POLITECNICO DI TORINO
Repository ISTITUZIONALE

Towards a data-driven profiles of household energy consumption in collective housing: Machine learning clustering with insights from Bab Ezzouar (Algiers)

Original
Towards a data-driven profiles of household energy consumption in collective housing: Machine learning clustering with insights from Bab Ezzouar (Algiers) / Ouazar, K., Mutani, G., Mansour, L.A.. - In: ENERGY AND BUILDINGS. - ISSN 0378-7788. - 26:(2026). [10.1016/j.enbuild.2026.117957]

Availability:
This version is available at: 11583/3013191 since: 2026-07-16T17:20:48Z

Publisher:
Elsevier

Published
DOI:10.1016/j.enbuild.2026.117957

Terms of use:

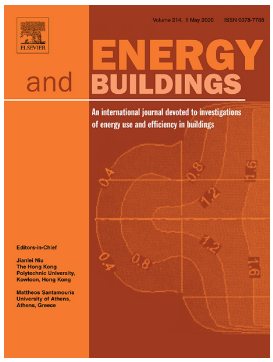
This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

12 November 2026

Journal Pre-proof



Towards a data-driven profiles of household energy consumption in collective housing: Machine learning clustering with insights from Bab Ezzouar (Algiers)

Kahina Ouazar, Guglielmina Mutani, Leila Abbas Mansour

PII: S0378-7788(26)01017-0

DOI: https://doi.org/10.1016/j.enbuild.2026.117957

Reference: ENB 117957

To appear in:

Received date: 31 March 2026

Revised date: 8 July 2026

Accepted date: 13 July 2026

Please cite this article as: K. Ouazar, G. Mutani and L.A. Mansour, Towards a data-driven profiles of household energy consumption in collective housing: Machine learning clustering with insights from Bab Ezzouar (Algiers), (2026), https://doi.org/10.1016/j.enbuild.2026.117957

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article. Please note that Elsevier's sharing policy for the Published Journal Article applies to this version, see: https://www.elsevier.com/about/policies-and-standards/sharing#4-published-journal-article. Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2026 Published by Elsevier B.V.

Journal Pre-proof

Towards a data-driven profiles of household energy consumption in collective housing: Machine learning clustering with insights from Bab Ezzouar (Algiers).

Kahina Ouazar [1]*, Guglielmina Mutani[2], Leila Abbas Mansour [3]

[1, 3] Department of Geography and Territorial Planning, University of Sciences and Technology Houari Boumediene (USTHB), Algiers 16000, Algeria (kouazar@usthb.dz, lmansour@usthb.dz).

[2] Department of Energy, Politecnico Di Torino, 10129 Torino, Italy (guglielmina.mutani@polito.it).

Correspondent author: Kahina Ouazar

Abstract

Residential energy consumption in North African collective housing remains insufficiently characterised, despite the need for energy-efficiency policies that account for both household practices and building-related constraints. This study addresses this gap in public collective housing in Bab Ezzouar, Algiers, Algeria, using a valid sample of 380 households. Five years of electricity and natural gas billing records were converted into average annual on-site energy consumption and combined with survey-based socio-demographic, behavioural and perceptual variables, dwelling geometry, equipment ownership and envelope thermal properties. The analysis combines univariate analysis, principal component analysis and the comparative evaluation of standard clustering algorithms.

K-means provided the most interpretable and internally consistent three-cluster solution, with a reasonable structure according to the Silhouette coefficient (0.513) and Davies–Bouldin index (0.71). The identified socio-technical profiles span approximately 14,200–19,100 kWh/year of annual on-site energy consumption. Cluster 03 corresponds to an affluent energy-intensive profile, Cluster 02 to a resource-constrained low-consumption profile, and Cluster 01 to a transition-oriented profile combining moderate consumption with favourable behavioural readiness. Cluster separation is mainly driven by dwelling size, equipment ownership and envelope thermal performance, particularly U-values, rather than by socio-demographic characteristics alone.

A central finding is the dissociation between declared willingness to save energy and realised consumption, showing that positive attitudes do not necessarily result in lower energy use when

Journal Pre-proof

physical, equipment-related and comfort-related constraints dominate. The findings support differentiated policy pathways: deep retrofits and demand-response measures for high-consuming households, energy-poverty support for constrained low-consumption households, and voluntary efficiency and renewable-energy programmes for transition-oriented households.

Keywords: Data-driven; Electricity and Gas consumption; Residential buildings; Machine Learning; Clustering; Households profiling; Algeria

Highlights

- Residential energy profiling in Algerian public collective housing.
- Five-year energy data combined with household survey and dwelling data.
- Socio-technical clustering of household energy-use patterns.
- Statistical analysis and comparative machine learning-based clustering evaluation.
- Policy pathways linked to behaviour, equipment ownership and building features.

1. Introduction

The building sector is central to global energy transition, accounting for over one-third of final energy consumption worldwide [1]. In residential contexts, improving efficiency remains challenging due to the combined influence of building attributes, energy systems, and occupant-related factors [2]. Capturing this socio-technical complexity is essential for reliable energy assessment, policy design, and targeted demand-side interventions [3,4]. Residential energy use emerges from the interplay of physical, technical, and behavioural determinants [5] . Physically, envelope characteristics: thermal transmittance, glazing, window-to-wall ratio, and shading; largely govern heat gains and losses, while geometry, orientation, and urban form modulate solar exposure and energy demand [6–9] . Technical systems, including HVAC, lighting, and on-site renewables, further shape performance [10–12] alongside climatic variability [13].

Occupant behaviour, comprising conscious and unconscious interactions with buildings and energy systems [14] ; remains a major source of variability and uncertainty [15]. Daily presence, thermostat settings, window operation, and appliance use generate substantial gaps between predicted and

Journal Pre-proof

observed consumption [16–18]. This behavioural heterogeneity is context-dependent and dynamic, challenging conventional energy modelling [19], and motivating analytical approaches capable of structuring occupant diversity [18].Clustering techniques have therefore emerged as a key tool for profiling residential energy use, reducing complexity while retaining meaningful heterogeneity[20].

Related work

Unsupervised machine learning has been widely employed for residential energy profiling, including K-means [21,22], K-medoids [23], hierarchical clustering [24,25] , Gaussian mixture models [26,27], DBSCAN [21], and self-organising maps [28]. Hierarchical clustering is valued for interpretability, whereas K-means is preferred for efficiency and scalability [29]. Comparative frameworks are common, evaluating multiple algorithms to select the most suitable approach. Recent studies increasingly adopt hybrid methods, combining clustering with statistical learning, simulation, or deep learning [30–33] integrating diverse data sources such as high-resolution smart meters and survey-based socio-demographics. While the former provides temporal granularity, the latter offer broader contextual insight.

Cluster validation is central to this literature and generally combines internal validity indices, downstream performance indicators, stability assessment and interpretability checks. Internal indices are commonly used to evaluate cohesion, separation and compactness [27,34], while downstream indicators may assess the usefulness of clusters for forecasting or demand-response applications [21]. Stability assessment provides an additional validation dimension by testing whether cluster partitions remain consistent under resampling or data perturbation, as emphasised by Hennig [35] and other validation studies [36]. Bootstrap-based approaches have also been proposed to support the selection of the number of clusters [37].

Feature engineering plays a central role in residential energy clustering, as the variables selected for analysis strongly influence the resulting household profiles. Beyond time-series indicators such as peak demand and load variability [38], recent studies increasingly incorporate contextual variables,

Journal Pre-proof

including building attributes, energy systems, appliance ownership and socio-demographic characteristics. This shift from signal-centred analysis to socio-technical profiling improves the interpretation of occupant–building interactions, but it also increases the complexity of cluster construction and validation. Previous research has shown that routines, household characteristics, appliance use and air-conditioning practices can substantially shape residential energy demand [22,39,40]. However, context-specific behavioural and socio-technical models remain limited, particularly in underrepresented regions [41].

The literature also shows several persistent limitations: a predominance of studies from developed countries, reliance on short-term or highly aggregated datasets, limited integration of physical and behavioural variables, and insufficient empirical evidence from North African collective housing [42–44]. Recent regional studies have begun to address these gaps from different perspectives. In Algeria, [45] used a bottom-up GIS-based approach to model heating and cooling needs across the national territory, while [46] applied clustering to multi-year electricity and natural gas consumption data from social housing in Algiers. In comparable Mediterranean and arid-climate contexts, [47] used clustering and open-data approaches to identify territorial patterns of residential energy consumption, and [48] applied K-means to high-resolution smart-meter data in Dubai to distinguish household load profiles, particularly in relation to air-conditioning use.

However, these studies mainly focus on territorial modelling, smart-meter load profiling, or temporal consumption patterns. Fewer studies combine long-term household billing data with socio-demographic characteristics, occupant behaviours and perceptions, dwelling geometry, equipment ownership, and envelope thermal properties within a single socio-technical clustering approach. This gap is particularly evident in Algerian public collective housing, where household-level empirical evidence remains limited.

In response, this study does not seek to introduce a new clustering algorithm. Rather, its originality lies in the empirical and contextual contribution provided by an original dataset from public collective housing in Bab Ezzouar, Algiers, Algeria. The study combines five years of electricity

Journal Pre-proof

and natural gas billing records, converted into annual on-site energy consumption, with structured household survey data from 380 valid respondents. The feature set integrates socio-demographic characteristics, occupant behaviours and perceptions, dwelling geometry, equipment ownership, and envelope thermal properties.

The contribution of this study is therefore threefold: first, it provides rare empirical evidence on residential energy profiles in an Algerian and North African collective-housing context; second, it combines long-term billing data with socio-technical and physical household-level variables; and third, it compares standard hierarchical and non-hierarchical clustering algorithms to assess the consistency and interpretability of the resulting profiles. This approach supports a more context-sensitive understanding of residential energy demand and provides evidence for targeted energy efficiency strategies in public collective housing.

2. Material and Methods

2.1 Case study

The study focuses on Bab Ezzouar, in eastern Algiers (Fig. 1). It is a municipality originally developed as a "cité-dortoir", primarily residential; and now characterised by dense collective housing estates. It represents one of the most illustrative examples of large-scale social housing implemented under national programmes since the 1970s, with early developments (1970s-1980s) coexisting alongside more recent projects (2000s-2015), reflecting sustained urban expansion. Today, the municipality comprises eight major urban units and hosts over 100,000 inhabitants across more than 17,000 collective housing units [49]. The study area lies within climatic zone A, as defined by DTR 2.3/4 ([50], corresponding to Algeria's coastal regions. It is characterised by a Mediterranean climate, with hot, dry summers and mild, relatively wet winters, which significantly influence residential energy demand[51].

At the Algerian national level, the residential sector represents a major share of final energy consumption[52], primarily supplied by electricity and natural gas [53]. Natural gas is primarily utilised for space heating, domestic hot water production, and cooking, whereas electricity is

predominantly used for air conditioning, lighting, and household appliances. In public collective housing, particularly in older developments; the absence of centralised or hybrid energy systems means that energy provision relies primarily on individual appliances at the dwelling level, resulting in fragmented and often suboptimal patterns of energy use.

