GAINING KNOWLEDGE FROM BIG DATA: ENERGY PERFORMANCE CERTIFICATE AS A SOURCE OF INFORMATION TO DECARBONIZE THE BUILT ENVIRONMENT

ABSTRACT

The decarbonization strategies for the built environment that policy-makers face today from the EU mandate risk being made with incomplete or insufficient information. The consequence of this could be ineffective choices, thus slowing down the ongoing ecological transition, or their high cost, whether borne by the state or citizens.

The progressive and unstoppable digitization of the built environment offers information collection and previously unthinkable management opportunities. The construction sector, traditionally lagging behind other industrial sectors, is beginning to produce large quantities of data that can be exploited thanks to the most modern techniques derived from the information technology sector.

Among the most promising data sources are energy performance certificates for buildings, which provide a snapshot of the characteristics of buildings, their fabric and plant components, and design forecasts of their energy performances. Analyzing the energy performance certificates through Artificial Intelligence techniques proves the effectiveness of using big data in the construction sector. In particular, in this study, unsupervised machine learning techniques led to an in-depth knowledge of a stock of buildings approaching two hundred thousand units distributed over an almost twenty-four thousand square kilometers area in northern Italy.

Keywords: Artificial Intelligence, Residential buildings, Energy performances, Decarbonisation, Open data

INTRODUCTION

The 'European Climate Law', in June 2021, established the aim of reaching a 55% reduction of greenhouse gas emissions (GHG) in the European Union (EU) by the year 2030 and a net zero GHG level by the year 2050 [1]. According to the European Commission, 2020, the highest and most cost-effective carbon reductions (at least 60% compared to 2015) can be made by buildings, which now account for 36% of greenhouse gas emissions and 40% of final energy in the EU [3]. Nowadays, 75% of the buildings in the EU are energy inefficient, and many households continue to utilize antiquated heating systems that burn dirty fossil fuels like coal and oil. It would be necessary for the renovation rate, which is currently approximately 1%, to quadruple or more in the years leading up to 2030 to fully realize the potential improvements [4].

In this context, massive databases kept by public agencies, such as national or regional departments of construction and infrastructure, are used to comprehend the existing building stock characteristics. For example, in Italy, building data, including conditioned floor area, energy consumption, space conditioning energy systems, and envelope features, are reported by the "Certificazione Energetica degli Edifici" DataBase (CENED DB). This dataset was developed using certification reports for buildings' energy efficiency that accredited energy consulting firms handled.

However, data collected in large datasets contain errors, missing values, and discrepancies that must be addressed to distill valuable information. Therefore, the exponential growth in volume, variety, velocity, and value of large datasets has led to the emergence of Big Data processing trends and, particularly, analytics and data science techniques [5]. Moreover, Artificial Intelligence (AI), especially Machine Learning (ML) techniques, are increasingly deployed to realize value from big data in decision support. ML is essential to overcome the obstacles presented by big data and transform its potential into actual value for commercial decision-making and scientific research [6].

In this study, we addressed the following research questions:

- What knowledge is it possible to gain from big data?
- What could be the role of AI in scouting information?

To answer these questions, we used unsupervised learning techniques, i.e., clustering, to characterize and understand the building stock of the Lombardy region.

MATERIALS AND METHODS

This research analyses the CENED DB, the energy cadastral of buildings in the Italian Lombardy region. The open DB includes information on the primary and net energy performances of buildings, as well as geometric data (such as volume, gross and net surface, window area, etc.) and installed technologies (primarily data on the average thermal transmittance of building components and details on the overall efficiency of thermal plants). The dataset contains around 1.52 million records described by 45 features. Among these, the most relevant to this study are: i) the gross and net heated surface; (ii) the gross and net volume; (iii) the envelope surface; (iv) the ratio between opaque and transparent envelope surface; (v) the average walls, windows, and roof thermal transmittance; and (vi) the primary energy for heating EPH. The data in the database is gathered by numerous people, making it untrustworthy and necessitating a data cleaning process summarized in *Table 1*.

Table 1 Data cleaning criteria.

