. https://doi.org/10.1109/ICSTCC50638.2020.9259635

Sayfiddinov, S., Akhmadiyorov, U. S., Razzokov, N. S. U., & Akhmedov, P. S. (2020). OPTIMIZATION OF MODELING WHILE INCREASING ENERGY EFFICIENCY OF BUILDING STRUCTURES OF PUBLIC BUILDINGS. *Theoretical & Applied Science*, *86*(06), 16–19. https://doi.org/10.15863/TAS.2020.06.86.3

Yang, C., Cheng, Q., Lai, P., Liu, J., & Guo, H. (2018). Data-Driven Modeling for Energy Consumption Estimation. In F. Aloui & I. Dincer (Eds.), *Exergy for A Better Environment and Improved Sustainability 2: Applications* (pp. 1057–1068). Springer International Publishing. https://doi.org/10.1007/978-3-319-62575-1_72

Yang, S., Wi, S., & Kim, S. (2018). Development Methodology of Web Crawling Based on Physical Properties DB of Building Materials for the Efficiency of Building Energy Simulation. 한국생활환경학회지, *25*(4), 467–475. https://doi.org/10.21086/ksles.2018.08.25.4.467

Zucker, G., Malinao, J., Habib, U., Leber, T., Preisler, A., & Judex, F. (2014). Improving energy efficiency of buildings using data mining technologies. *2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE)*, 2664–2669. https://doi.org/10.1109/ISIE.2014.6865041