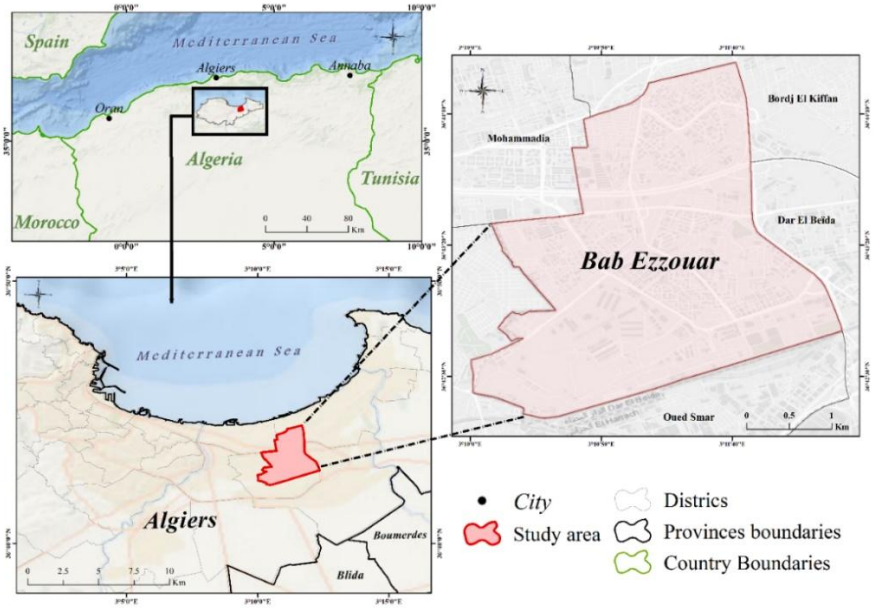


Fig. 1: Case study location.

From a morphological perspective, the housing stock is highly standardised, comprising primarily F3 (two-bedroom) and F4 (three-bedroom) apartments within mid-rise (G+4) and high-rise (G+9 to G+14) buildings, with Table 1 summarising the key characteristics of these typical dwelling types. These buildings were constructed using similar materials and construction techniques, typically featuring reinforced concrete structural frames with double-leaf external walls made of hollow bricks and floor slabs composed of solid or hollow-core concrete elements. This architectural homogeneity is accompanied by relatively uniform socio-economic conditions, characterised mainly by middle-income households and comparatively large family sizes. Nevertheless, variations in occupants' daily practices and behavioural patterns introduce additional diversity in residential energy consumption.

Table 1: Characteristics of typical dwellings in the study area.

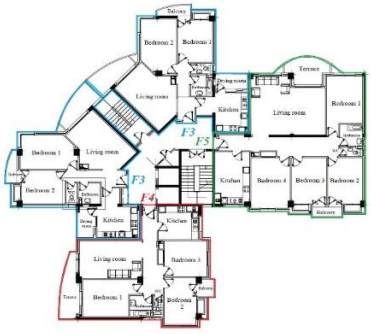
Floor plan	Dimensions

Type 01: 5-story multi-family residential buildings with 2 dwelling units per floor, built using brick masonry construction (1970-1995)

	F3	F4
Total useful area	69	79
Wall area	46	55
Ceiling height	2.6	2.6
Total windows area	13	16
S/V	0.4	0.45

Type 02: A 14-storey multi-family residential building with four dwelling units per floor, built using brick masonry construction (2001-2003)

	F3	F4
Total useful area	72	87
Wall area	50	70
Ceiling height	2.7	2.7
Total windows area	8	10
S/V	0.4	0.5



Type 03: A 9-story multi-family residential building with four dwelling units per floor, built using brick masonry construction (2015)

	F3	F4	F5
Total useful area	75	88	93
Wall area	65	78	84
Ceiling height	3	3	3
Total windows area	13.2	12.8	19
S/V	0.4	0.5	0.44

Given these characteristics, Bab Ezzouar provides a representative context for analysing residential energy use in Algerian collective housing. The analysis focuses on five eastern districts selected for their high density and availability of architectural data, reflecting the dominant housing typologies within the municipality (Fig. 2).

2.2 Methodology

To identify household energy consumption profiles in collective housing, a structured workflow was followed: data collection, preprocessing, PCA-based dimensionality reduction, clustering with multiple algorithms, internal validation via indices, and cluster interpretation (Fig. 3)

Journal Pre-proof

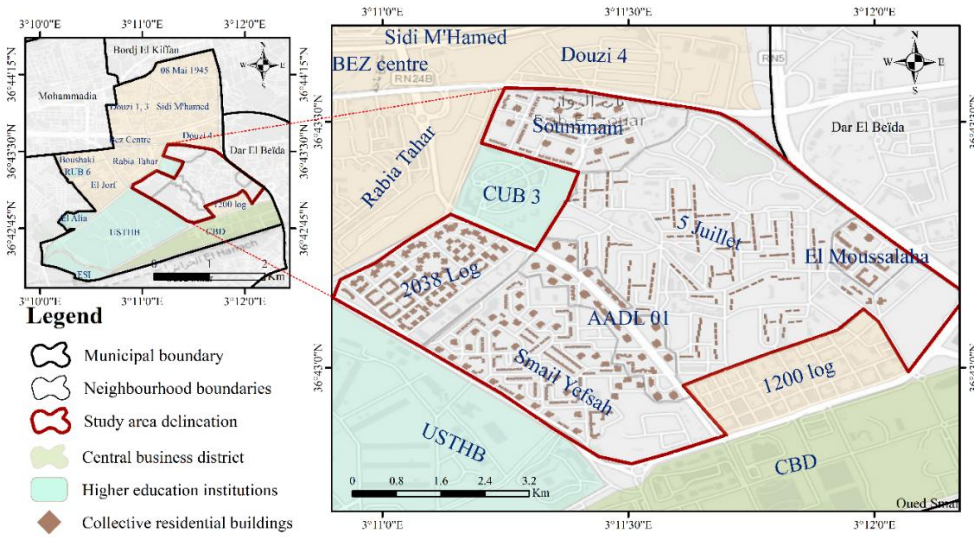


Fig. 2: Study area delineation.

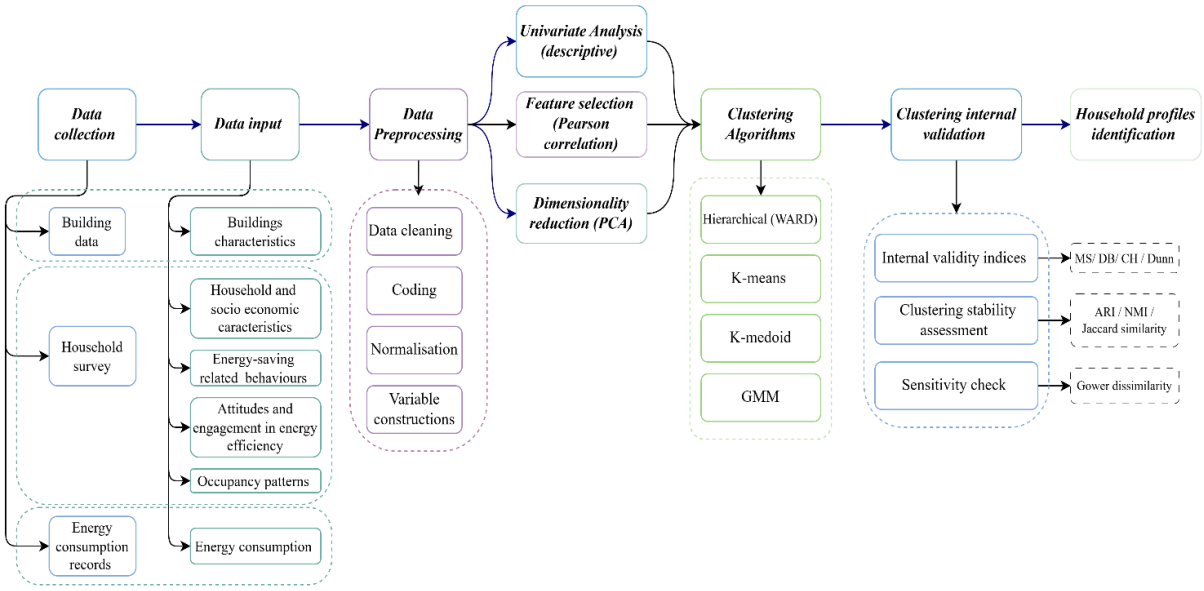


Fig. 3*:* Methodological framework adopted in the study

2.2.1 Data collection

- **Questionnaire survey**

A survey was conducted between November 2024 and June 2025 across the study area, focusing on public collective housing. A spatially targeted sampling approach was adopted, concentrating on districts with a high density of collective residential buildings and available building documentation. Household responses were collected through door-to-door surveys based on voluntary participation, potentially introducing self-selection bias. The survey was administered orally through face-to-face

interviews using a structured questionnaire comprising six thematic sections (illustrated in Fig. 4 and detailed in Supplementary Material, Appendix A), delivered by trained personnel. It covered basic household characteristics, energy systems and appliances in use, adoption of energy-saving measures, attitudes and perceptions regarding energy efficiency, as well as occupancy patterns and housing (inspired from [4,20,54,55] and adapted to the local context. The survey yielded 570 valid responses. All procedures were carried out by trained personnel and adhered to established ethical research standards, ensuring informed consent, confidentiality, and participant anonymity. The study complied with the Declaration of Helinski [56] and did not require formal ethics approval under Algerian regulations governing anonymous, non-interventional research.

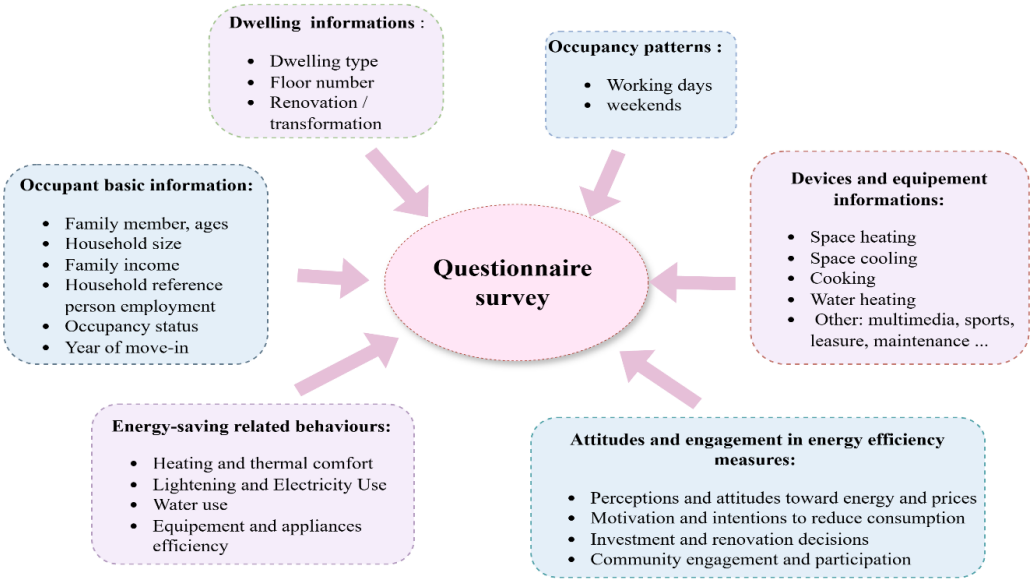


Fig. 4: Questionnaire design used in the study

- **Dwelling information**: This category includes geometric, morphological and envelope-related characteristics of the dwellings. Data were obtained from public institutions and design offices responsible for the construction and management of the residential stock, and were extracted from architectural drawings, AutoCAD files and civil engineering documents. Dwelling-level indicators were derived from technical drawings using manual measurements and AutoCAD-based calculations (Table1). These variables include direct geometric measurements (floor area, perimeter, height, heated volume, window area and floor level) and derived indicators describing dwelling

compactness, façade exposure and potential heat-transfer pathways. Their definitions, units and nomenclature, including SVR, WWR, THLA and THLP, are provided in Table 2.