Selecting criteria	*Number of remaining records*
Deleting all records not pertaining to residential assets, “E.1 (1)” and “E.1 (2)” according to CENED classification	1,283,838
Deleting records with EP_H <=0 or EP_H >=300 kWh/m^{2}y	1,048,829
Deleting records with gross or net heated surface <=20 m^2	1,045,947
Deleting records with gross or net heated volume <=30 m^3	1,045,103
Deleting records with envelope surface <=20 m^2	967,971
Deleting records with the ratio between transparent and opaque envelope surface <=0	964,221
Deleting records with average walls transmittance U_{walls}<=0 W/m^{2}K or U_{walls}>=40 W/m^{2}K	963,251
Deleting records with average windows transmittance U_{win}<=0 W/m^{2}K or U_{win}>=40 W/m^{2}K	956,218
Deleting records with year of construction y<=1800 or y>=2021	956,143

The remaining records are then analyzed using clustering techniques. Clustering is a crucial tool for data mining and pattern recognition [7]. This strategy aims to categorize a group of objects into classes or clusters so that objects within the same cluster can be inferred to be of the same type, and objects within different clusters may be inferred to be different types [8]. There are three main uses for data clustering [9]:

- Understand the fundamental structure: understanding the data, formulating hypotheses, spotting abnormalities, and identifying crucial elements.
- Natural classification: assessing how closely different species or shapes resemble one another.
- Compression: employing cluster prototypes to organize and summarize data.

In this study, clustering divides the entire building stock in Lombardy into groups with similar energy characteristics.

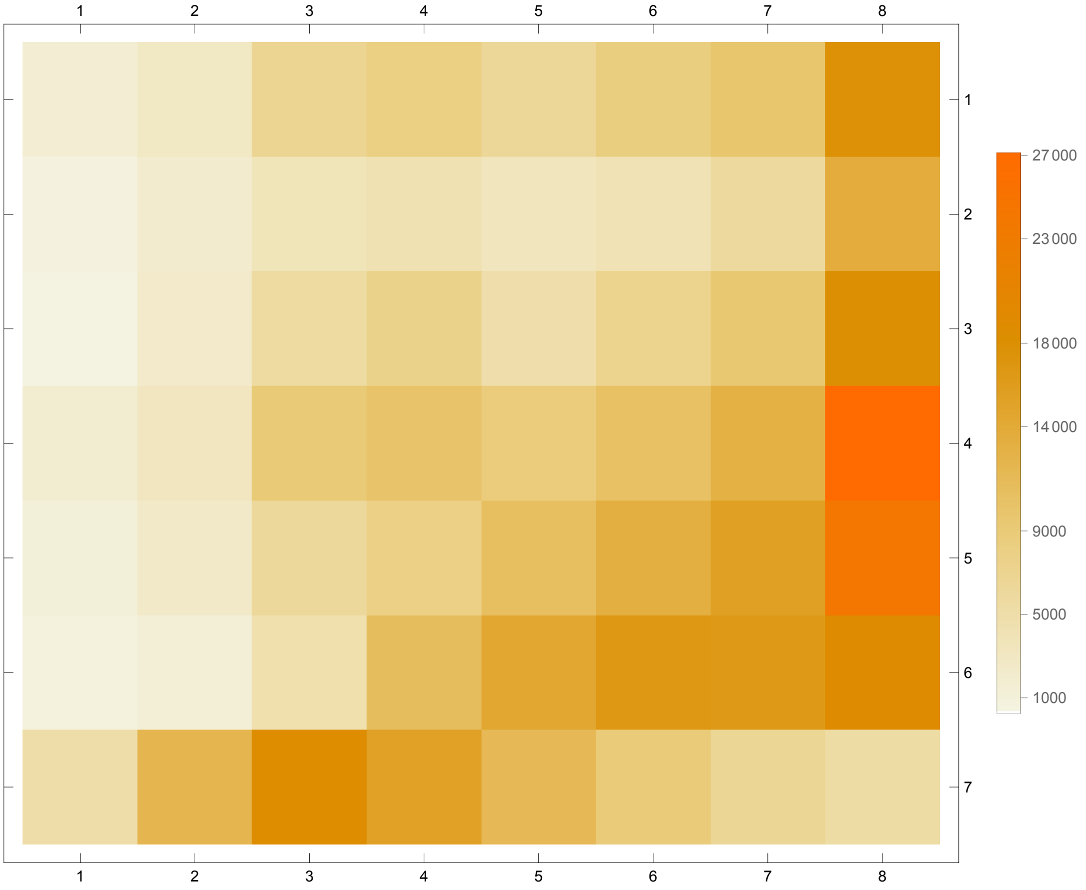
RESULTS

The results of this study are twofold and inspire the structure of the next subsections. The Exploratory Data Analysis (EDA) provides the first results, which answer the first research question, whereas the second question is answered in the last subsection of this paragraph and involves clustering techniques.