Table 2: Nomenclature and description of geometric and envelope-related variables used in the clustering framework

Variable	Unit	Description
Surface-to-Volume Ratio (SVR)	m^2/m^3	Ratio between the dwelling envelope surface area and the heated volume. This indicator reflects the compactness of the dwelling and its potential thermal performance.
Window-to-Wall Ratio (WWR)	%	Ratio of total window area to total external wall area. It is commonly used to assess daylight availability, solar gains, and heat losses through glazing.
Total Heat Loss Area (THLA)	m^2	Total area of the dwelling envelope contributing to heat transfer, including external walls, windows, doors, and other exposed envelope components.
Total Heat Loss Perimeter (THLP)	m	Total length of envelope boundaries exposed to heat transfer, representing the extent of thermal exchange between the dwelling and its surroundings.

- **Thermal transmittance calculation:** The thermal properties of the building envelope components considered in this study, including walls, roofs, floors, and windows, were determined in accordance with the Algerian thermal building regulation DTR C3.2/4 [50]. Thermal transmittance coefficients (U-values) were calculated for both summer and winter conditions. Detailed calculation procedures, equations, and parameter definitions are provided in Supplementary Material, Appendix B.

- **Energy consumption records:** Longitudinal electricity and natural gas consumption data were obtained from SONELGAZ, the state-owned energy supplier in Algeria [57]. The dataset covers the period 2020–2024 and includes quarterly consumption records for each household. Electricity was recorded in kilowatt-hours (kWh), while natural gas was recorded in Thermies (Th) and converted into kWh using a fixed factor of (1 Th ≈ 1.162 kWh). Annual energy consumption was then calculated by summing electricity and converted gas consumption for each year, resulting in five annual final consumption values per household.

2.2.2 Preprocessing data

Data cleaning addressed missing values, incomplete responses and inconsistencies, resulting in a validated sample of 380 households from the initial 570 surveyed households. For each household,

the five annual final energy consumption values were averaged to obtain a single five-year mean consumption indicator. This variable was used to represent medium-term household energy use and was log-transformed using base 10 to reduce skewness and limit the influence of extreme values.

Survey and dwelling variables were coded prior to analysis. Categorical and ordinal variables were recoded according to predefined rules: dwelling types such as F3, F4 and F5 were coded according to the number of main rooms, *HRPjob* refers to the occupation of the household reference person, and binary variables such as air-conditioning ownership, renovation status and window renewal were coded as 0/1 variables. Continuous variables and composite indicators were standardised using the z-score transformation (Eq.1), where X is the original value, μ the sample mean and σ the standard deviation.

$$Z = (X - \mu)/\sigma \qquad (1)$$

Additional engineered variables were constructed from survey items to capture behavioural, perceptual and equipment-related dimensions. The construction rules are provided in Table 3. Equipment ownership was described through seven category-specific variables and one aggregate Household Equipment Index (HEI), calculated as the total number of appliances owned across all categories. These preprocessing steps ensured consistency and comparability before dimensionality reduction and clustering.

Journal Pre-proof

Table 3: Coding rules and construction of engineered features and composite indicators derived from survey data

Construct / Variable	Description	Constituent Items	Construction Method
Occupancy Ratio	Average daily dwelling occupancy level	Mean occupancy rate calculated across four daily time periods	Derived indicator
Weekend/Weekday Occupancy Ratio	Relative occupancy intensity during weekends compared with weekdays	Weekend occupancy rate / Weekday occupancy rate	Derived indicator
Household Equipment Index (HEI)	Overall equipment ownership level within the dwelling	Aggregated equipment categories listed below	Composite score

Large Appliances	Major domestic appliances with significant electricity demand	Refrigerator, electric oven, washing machine, dishwasher, cooker	Summative score
Domestic Maintenance Equipment	Equipment used for household maintenance activities	Iron, vacuum cleaner, hair dryer	Summative score
Multimedia and Communication Equipment	Information, communication and entertainment devices	Television, landline phone, mobile phone, computer, printer	Summative score
Leisure and Comfort Equipment	Appliances associated with comfort and recreational activities	Air conditioner, sport and leisure equipment	Summative score
Energy Sobriety Index (ESI)	Adoption of everyday energy-saving practices	12 energy-saving actions (heating, lighting, appliance use, hot water consumption, equipment maintenance, etc.)	Composite scored index
Willingness and engagement to energy efficiency index	Individual and collective commitment to energy transition and efficiency actions	13 items (Financial commitment, renovation willingness, participation in awareness programmes, adoption of best practices, investment intentions)	Composite scored index

2.2.3 Filtering, verification and dimensionality reduction of variables (PCA)

The initial dataset consisted of a range of explanatory variables (see Table C.1 in Supplementary material, Appendix C), which were processed, coded, and structured into seven distinct categories. To mitigate redundancy caused by multicollinearity, a Pearson correlation analysis [58] was first carried out, adopting a 0.95 threshold [59] Subsequently, a Principal Component Analysis (PCA) was employed to further reduce the dimensionality of the data[60]. Prior to running the PCA, two standard diagnostic tests were performed to verify the adequacy of the dataset: The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy[61]**,** calculated using Eq.2, and Bartlett's test of sphericity [62]**.** Both confirmed the relevance of inter-variable correlations and supported the use of PCA. Variables presenting an individual KMO value below 0.7 were excluded to ensure the robustness of the factor structure and improve the quality of the retained components [61].

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} p_{ij}^2} \qquad (2)$$

where r_{ij} denotes the correlation coefficient between variables i and j, p_{ij} represents the corresponding partial correlation coefficient, and $\sum_{i \neq j}$ denotes the summation over all variable pairs.

2.2.4 Profiles clustering

Journal Pre-proof

The component scores obtained from PCA were used as input variables for the clustering analysis. Four unsupervised clustering methods were compared to avoid relying on a single algorithm and to examine whether the household profiles remained coherent across different clustering approaches.

First, **Ward's hierarchical clustering** method, implemented in R using the Ward.D2 criterion, was applied as an exploratory technique. This method progressively merges the most similar observations based on squared Euclidean distances [63], while minimising within-cluster variance. It produces a dendrogram that helps visualise the structure of the data and identify plausible cluster configurations without requiring the number of clusters to be specified in advance [64].

K-means clustering was then used as a standard partitioning method and benchmark approach. It assigns households to a predefined number of clusters by minimising the distance between each observation and the centroid of its cluster [65]. This method was included because it is widely used in residential energy profiling and produces compact and easily interpretable groups.

K-medoids clustering, based on the Partitioning Around Medoids (PAM) algorithm [66] was also tested as a more robust alternative to K-means. Unlike K-means, which represents each cluster by a calculated centroid, K-medoids uses an actual observation as the representative centre of each cluster [67] .This makes it less sensitive to extreme observations. The average silhouette method was used to assess the balance between intra-cluster cohesion and inter-cluster separation [68].

Finally, **Gaussian Mixture Models (GMM)** were applied to test a probabilistic clustering approach. In contrast to deterministic methods, GMM estimates the probability that each household belongs to each cluster, allowing for the possibility of overlapping group structures [69]. Model selection was guided by the Bayesian Information Criterion (BIC), which balances model fit and complexity.

For each method, several values of k were tested. The number of clusters was selected by combining graphical and statistical criteria, including the elbow method, the average silhouette score, and

internal validation indices. The retained clustering solution was chosen based on both statistical performance and the interpretability of the resulting household energy profiles.

2.2.5 Cluster validation and model comparison

2.2.5.1 Internal validation

The PCA-based clustering algorithms were evaluated using four complementary internal validation indices (Table 4), capturing cohesion, separation, and compactness. Combined, these metrics provide an objective and consistent framework to assess algorithm performance and identify the most robust cluster structure.

Table 4: Internal validation indices used to evaluate clustering performance

Index	Function	Formula	Reference
Average silhouette coefficient	Assesses intra-cluster cohesion and inter-cluster separation.	$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$ (3) a_i is the average intra-cluster distance for observation i, and b_i is its minimum average distance to the nearest cluster.	[68]
Calinski-Harabasz index (CH)	Evaluates the ratio of between-cluster to within-cluster dispersion.	$CH(K) = \frac{B(K)/(K-1)}{W(k)/(n-K)}$ (4) $B(K)$ represents the between-cluster dispersion, $W(K)$ the within-cluster dispersion, K the number of clusters, and n the total number of observations.	[72]
The Davies–Bouldin index (DB)	Measures the average similarity between clusters based on compactness and separation.	$DB = \frac{1}{k}\sum_{i=1}^{k} max_{(j \neq i)} \left(\frac{S_i + S_j}{d_{ij}}\right)$ (5) k is the number of clusters, S_i the average intra-cluster distance for cluster i, and d_{ij} the inter-centroid distance between clusters i and j	[73]
Dunn index	Favors cluster configurations with high inter-cluster separation and low intra-cluster dispersion	$DI = \frac{\min_{i \neq j} d(c_i, c_j)}{\max_k \Delta_k}$ (6) $d(c_i, c_j)$ is the distance between the centroids of clusters i and j, and Δ_k denotes the maximum distance between observations within cluster k	[74]

2.2.5.2 Resampling-based cluster stability assessment

After comparing the clustering algorithms using internal validity indices, the method providing the best balance between statistical performance and interpretability was retained for further assessment. A resampling-based stability analysis was then conducted to evaluate whether the selected clustering solution remained consistent when the sample composition was moderately perturbed.

Since the clustering are unsupervised methods, no ground-truth labels were available to compute conventional classification error metrics. Stability was therefore assessed by comparing the reference partition with partitions obtained from repeated subsampling. A total of 100 random subsamples were generated from the retained clustering solution, each containing 80% of the observations without replacement. For each subsample, the clustering procedure was re-estimated using the same number of clusters, and the resulting partition was compared with the corresponding subset of the reference partition. Table 5 shows the agreement metrics used in this step

Table 5: Metrics used for resampling-based cluster stability assessment

Metric	Function	Reference
Adjusted Rand Index (ARI),	Measures the similarity between two partitions while adjusting for chance agreement	[75]
Normalised Mutual Information (NMI)	Evaluates the amount of shared information between the reference and resampled partitions	[76]
Pairwise Jaccard similarity	Assesses the proportion of observation pairs consistently assigned to the same cluster	[35]

Nb: values close to 1 indicate high stability of the clustering solution.

2.2.5.3 Mixed-data sensitivity check using Gower dissimilarity

Because the retained feature set included variables of different types, a mixed-data sensitivity check was added using Gower dissimilarity[70]. This method is suitable for datasets combining quantitative, ordinal, binary and categorical variables, as it computes pairwise dissimilarities while accounting for the measurement scale of each variable[71]. Quantitative variables, including log-transformed annual energy consumption, behavioural and engagement scores, geometric indicators, thermal properties and equipment count, were treated as continuous. Binary and categorical variables, such as air-conditioning ownership, renovation works, window and door renewal, dwelling type and household socio-economic categories, were treated as categorical.

Agglomerative hierarchical clustering with average linkage was then applied directly to the Gower dissimilarity matrix, without PCA transformation. This analysis was used as a sensitivity check to verify whether the preferred cluster structure remained consistent when the mixed nature of the dataset was explicitly considered.

Journal Pre-proof

2.2.6 Household profile identification and socio-technical interpretation

After selecting the retained clustering solution, household profiles were identified and interpreted using the original variables rather than the reduced factorial coordinates. This step aimed to translate the statistical partitions into meaningful socio-technical profiles. For quantitative variables, both mean values and standard deviations were reported to describe central tendencies and within-cluster dispersion. For categorical variables, frequency distributions were used where applicable. These descriptive statistics were then used to construct an operational typology of household energy profiles. All analyses were performed using R and Python. R was used for preprocessing, statistical testing, PCA, clustering and internal validation, while Python was used for the Gower-based sensitivity analysis, cluster stability assessment and additional visualisation outputs.