Exploratory Data Analysis

In statistics, EDA, introduced in [10], is a way of evaluating data sets to summarize their major features, frequently done using statistical graphics and other data visualization approaches. A statistical model may or may not be utilized, but the primary goal of EDA is to explore what the data may tell us beyond formal modeling, contrasting typical hypothesis testing.

The analysis of more than 900 thousand buildings needs data visualization techniques to extract useful information. Grouping the building stock by year of construction and energy performances gives important insight into the overall status of the Lombardy's built environment (Figure 1). It is noteworthy that the time frame for "Year of construction" (YoC) groups is based on introducing new and different energy legislation in Italy. Most of the remaining constructions come after the second world war, especially in the window frame from 1961 to 1976, where Italy was living an economic boom that involved a thriving business in all sectors, including construction. However, most assets built before 1992 were significantly inefficient from an energy performance point of view. Only after adopting the first national law to lower buildings' energy demand did the performance slightly increase.



	1	2	3	4	5	6	7	8
YoC	Before 1930	1930-45	1946-60	1961-76	1977-92	1993-2006	After 2006	
EL	A+	A	B	C	D	E	F	G

Figure 1. Year of construction (Y axes) and energy performances (X axes) of residential buildings in Regione Lombardia

Moreover, maintaining the same grouping philosophy, it is interesting to see the distribution of energy performance indicators, such as EPH and ETH, and sustainability indicators, like CO2 emissions (Figure 2). Undoubtedly, building performances in energy class G improved by the time, probably thanks to increasingly efficient building technologies and more strict code requirements. However, compared to the buildings classified inside the “G” category, the distribution of the other assets’ parameters across different YoC shows slight significant variation. One possible explanation is that buildings under better energy certification have undergone renovations and present better performances, despite their original structure being dated earlier.