3. Results

The initial dataset included 70 variables combining survey responses, energy consumption records and dwelling characteristics, grouped into seven categories (detailed in Supplementary material, appendix C). After preprocessing and feature engineering, a refined dataset of 47 coded and non-redundant variables was obtained. This dataset was used for the univariate analysis and PCA-based dimensionality reduction.

3.1 Univariate analysis, feature selection and PCA

Fig.5 shows the distribution of annual on-site energy consumption after log10 transformation, which reduced skewness and limited the influence of high-consumption households.

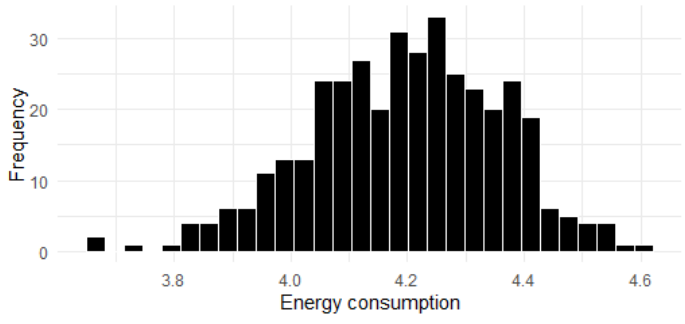


Fig. 5: Log Five-year (2020-2024) average annual energy consumption (Gas & Electricity) (kWh)

Before PCA, a Pearson correlation matrix was used to identify redundant variables and reduce multicollinearity (Fig. 6). Only *Windows_area* and WWR exceeded the predefined threshold, showing a near-perfect correlation (r = 0.995). Since WWR partly derives from window area and overlapped with other envelope-exposure indicators, it was removed, while *Windows_area* was retained as a direct geometric measure of glazed surface.

Data adequacy for PCA was assessed using the Kaiser–Meyer–Olkin (KMO) measure and Bartlett's test of sphericity. The initial global KMO value was 0.65, indicating moderate suitability for PCA. To improve the quality of the factor structure, variables with individual KMO values below 0.70 were removed, resulting in the exclusion of 21 variables. The retained set of 26 variables increased the global KMO to 0.79, indicating good sampling adequacy (Table 6).

Table 6: Individual KMO scores of the dataset's variables

Overall MSA = 0.79					
Energy consumption	Willingness and Engagement index	ESI	Year_ housing	Building construction period	Dwelling type
0.71	0.79	0.7	0.83	0.77	0.82
Number floor	Dwelling perimeter	THLP	THLA	WWR	U_{wn}
0.73	0.89	0.78	0.83	0.84	0.72
U_{door}	Age_18-29	Age_30-49	Income level	HRPjob	HEI
0.75	0.74	0.86	0.74	0.73	0.77
Space Cooling (air-conditioning)	Retrofit interventions	Household size	Renovation	Windows substitution	Door replacement
0.72	0.74	0.7	0.74	0.77	0.8
Balcony extension	Water heating system				
0.83	0.73				

MSA: mean silhouette average

Journal Pre-proof

Journal Pre-proof

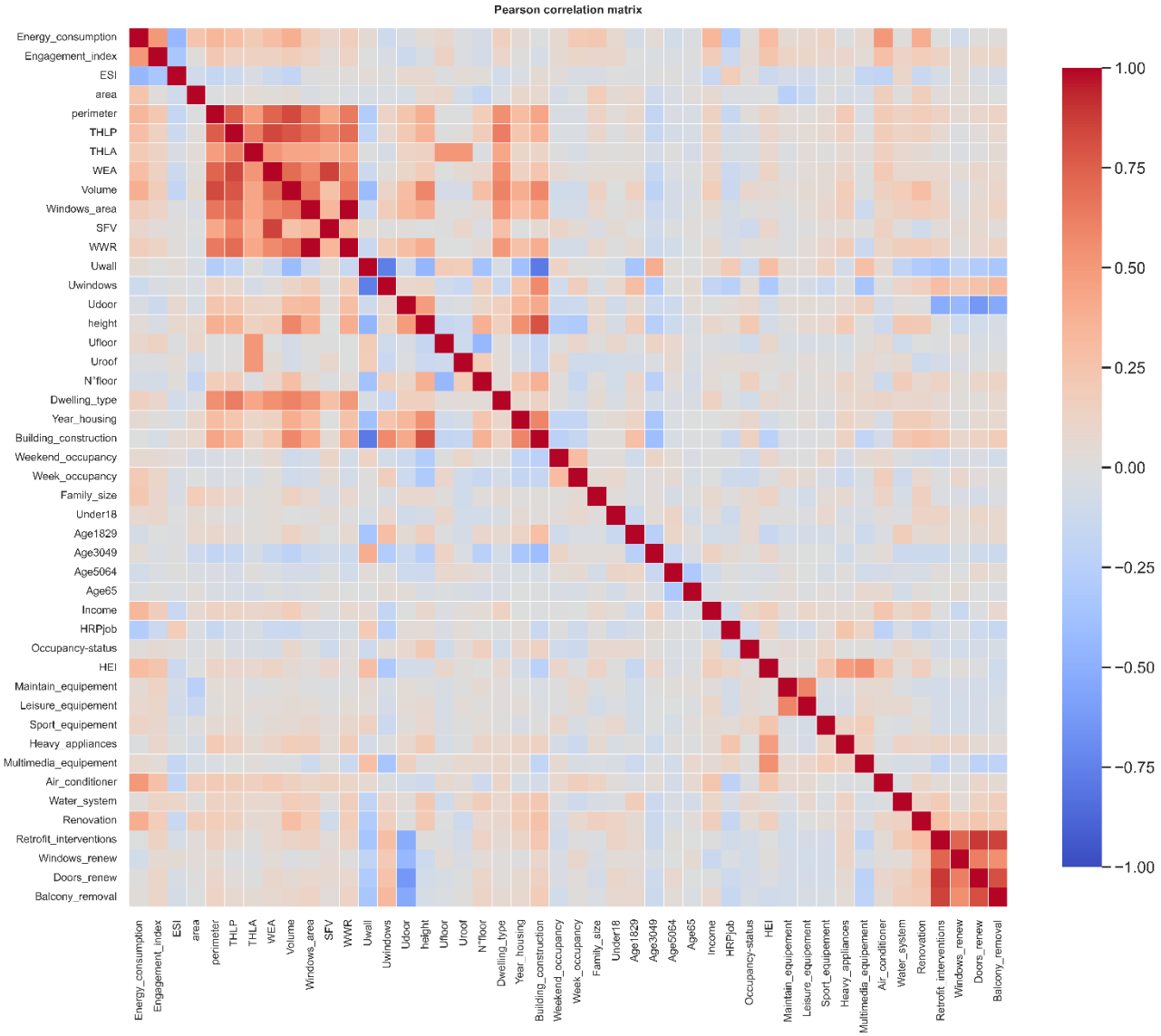


Fig. 6: Pearson Correlation matrix

Bartlett's test further confirmed the suitability of the dataset for PCA, with significant inter-variable correlations ($\chi^2 = 4630.214$, df = 325, $p < 0.001$). Based on the retained variables, PCA identified ten components explaining 74.5% of the total variance (Fig. 7). These components were then used as reduced-dimensional inputs for the subsequent clustering analysis.

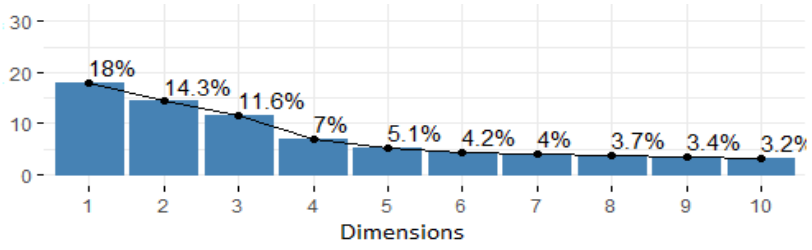


Fig. 7: Distribution of Explained Variance across the First Ten Principal Components

This initial stage, comprising variable transformation, univariate statistical analysis, and principal component analysis (PCA); is illustrated in Fig. 8, which provides a schematic overview of the evolution of variable categories throughout the study, from raw input data to pre-processed variables and, ultimately, to the final set of features retained after dimensionality reduction.

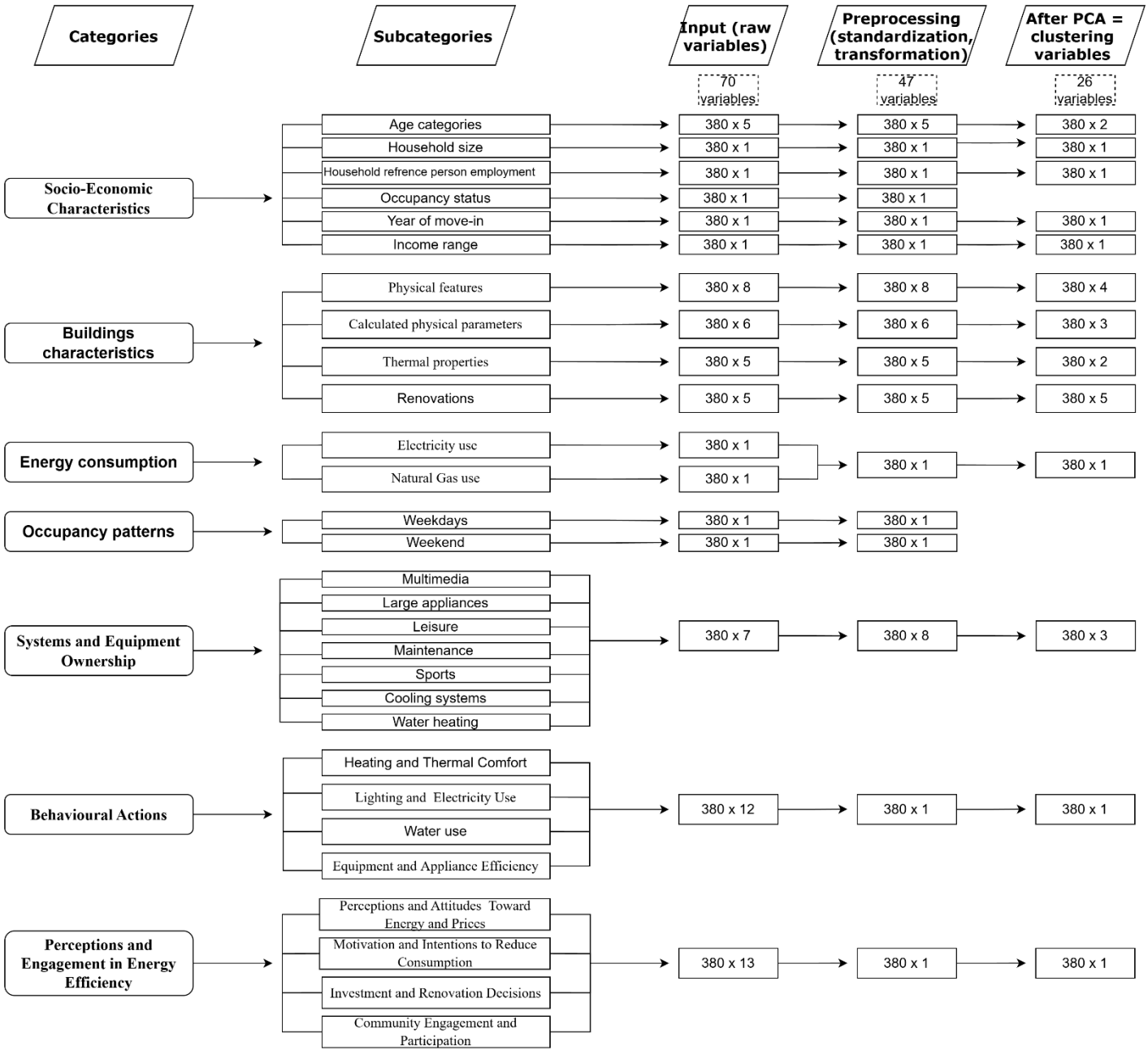


Fig. 8: Workflow of Variable Transformation and Dimensionality Reduction Across Data Processing Stages

3.2 Clustering

The dataset obtained at this initial stage subsequently forms the basis for the clustering analyses

3.2.1 Hierarchical clustering

Ward's hierarchical clustering was applied to explore the internal structure of the PCA-reduced dataset. The average silhouette analysis indicated two plausible solutions, with local maxima at k = 3 and k = 4 (Fig. 9). Since the difference between these two configurations was relatively limited, an additional sensitivity comparison was conducted using internal validation indices.

As shown in Table 7, the k = 3 solution performed better than the k = 4 solution across all validation criteria. It achieved a higher average silhouette score (0.448 versus 0.415), a lower Davies–Bouldin index (0.728 versus 0.774), a higher Calinski–Harabasz index (342.11 versus 314.29), and a higher Dunn index (0.025 versus 0.022). These results indicate that the three-cluster solution provides a better balance between intra-cluster cohesion and inter-cluster separation. Therefore, k = 3 was retained as the most parsimonious and interpretable Ward partition.

Table 7: Sensitivity analysis of Ward hierarchical clustering solutions for k = 3 and k = 4

	K=3	K=4
Average silhouette	0.448	0.415
DB	0.728	0.774
CH	342.11	314.29
Dunn	0.025	0.022

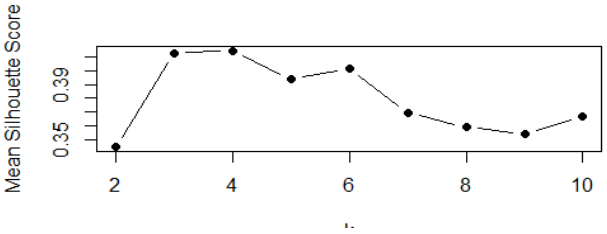


Fig. 9: Value of the optimal number of clusters f for hierarchical clustering

The retained Ward solution with k = 3 is shown in Fig. 10(a). Each household was assigned to a single cluster, providing an initial exploratory segmentation of the sample. Fig. 10(b) illustrates the distribution of observations within the Ward clustering structure and supports the interpretation of three broad household energy profiles.

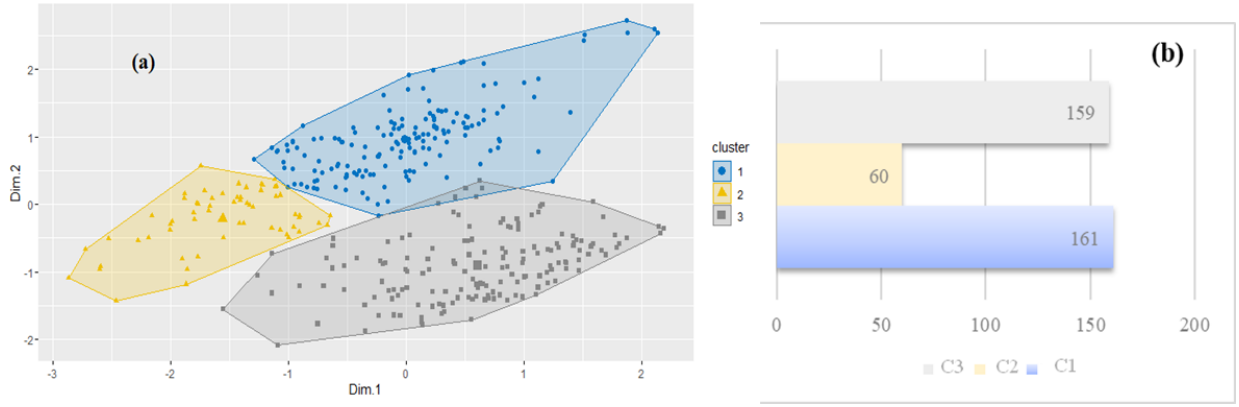


Fig. 10: Hierarchical clustering plot results; **(a)** clusters visualisation , **(b)** observations distribution in the clusters

3.2.2 K-means

The optimal number of clusters for K-means was determined using the Elbow (Fig. 11.a) and average silhouette methods Fig. 11.b), both indicating k = 3 with a mean silhouette near 0.5, supporting a clear and stable partition. Final clustering strictly assigned each observation to one cluster, with Fig. 12 illustrating both the overall distribution and the cluster-wise observation spread, confirming interpretability and robustness.

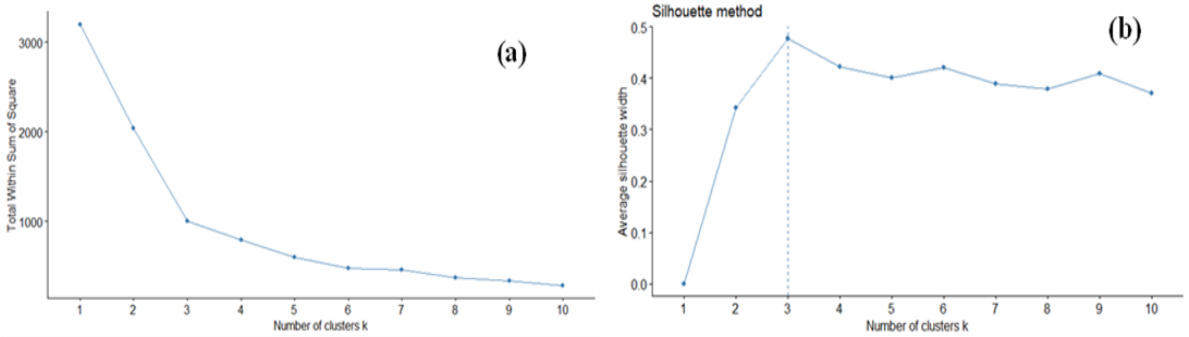


Fig. 11: Optimal number of clusters with **(a)** the Elbow method; **b)** average silhouette method

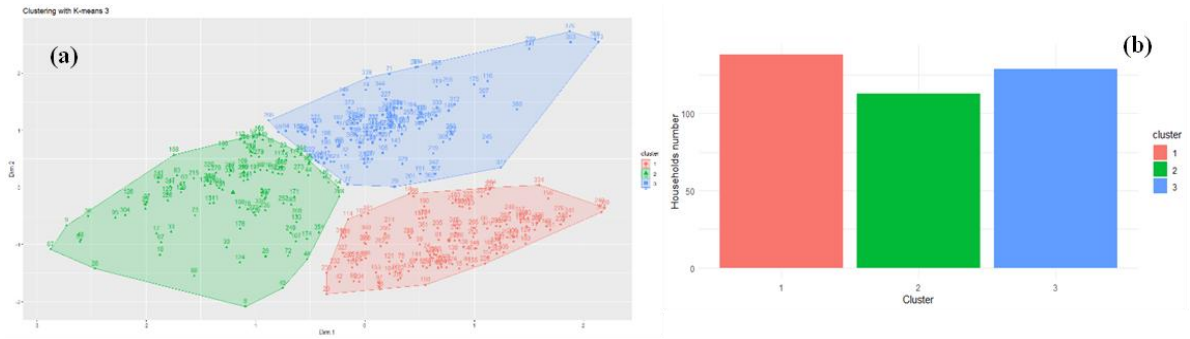


Fig. 12: K-means clustering results **(a)** clusters partition, **(b)** observations distribution in the clusters.

3.2.3 K-medoid

A **(**PAM algorithm) was applied as a robust alternative to K-means, defining clusters around actual observations (medoids) rather than centroids, which reduces sensitivity to outliers. Using the average silhouette criterion, the silhouette curve Fig. 13(a) clearly peaks at $k = 3$, indicating a compact and well-defined partition. The resulting three stable clusters, shown in Fig. 13(b), confirm the method's effectiveness in capturing the dataset's underlying structure without outliers.

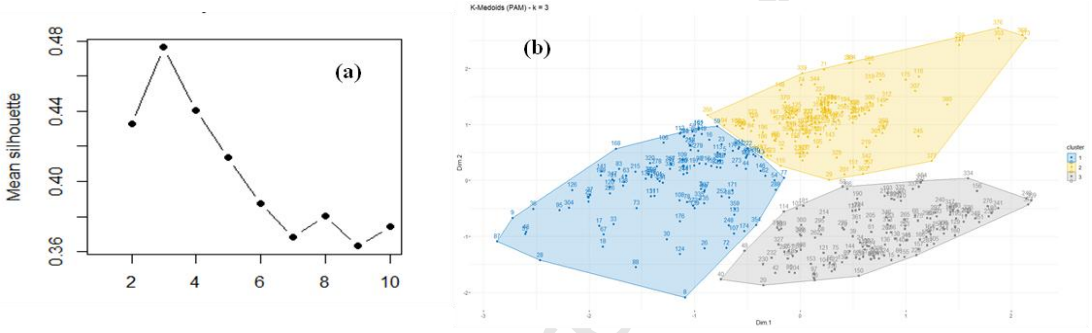


Fig. 13: (a) search for optimal k number with K-medoid clustering, **(b)** K-medoid clustering results.

3.2.4 The Gaussian Mixture Model (GMM)

The GMM was fitted using the Expectation-Maximization algorithm, with the Bayesian Information Criterion (BIC) selecting a VVE covariance structure (ellipsoidal clusters with equal orientation) and three components as optimal. The model yielded a log-likelihood of –904.57, BIC = –1898.25, and ICL = –1942.89, confirming partition stability. The 380 observations were assigned to three groups (n = 75, 220, and 85), consistent with previous clustering results, demonstrating the robustness of the three-cluster solution and GMM's ability to capture clusters of varying shapes and volumes. Fig. 14 illustrates the classification**.**

Journal Pre-proof

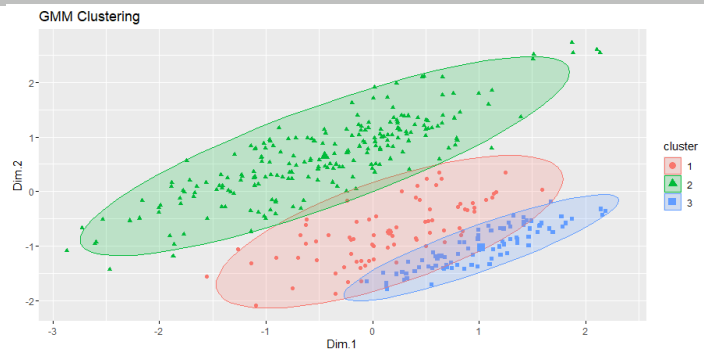


Fig. 14: GMM clustering results

3.3 Evaluation and comparison of clustering methods

3.3.1 Internal validation

All four clustering algorithms supported a three-cluster solution, although their validation scores differed (Table 8). K-means showed the best overall balance, with the highest mean Silhouette score (0.513) and the lowest Davies–Bouldin index (0.710). K-medoids produced a comparable partition, with the highest Calinski–Harabasz index (416.666) and a Davies–Bouldin value close to K-means. Ward's method obtained the highest Dunn index, while GMM showed weaker performance, with a lower mean silhouette score (0.281) and a higher DB index (1.354). However, this result should be interpreted cautiously, since GMM is a probabilistic clustering approach that allows partial membership and overlapping group structures. Its three-cluster solution remained broadly consistent with the overall segmentation pattern, but the overlap between groups was likely penalised by indices designed for hard partitions.