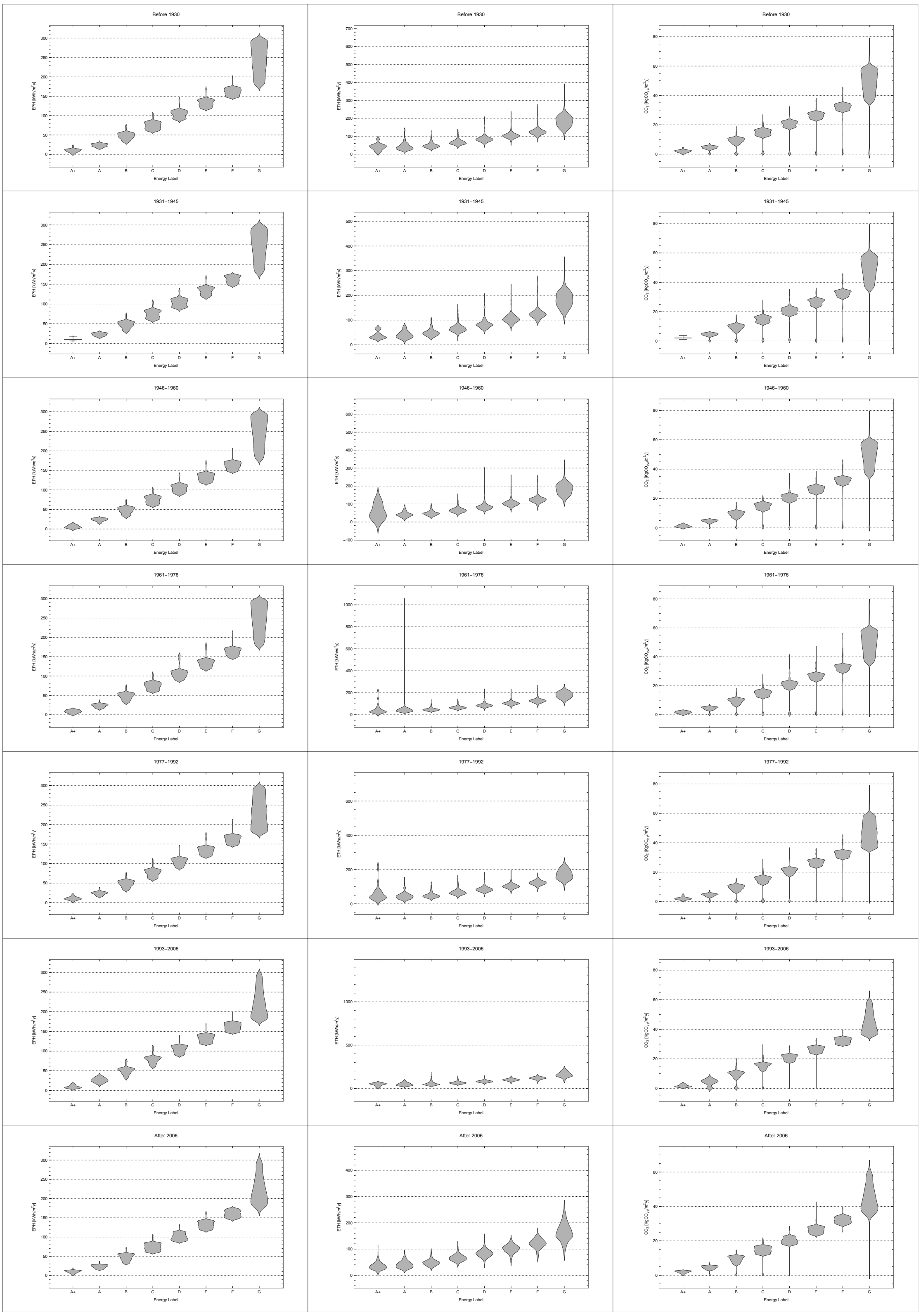


Figure 2 Statistical distribution of EPH, ETH and CO2 for each "Year of construction" group.

Figure 3 shows the geographical distribution of the buildings based on YoC and EPH values. The high number of points on the map hinders some information. However, it is visible that older and less efficient assets are primarily located in the historical center of the cities (the pattern is evident in Milan).

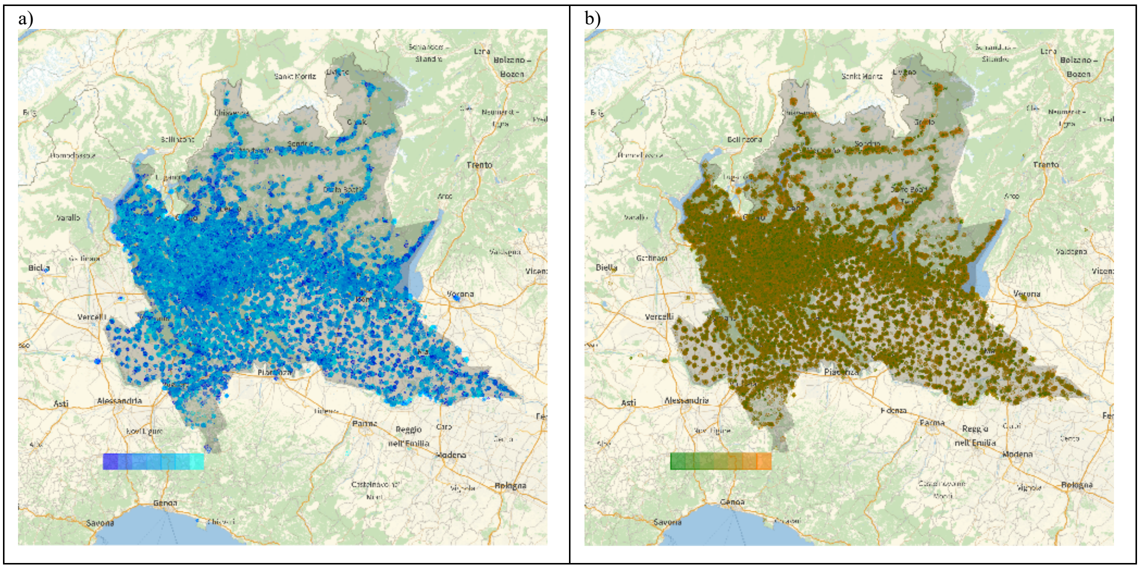


Figure 3 Geographical distribution of a) Year of Construction (the more intense is the blue color, the older the building), and b) EPH (red dots have higher EPH values than the green ones).