K-means was therefore retained as the main clustering solution because it provided the clearest compromise between statistical performance and socio-technical interpretability. The mean Silhouette value of 0.513 indicates a reasonable, rather than strong, cluster structure [66]. The retained clusters are therefore interpreted as meaningful household profiles with moderate separation, rather than perfectly discrete categories.

Table 8: Internal validation indices for the compared clustering algorithms

Clustering Algorithm	Mean silhouette (MS)	Calinski-Harabasz (CH)	Davies-Bouldin (DB)	Dunn Index
Hierarchical by Ward	**0.488**	**342.117**	0.728	**0.025**

Journal Pre-proof

K-means	**0.513**	411.447	**0.710**	0.019
K-medoid	0.477	**416.666**	0.715	0.018
GMM	0.281	168.310	1.354	0.008

3.3.2 Cluster stability analysis

The resampling-based stability analysis confirms the consistency of the retained clustering solution. As shown in Table 9, the three agreement metrics present high mean values across the 100 subsampling iterations, with an ARI of 0.954, a NMI of 0.942 and a pairwise Jaccard similarity of 0.942. These values indicate strong agreement between the reference partition and the partitions obtained from repeated subsamples.

The low standard deviations further suggest limited variability across iterations. In addition, the minimum values remain high for all three metrics, with ARI = 0.889, NMI = 0.862 and Jaccard similarity = 0.862. This indicates that even in the least favourable subsamples, the cluster assignments remained largely consistent. Overall, the results support the stability of the identified household profiles under moderate changes in sample composition.

Table 9: Resampling-based stability results for the retained clustering solution

	ARI	NMI	Jaccard Similarity
mean	0.954	0.941	0.942
std	0.033	0.041	0.042
min	0.889	0.862	0.862
max	1	1	1

3.4 Gower-distance-based sensitivity analysis

The Gower-distance-based sensitivity analysis supported the selection of a three-cluster solution. Among the tested values of k, the highest Silhouette value was obtained for k = 3 (0.458), followed by k = 4 (0.438) and k = 2 (0.432) (Table 10). Although the difference between these solutions remains moderate, k = 3 provides the best balance between cluster separation, parsimony and interpretability. The decrease in Silhouette values from k = 5 onwards suggests that additional clusters would mainly fragment the sample without improving the overall structure. These results therefore indicate that the three-profile solution is not solely an artefact of the PCA-based procedure, but remains consistent when mixed-type variables are analysed using Gower dissimilarity.

Table 10: Gower-distance-based sensitivity analysis for mixed-type variables

k	Silhouette Gower
2	0.432
3	**0.458**
4	0.438
5	0.397
6	0.386
7	0.387

3.5 Profiles identification

The clustering analysis identified three household profiles characterised by distinct combinations of energy consumption, dwelling characteristics, socio-economic attributes, and behavioural indicators. Comparison of cluster centroids reveals distinct energy demand profiles, with Cluster 01 (n = 138) occupying an intermediate position, Cluster 02 (n = 113) representing the lowest-consuming households, and Cluster 03 (n = 129) comprising the highest-consuming households (Table 11, Fig.15).

Table 11: Silhouette means and cluster sizes

(K-Means)

	Cluster 01	Cluster 02	Cl
Silhouette width	0.56	0.44	(
Observations	138	113	
Percentage (%)	36.31	29.73	3

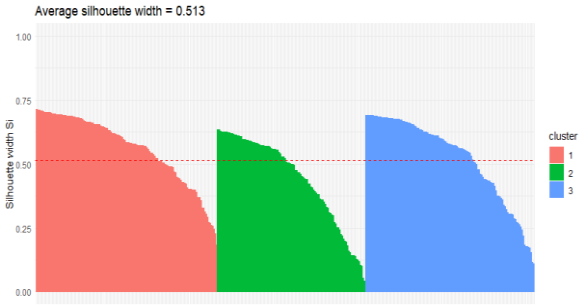


Fig. 15: average silhouette width of K-means clustering

Quantitative variables were summarised in Table 12, for each cluster using mean values and standard deviations in order to describe both central tendencies and within-cluster dispersion. For categorical, binary and ordinal survey variables, response distributions were calculated by cluster and reported as counts and percentages in Supplementary material, Appendix C (Table C.2). These descriptive statistics were used to construct an operational typology of household energy profiles and to assess the comparability of the three clusters.

Table 12: The mean values of analysed quantitative variables in each cluster

Variable	Cluster 01	Cluster 02	Cluster 03

Journal Pre-proof

	Mean	SD	Mean	SD	Mean	SD
Energy consumption (kWh)	16870.5	0.15	14244.8	0.17	19055.7	0.16
Willingness and engagement to energy efficiency index	6.52	2.23	5.46	2.47	6.52	2.61
Energy sobriety index (ESI)	8.78	1.85	9.33	2.34	8.88	2.05
Building's period of construction	1999	4.98	1990	6.51	1998	7.53
Dwelling area (m^2)	79.97	8.05	69.25	5.00	84.26	9.99
Dwelling perimeter (m)	42.27	3.34	36.13	4.17	44.12	5.17
Windows area (m^2)	12.26	3.22	9.02	2.49	13.04	3.79
Volume gross (m^3)	237.92	25.57	198.59	16.67	253.96	33.91
THLP (m)	34.40	6.39	27.38	5.01	38.05	5.23
THLA (m^2)	115.59	36.31	93.48	32.48	131.08	36.32
U windows (W/m²/°C)	2.98	0.54	2.30	0.46	2.60	0.76
U door (W/m²/°C)	2.01	0.21	2.68	1.12	4.09	0.93
Family size	5.16	1.61	4.45	1.76	4.76	2.12
Weekend occupancy ratio	0.82	0.29	0.91	0.20	0.82	0.28
Week Occupancy ratio	0.80	0.29	0.81	0.27	0.77	0.29

Cluster 01: Intermediate households, with moderate consumption and a strong readiness to invest in energy efficiency.

Represents households with intermediate energy consumption (≈16,870 kWh annually). Dwellings are of medium size (≈80 m²), with construction periods concentrated around the late 1990s. Households exhibit moderate energy sobriety (0.23) but relatively strong willingness to engage in energy efficiency (6.52). Retrofit activity (0.88) and window replacement (0.75) are comparatively common, indicating a degree of proactive improvement of building components. Equipment ownership and air-conditioning use remain moderate but contribute to overall energy demand.

Cluster 02: Energy-sober households, low consumption, and limited willingness to pay for efficiency improvements.

Corresponds to the lowest energy-consuming households, with an average annual consumption of about 14,245 kWh. These dwellings are smaller (≈69 m²) and exhibit lower building volumes and envelope dimensions. Households also show slightly smaller family sizes (4.45 persons) and lower equipment ownership. Despite recording the highest Energy Sobriety Index (0.54), engagement in energy efficiency measures is comparatively lower (5.45), and retrofit activities are less frequent.

Journal Pre-proof

Combined with lower income levels, these characteristics suggest that reduced energy consumption may partly reflect structural and economic constraints rather than proactive efficiency behaviour.

Cluster 03: Energy-intensive households, less sustainable in behaviour but with a stronger willingness to engage in efficiency measures.

Represents the most energy-intensive households, with an average annual consumption of approximately 19,000 kWh (Fig. 16). These dwellings are the largest in the sample (≈84 m²) and display the highest building volumes and envelope dimensions. Larger family sizes (5.22 persons), high equipment ownership (HEI = 9.91), and widespread air-conditioning use (0.87) are also observed. Although households exhibit the lowest energy sobriety index (0.20), their willingness to engage in energy efficiency measures is relatively high (6.52). Retrofit interventions are relatively frequent (0.84), though window replacement remains limited (0.09), suggesting that improvements may focus more on equipment upgrades than on envelope retrofits.

**A comparison of the three clusters reveals a socio-technical gradient in residential energy demand (Fig. 16) Cluster 03 exhibits higher consumption due to larger dwellings, bigger households, and higher incomes, while Cluster 02 comprises lower-income households in smaller homes with fewer appliances. Occupancy patterns differ, with Cluster 02 more present at weekends and Cluster 03 slightly higher on weekdays. These findings indicate that structural and socio-economic factors, dwelling size, household composition, and equipment ownership; largely drive energy use, as similar levels of willingness to adopt efficiency measures do not translate into equivalent consumption.

4. Discussion

The three clusters are interpreted here as socio-technical household profiles rather than simple consumption categories. They reflect different combinations of household resources, dwelling characteristics, equipment ownership and energy-related practices.

Accordingly, Cluster 03 is interpreted as an affluent energy-intensive profile, Cluster 02 as a resource-constrained low-consumption profile, and Cluster 01 as a transition-oriented profile characterised by moderate consumption and favourable behavioural readiness.

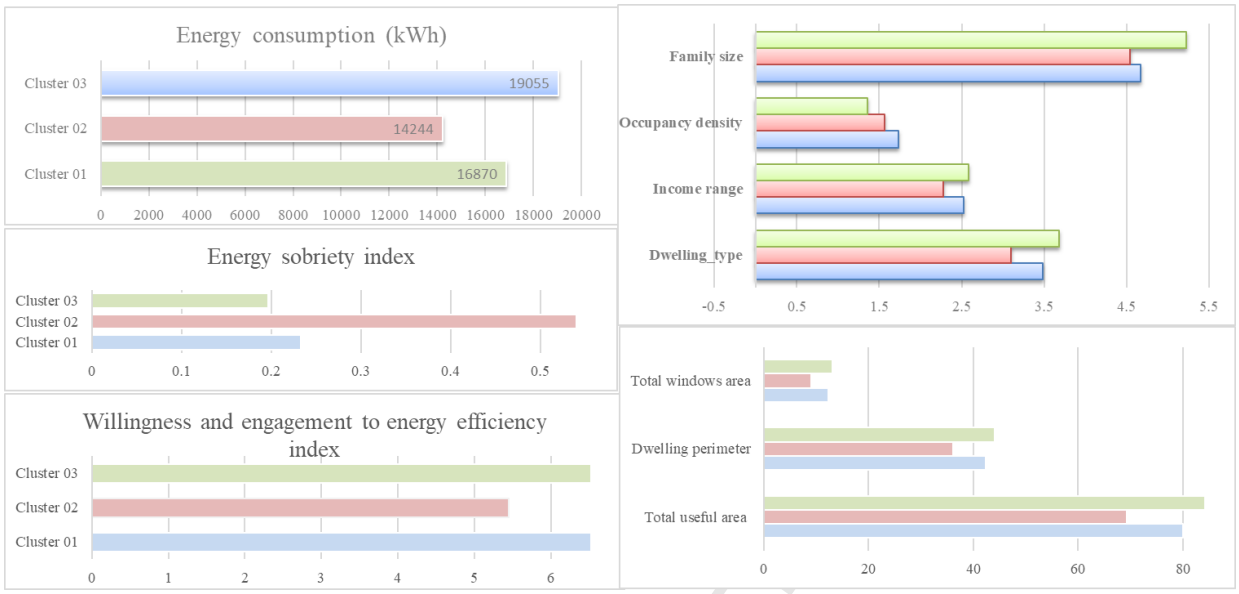


Fig. 16: Multi-dimensional characterisation of household clusters

4.1 Residential energy profiles reflect socio-technical lifestyles rather than isolated determinants

The three clusters indicate that residential energy demand is shaped by socio-technical lifestyles rather than by isolated socio-demographic, behavioural or building-related variables. In this study, consumption profiles emerged from the combined effects of dwelling geometry, envelope characteristics, equipment ownership, household composition, income and everyday practices. This supports previous work showing that socio-demographic factors and behavioural patterns jointly influence residential energy use [77], while strong interactions between building size, occupancy and practices make it difficult to isolate independent effects [78].

Building-related variables also need to be interpreted within this broader configuration. Larger areas, volumes and window-related dimensions were associated with higher demand, in line with studies linking dwelling geometry to heating, cooling, hot water and ventilation needs [79,80]. However, thermal indicators such as U-values provide only partial insight when considered separately [55].

Journal Pre-proof

High consumption may persist where envelope weaknesses, intensive occupancy and equipment-intensive routines interact.

These findings support an interpretation of the clusters as integrated socio-technical energy lifestyles. For policy, this implies that efficiency measures should not target single determinants alone, but rather the combined household, building and behavioural configurations through which energy is actually consumed [14,78,81,82].

4.2 Low consumption may indicate constrained efficiency rather than sustainable behaviour

The results challenge the assumption that low energy consumption necessarily reflects sustainable or environmentally conscious behaviour. In this study, the lowest-consumption profile was also associated with lower income, limited equipment ownership, few or no dwelling retrofits, and weaker willingness to invest in energy-efficiency measures. Conversely, low consumption may reflect income constraints rather than stronger environmental commitment.

This interpretation is consistent with studies showing that low-income households may delay, reduce or avoid the use of energy services, particularly cooling, because of budget constraints despite comparable comfort needs [83]. Such energy-limiting practices can make households appear efficient in consumption data, while actually reflecting under-consumption and restricted access to adequate thermal comfort [84]. Similarly, income- or bill-based indicators may overlook households that maintain low bills through severe self-restriction [85].

Therefore, Cluster 02 should not be interpreted simply as an energy-efficient or sustainable group. Its lower energy demand may indicate hidden energy deprivation, whereby households reduce consumption at the expense of comfort, health, or everyday well-being. From a policy perspective, this implies that low-consumption households should not be excluded from energy-efficiency support. Instead, interventions should distinguish between voluntary energy sobriety and constrained

Journal Pre-proof

under-consumption, combining efficiency improvements with measures that protect minimum comfort levels and reduce the risk of energy poverty.

4.3 Declared willingness for energy efficiency does not necessarily reduce realised consumption

A key finding of this study is that declared willingness to engage in energy efficiency does not automatically translate into lower realised energy consumption. The comparison between Cluster 01 and Cluster 03 illustrates this attitude-behaviour gap: both profiles show relatively high willingness towards energy efficiency, yet Cluster 03 records substantially higher consumption. This suggests that positive attitudes are not sufficient to offset the effects of larger dwellings, higher equipment ownership, comfort expectations and energy-intensive routines.

This finding is consistent with previous studies showing that environmental concern and energy-saving intentions do not reliably predict actual household energy use. Attitudes may influence intentions and low-impact self-reported behaviours, but their relationship with measured consumption remains weak [86]. Similar attitude-behaviour gaps have been reported in Beijing [87], while evidence from France suggests that energy-saving behaviours and deep efficiency investments may diverge, partly due to rebound effects and comfort-seeking practices[88]. In the UK, environmental viewpoints were weakly associated with specific efficiency technologies or retrofit intentions [89]. Cross-national evidence also shows that reductions in household energy expenditure are mainly achieved through curtailment, whereas efficient product ownership has a more limited effect [90].

Therefore, willingness should be interpreted as an enabling condition rather than a direct predictor of lower consumption. In this study, high willingness among high-consuming households may reflect awareness without sufficient structural, financial or technical capacity to reduce demand. Conversely, low consumption may reflect involuntary efficiency linked to income constraints rather than stronger environmental commitment [91,92].

Journal Pre-proof

From a policy perspective, awareness campaigns alone are unlikely to reduce demand where consumption is driven by dwelling size, equipment intensity, comfort needs or inefficient envelopes. Moderate-consuming households with high behavioural readiness, such as Cluster 01, may represent a strategic leverage point for demand-side interventions, provided that willingness is supported by targeted incentives, technical assistance and accessible retrofit programmes [93,94].

4.4 Segmented energy lifestyles require differentiated policy pathways

The three socio-technical profiles suggest that energy policies for collective housing should be differentiated according to household lifestyles, rather than based on uniform interventions. Since the clusters combine income, equipment ownership, dwelling characteristics, behavioural readiness and realised consumption in different ways, each profile requires a specific policy pathway.

For Cluster 03, the affluent high-consumption profile, interventions should focus on reducing energy intensity through smart metering, dynamic tariffs, demand-response programmes, deep retrofits and efficient appliance replacement, while limiting potential rebound effects. For Cluster 02, lower consumption may reflect constrained use rather than successful efficiency. This profile should therefore be prioritised for energy-poverty programmes, financial support and subsidised retrofits aimed at improving both efficiency and thermal comfort. Cluster 01, with moderate consumption and relatively favourable behavioural readiness, appears better suited to voluntary efficiency programmes, behavioural interventions and renewable energy adoption, provided that information is combined with practical incentives and technical support.

Overall, lifestyle-based segmentation can help align energy policies with household constraints, capacities and practices, improving both policy effectiveness and social equity [78,84,86].

5. Conclusion

This study developed a multidimensional socio-technical approach to profile residential energy consumption in public collective housing in Bab Ezzouar, Algiers. By combining five-year electricity and natural gas consumption data with household survey responses, dwelling geometry

Journal Pre-proof

and envelope-related indicators, the analysis identified three household energy profiles. Among the tested clustering algorithms, K-means provided the best compromise between internal validity and socio-technical interpretability, with a reasonable three-cluster structure.

The findings show that residential energy demand results from the interaction between building, socio-demographic and behavioural factors. The three clusters therefore represent distinct socio-technical lifestyles requiring differentiated policy responses. Cluster 03, the affluent high-consumption profile, should be prioritised for deep retrofits, efficient appliance replacement, smart metering and demand-response measures, while limiting rebound effects. Cluster 02, the resource-constrained low-consumption profile, requires energy-poverty support, financial assistance and subsidised retrofits to improve both efficiency and thermal comfort. Cluster 01, the transition-oriented profile, appears most suitable for voluntary efficiency programmes, behavioural interventions and renewable-energy adoption support.

The main contribution of this study is therefore empirical and contextual. It provides evidence from an underrepresented Algerian and North African collective-housing context and demonstrates that household energy profiling can support more differentiated and socially sensitive policy design. Rather than applying uniform awareness campaigns or generic retrofit programmes, energy-efficiency strategies should be aligned with household capacities, constraints and energy-use practices.

This study has several limitations. The sample is limited to collective housing in one municipality and relies partly on voluntary participation, which may introduce response bias. Building performance was described mainly through geometric and envelope-related indicators, including U-values, while thermal inertia, air infiltration, solar gains and indoor comfort conditions were not directly measured. In addition, electricity and natural gas were aggregated into a single annual on-site energy consumption variable, preventing end-use distinctions between heating, cooking, cooling and appliances.

Journal Pre-proof

Methodologically, bootstrap-based cluster stability analysis and post-COVID sensitivity testing were not conducted, although the 2020–2021 period may have influenced residential demand. Finally, the study did not simulate retrofit scenarios or estimate potential kWh savings. Future research should therefore combine household profiling with end-use disaggregation, building energy simulation, longitudinal monitoring and scenario-based retrofit assessment.

Declaration of AI-assisted technologies in Scientific writing

During the preparation of this manuscript, the authors used Grammarly to improve language accuracy and readability. All content was carefully reviewed and edited by the authors. The analyses, interpretations, and conclusions are entirely the authors' own work. AI tools were used solely to enhance presentation, under full human oversight, and the authors take full responsibility for the content of the published article.

CRediT authorship contribution statement

Kahina Ouazar: conceptualization, data curation, methodology and design of the study. **Guglielmina Mutani**: Writing- review and editing, Final approval of the version to be submitted. **Leila Abbas Mansour**: supervision.

Declaration of competing interest

The authors declare that they have no financial or personal conflicts of interest that could inappropriately influence or bias their work.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request.

References

[1] IEA, IEA – International Energy Agency - IEA, (2023). https://www.iea.org/data-and-statistics (accessed January 2, 2026).

[2] X. Xu, H. Yu, Q. Sun, V.W.Y. Tam, A critical review of occupant energy consumption behavior in buildings: How we got here, where we are, and where we are headed, Renewable and Sustainable Energy Reviews 182 (2023) 113396. https://doi.org/10.1016/J.RSER.2023.113396.

[3] S. Hu, X. Zhou, D. Yan, F. Guo, T. Hong, Y. Jiang, A systematic review of building energy sufficiency towards energy and climate targets, Renewable and Sustainable Energy Reviews 181 (2023) 113316. https://doi.org/10.1016/J.RSER.2023.113316.

[4] K. Ouazar, L.A. Mansour, E. Eveno, K. Rahmani, A. Zemmit, A.A. Arafat, B.A. Jemberie, M. Alsharef, A. Althobaiti, A GIS and statistical based approach to assessing citizen awareness of the energy transition and climate change across algerian climatic zones, Scientific Reports 2026 (2026). https://doi.org/10.1038/s41598-026-43948-0.

[5] H. Yoshino, T. Hong, N. Nord, IEA EBC annex 53: Total energy use in buildings—Analysis and evaluation methods, Energy Build. 152 (2017) 124–136. https://doi.org/10.1016/J.ENBUILD.2017.07.038.

[6] H. Fang, W.K. Li, P. Dai, Parametric analysis of passive ultra-low energy building envelope performance in existing residential buildings, Scientific Reports 2025 15:1 15 (2025) 23407-. https://doi.org/10.1038/s41598-025-07421-8.

Journal Pre-proof

[7] X. Chen, Z. Xu, Y. Qiu, H. Hu, X. Wang, China's newest design of apartment buildings with modernized façade: A comparative evaluation of its energy performance in five major climate zones, Sustain. Cities Soc. 117 (2024) 105954. https://doi.org/10.1016/J.SCS.2024.105954.

[8] H. Samuelson, S. Claussnitzer, A. Goyal, Y. Chen, A. Romo-Castillo, Parametric energy simulation in early design: High-rise residential buildings in urban contexts, Build. Environ. 101 (2016) 19–31. https://doi.org/10.1016/J.BUILDENV.2016.02.018.

[9] H. Li, J. Zheng, Y. Zhang, R. Zhang, Urban morphology and residential building net-zero energy performance: A global meta-analysis, Energy Build. 346 (2025) 116142. https://doi.org/10.1016/J.ENBUILD.2025.116142.

[10] R. Alimohamadi, M.H. Jahangir, Multi-Objective optimization of energy consumption pattern in order to provide thermal comfort and reduce costs in a residential building, Energy Convers. Manag. 305 (2024) 118214. https://doi.org/10.1016/J.ENCONMAN.2024.118214.

[11] R. Lee, M. Choi, J. Yoon, D. Kim, Impacts of lighting and plug load variations on residential building energy consumption targeting zero energy building goals, Journal of Building Engineering 75 (2023) 106962. https://doi.org/10.1016/J.JOBE.2023.106962.

[12] S. Habibi, Energy-efficient building design optimization and renewable energy integration for zero-energy residential buildings using simulations, Energy & Environment (2025). https://doi.org/10.1177/0958305X251343057.

[13] A. Remizov, S.A. Memon, J.R. Kim, Novel building energy performance-based climate zoning enhanced with spatial constraint, Appl. Energy 355 (2024) 122238. https://doi.org/10.1016/J.APENERGY.2023.122238.