Clustering

Clustering is a standard approach for statistical data analysis used in many domains. It is the process of grouping a collection of objects so that items in the same group (called a cluster) are more similar than those in other groups.

In this study, we used clustering techniques to discover buildings group that might impact the sustainability of the Built Environment in the case of retrofitting. To create the clusters, we used a Gaussian mixture model that found 8 clusters from the datasets. The division of the datasets into clusters gave the following insights (Figure 4a):

- Cluster 4 is composed of the most recent buildings with the highest performance; therefore, they should not be considered in a retrofit scenario with a limited budget;
- Clusters 1, 5, 6, and 7 present the worst transmittance value and the highest EPH; therefore, assets included in these clusters should be prioritized in a retrofit scenario. Unsurprisingly, most of the buildings included in these clusters come from the oldest YoC groups, confirming the findings retrieved from the EDA;
- Cluster 8 includes buildings that have bad transmittances but decent EPH values. Probably, buildings in this cluster underwent a systems renovation that did not involve the envelope and fixtures;
- Buildings in Cluster 3 present good characteristics; therefore, they should be considered for retrofit only in case of large budget availability.

It is noteworthy that the cluster divisions present a strong correlation among transmittances value (Figure 4b).

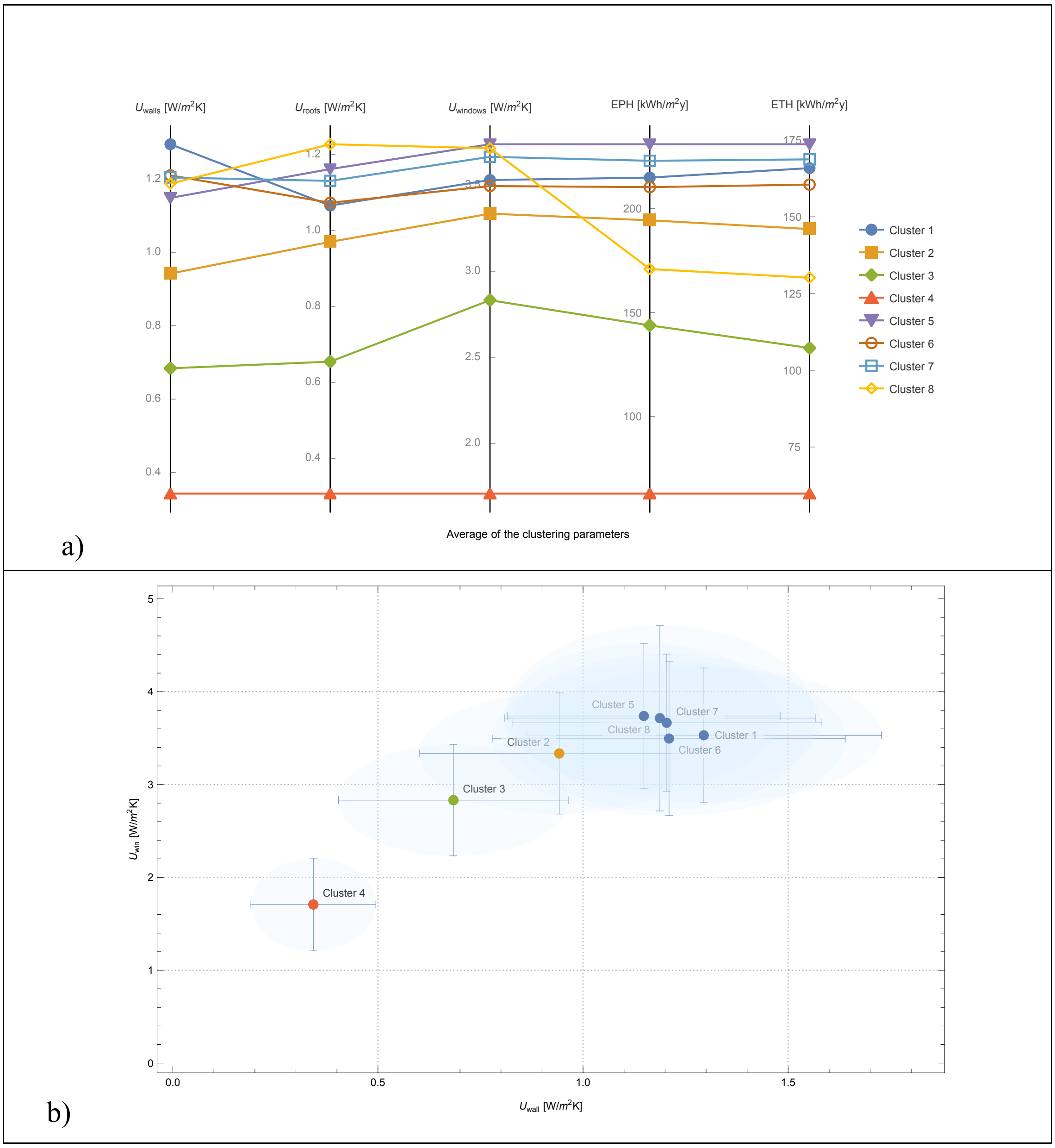


Figure 4 a) Distribution of the average values of the clusters, and b) distribution of the cluster based on wall and windows transmittance (U).

CONCLUSION

Decisions on energy retrofitting building portfolios must be supported by data and resources that are usually available in an open format and are collected by specific Institutions. Recently, the growing availability of data required specialized techniques to retrieve information from them. In this study, we used data mining and clustering techniques (a particular subfield of Machine Learning called unsupervised learning) to discover valuable insights from an open, large dataset in Italy. In particular, the methodology has been deployed effectively on an extensive portfolio: the residential assets developed in the Lombardy Region recorded in the CENED database, totaling over 900,000 records.

Using those techniques helped to find valuable insights that can ease decision-making processes that should select which assets need to be retrofitted first. On the one hand, the EDA depicted the distribution of the assets according to energy parameters and revealed which buildings might undergo retrofit interventions. On the other hand, clustering techniques grouped assets with similar characteristics. This step allows us to characterize further the building portfolio: for instance, buildings in cluster 8 have decent EPH values but bad transmittances, suggesting that those assets have efficient systems. Overall, the discovered clusters can be the backbone of a decision-making process policy that aims to retrofit the lower energy-efficient buildings. It is noteworthy that although scientific literature has examples of energy models for retrofitting districts or entire cities, they are complicated, difficult to calibrate, and challenging to update over time. Therefore, the data-driven approach adopted in this study depicts an overall picture of the energy performance building portfolio without the necessity of a complex energy model.

REFERENCES

[1] European Parliament and European Council, "European Climate Law 2021/1119." 2021.

[2] European Commission, "Stepping up Europe's 2030 climate ambition: Investing in a climate-neutral future for the benefit of our people." 2020.

[3] T. M. Gulotta, M. Cellura, F. Guarino, and S. Longo, "A bottom-up harmonized energy-environmental models for europe (BOHEEME): A case study on the thermal insulation of the EU-28 building stock," Energy Build., vol. 231, p. 110584, Jan. 2021, doi: 10.1016/j.enbuild.2020.110584.

[4] C. Fetting, "The European Green Deal," Vienna, 2020.

[5] J. Ranjan and C. Foropon, "Big Data Analytics in Building the Competitive Intelligence of Organizations," Int. J. Inf. Manage., vol. 56, p. 102231, Feb. 2021, doi: 10.1016/J.IJINFOMGT.2020.102231.

[6] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," Neurocomputing, vol. 237, pp. 350–361, May 2017, doi: 10.1016/J.NEUCOM.2017.01.026.

[7] F. Re Cecconi, A. Khodabakhshian, and L. Rampini, "Data-driven decision support system for building stocks energy retrofit policy," J. Build. Eng., vol. 54, p. 104633, Aug. 2022, doi: 10.1016/j.jobe.2022.104633.

[8] D. Pfitzner, R. Leibbrandt, and D. Powers, "Characterization and evaluation of similarity measures for pairs of clusterings," Knowl. Inf. Syst., vol. 19, no. 3, pp. 361–394, Jun. 2009, doi: 10.1007/s10115-008-0150-6.

[9] A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.

[10] J. W. Tukey, "Exploratory Data Analysis by John W. Tukey," Biometrics, vol. 33. 1977.