[14] B.F. Balvedi, E. Ghisi, R. Lamberts, A review of occupant behaviour in residential buildings, Energy Build. 174 (2018) 495–505. https://doi.org/10.1016/J.ENBUILD.2018.06.049.

[15] B. Yan, W. Yang, F. He, W. Zeng, Occupant behavior impact in buildings and the artificial intelligence-based techniques and data-driven approach solutions, Renewable and Sustainable Energy Reviews 184 (2023) 113372. https://doi.org/10.1016/J.RSER.2023.113372.

[16] M.N. Uddin, H.-H. Wei, H.-L. Chi, M. Ni, N. Tamanna, Building Layout Influence on Occupant's Energy Consumption Behaviour: An Agent-Based Modeling Approach, in: The 9th International Conference on Sustainable Development, MDPI, Basel Switzerland, 2022: p. 22. https://doi.org/10.3390/environsciproc2022015022.

[17] Y. Yuan, L. Gao, K. Zeng, Y. Chen, Space-Level air conditioner electricity consumption and occupant behavior analysis on a university campus, Energy Build. 300 (2023) 113646. https://doi.org/10.1016/J.ENBUILD.2023.113646.

[18] A. Mylonas, A. Tsangrassoulis, J. Pascual, Modelling occupant behaviour in residential buildings: A systematic literature review, Build. Environ. 265 (2024) 111959. https://doi.org/10.1016/J.BUILDENV.2024.111959.

[19] J. Rouleau, L. Gosselin, P. Blanchet, Robustness of energy consumption and comfort in high-performance residential building with respect to occupant behavior, Energy 188 (2019) 115978. https://doi.org/10.1016/J.ENERGY.2019.115978.

[20] J. Duan, N. Li, J. Peng, Q. Liu, T. Peng, S. Wang, Clustering and prediction of space cooling and heating energy consumption in high-rise residential buildings with the influence of occupant behaviour: Evidence from a survey in Changsha, China, Journal of Building Engineering 76 (2023) 107418. https://doi.org/10.1016/J.JOBE.2023.107418.

[21] V. Michalakopoulos, E. Sarmas, I. Papias, P. Skaloumpakas, V. Marinakis, H. Doukas, A machine learning-based framework for clustering residential electricity load profiles to enhance demand response programs, Appl. Energy 361 (2024) 122943. https://doi.org/10.1016/j.apenergy.2024.122943.

Journal Pre-proof

[22] X. Cui, M. Lee, M.N. Uddin, X. Zhang, V.G. Zakka, Analyzing different household energy use patterns using clustering and machine learning, Renewable and Sustainable Energy Reviews 212 (2025) 115335. https://doi.org/10.1016/j.rser.2025.115335.

[23] W. Tang, H. Wang, X.L. Lee, H.T. Yang, Machine learning approach to uncovering residential energy consumption patterns based on socioeconomic and smart meter data, Energy 240 (2022) 122500. https://doi.org/10.1016/j.energy.2021.122500.

[24] L. Czétány, V. Vámos, M. Horváth, Z. Szalay, A. Mota-Babiloni, Z. Deme-Bélafi, T. Csoknyai, Development of electricity consumption profiles of residential buildings based on smart meter data clustering, Energy Build. 252 (2021) 111376. https://doi.org/10.1016/j.enbuild.2021.111376.

[25] E. Zaidan, A. Abulibdeh, A. Alban, R. Jabbar, Motivation, preference, socioeconomic, and building features: New paradigm of analyzing electricity consumption in residential buildings, Build. Environ. 219 (2022) 109177. https://doi.org/10.1016/j.buildenv.2022.109177.

[26] K. Li, Z. Ma, D. Robinson, J. Ma, Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering, Appl. Energy 231 (2018) 331–342. https://doi.org/10.1016/J.APENERGY.2018.09.050.

[27] S. Kallel, M. Amayri, N. Bouguila, Clustering and Interpretability of Residential Electricity Demand Profiles, Sensors 2025, Vol. 25, 25 (2025). https://doi.org/10.3390/s25072026.

[28] G. Petricli, T. Inkaya, G. Gokay Emel, Identifying green citizen typologies by mining household-level survey data, Renewable and Sustainable Energy Reviews 189 (2024) 113957. https://doi.org/10.1016/J.RSER.2023.113957.

[29] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches for prediction and classification of building energy consumption, Renewable and Sustainable Energy Reviews 82 (2018) 1027–1047. https://doi.org/10.1016/j.rser.2017.09.108.

[30] Y. Wang, T. Cao, H. Wang, Q. Fu, J. Chen, Y. Lu, An interactive clustering method via metric learning for analyzing energy consumption patterns, Journal of Building Engineering 98 (2024) 111143. https://doi.org/10.1016/j.jobe.2024.111143.

[31] V.R. Chifu, T. Cioara, I. Danci, C.B. Pop, I. Anghel, Whale optimization-based clustering for buildings' energy consumption profiles, Energy Reports 14 (2025) 1207–1221. https://doi.org/10.1016/j.egyr.2025.07.034.

[32] B. Kim, Y. Yamaguchi, Y. Shimoda, Physics-based modeling of electricity load profile of commercial building stock considering building system composition and occupancy profile, Energy Build. 304 (2024) 113813. https://doi.org/10.1016/J.ENBUILD.2023.113813.

[33] M.H. Khajedehi, J. Vivian, A. Zarrella, Identification of Italian user archetypes for building energy simulations based on a national survey, Energy Build. 344 (2025) 115924. https://doi.org/10.1016/j.enbuild.2025.115924.

[34] E. Sarmas, A. Fragkiadaki, V. Marinakis, Explainable AI-Based Ensemble Clustering for Load Profiling and Demand Response, Energies 2024, Vol. 17, Page 5559 17 (2024) 5559. https://doi.org/10.3390/EN17225559.

[35] C. Hennig, Cluster-wise assessment of cluster stability, Comput. Stat. Data Anal. 52 (2007) 258–271. https://doi.org/10.1016/J.CSDA.2006.11.025.

[36] D. Hsu, Identifying key variables and interactions in statistical models of building energy consumption using regularization, Energy 83 (2015) 144–155. https://doi.org/10.1016/J.ENERGY.2015.02.008.

[37] Y. Fang, J. Wang, Selection of the number of clusters via the bootstrap method, Comput. Stat. Data Anal. 56 (2012) 468–477. https://doi.org/10.1016/J.CSDA.2011.09.003.

[38] R. Kaur, D. Gabrijelčič, Behavior segmentation of electricity consumption patterns: A cluster analytical approach, Knowl. Based. Syst. 251 (2022) 109236. https://doi.org/10.1016/j.knosys.2022.109236.

[39] N. Carbonare, T. Pflug, A. Wagner, Clustering the occupant behavior in residential buildings: a method comparison, Bauphysik 40 (2018) 427–433. https://doi.org/10.1002/bapi.201800023.

Journal Pre-proof

[40] R. Tolas, R. Portase, R. Potolea, From Individual Device Usage to Household Energy Consumption Profiling, Electronics 2024, Vol. 13, 13 (2024). https://doi.org/10.3390/electronics13122325.

[41] D. Krishnan, S. Kelly, Y. Kim, A Meta-Analysis Review of Occupant Behaviour Models for Assessing Demand-Side Energy Consumption, Energies 2022, Vol. 15, 15 (2022). https://doi.org/10.3390/en15031219.

[42] G. Ruellan, S. Attia, G. Haesbroeck, Clustering of archetypal building-inhabitant pairs to improve energy efficiency: The case of the Walloon region in Belgium, Energy Build. 335 (2025) 115549. https://doi.org/10.1016/j.enbuild.2025.115549.

[43] M. De Simone, L. Callea, G. Fajilla, Surveys and inferential statistics to analyze contextual and personal factors influencing domestic hot water systems and usage profiles in residential buildings of Southern Italy, Energy Build. 255 (2022) 111660. https://doi.org/10.1016/j.enbuild.2021.111660.

[44] T. Malatesta, J.K. Breadsell, Identifying Home System of Practices for Energy Use with K-Means Clustering Techniques, Sustainability 2022, Vol. 14, 14 (2022). https://doi.org/10.3390/su14159017.

[45] L. Messoudi, A. Gouareh, B. Settou, C. Mokhtara, N. Settou, Modeling and forecasting energy consumption in Algerian residential buildings using a bottom-up GIS approach, Energy Build. 317 (2024) 114370. https://doi.org/10.1016/J.ENBUILD.2024.114370.

[46] M. Afaifia, M. Boulahia, K.A. Djiar, N.A. Lamraoui, A.N. Mansouri, L. Milat, S.C. Serrai, J. Teller, Energy Benchmarking Analysis of Multi-Family Housing Unit in Algiers, Algeria, Sustainability 2025, Vol. 17, Page 4120 17 (2025) 4120. https://doi.org/10.3390/SU17094120.

[47] J. García-López, S. Domínguez-Amarillo, J.J. Sendra, Clustering Open Data for Predictive Modeling of Residential Energy Consumption across Variable Scales: A Case Study in Andalusia, Spain, Buildings 2024, Vol. 14, Page 2335 14 (2024) 2335. https://doi.org/10.3390/BUILDINGS14082335.

[48] H. Rafiq, P. Manandhar, E. Rodriguez-Ubinas, J.D. Barbosa, O.A. Qureshi, Analysis of residential electricity consumption patterns utilizing smart-meter data: Dubai as a case study, Energy Build. 291 (2023) 113103. https://doi.org/10.1016/J.ENBUILD.2023.113103.

[49] ONS (2008), National Statistics Office, (n.d.). https://www.ons.dz/ (accessed March 26, 2026).

[50] CNERIB (2016) DTR C3.2/4, Building Thermal Regulation, (n.d.). https://www.cnerib.edu.dz/ (accessed December 1, 2025).

[51] M. Traich, A. Rahmane, A Decomposition Analysis of Algeria's Residential Energy Consumption Change: (2000–2020), Strategic Planning for Energy and the Environment 43 (2024) 151–176. https://doi.org/10.13052/spee1048-5236.4317.

[52] APRUE, National Agency for the Promotion and Rationalisation of Energy Use, (2023). https://aprue.org.dz/fr/ (accessed March 28, 2026).

[53] M. Afaifia, K.A. Djiar, N. Bich-Ngoc, J. Teller, An energy consumption model for the Algerian residential building's stock, based on a triangular approach: Geographic Information System (GIS), regression analysis and hierarchical cluster analysis, Sustain. Cities Soc. 74 (2021) 103191. https://doi.org/10.1016/j.scs.2021.103191.

[54] S. Hu, D. Yan, S. Guo, Y. Cui, B. Dong, A survey on energy consumption and energy usage behavior of households and residential building in urban China, Energy Build. 148 (2017) 366–378. https://doi.org/10.1016/J.ENBUILD.2017.03.064.

[55] L. Li, Y. Wang, M. Wang, W. Hu, Y. Sun, Impacts of multiple factors on energy consumption of aging residential buildings based on a system dynamics model--Taking Northwest China as an example, Journal of Building Engineering 44 (2021) 102595. https://doi.org/10.1016/j.jobe.2021.102595.

[56] Declaration of Helsinki, WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Participants – WMA – The World Medical Association, (1964). https://www.wma.net/policies-post/wma-declaration-of-helsinki/ (accessed March 26, 2026).

[57] SONELGAZ, Sonelgaz, (2024). https://www.sonelgaz.dz/fr (accessed December 1, 2025).

Journal Pre-proof

[58] K. Pearson, VII. Note on regression and inheritance in the case of two parents, Proceedings of the Royal Society of London 58 (1895) 240–242. https://doi.org/10.1098/RSPL.1895.0041.

[59] G. Mutani, S. Santantonio, A. Tartaglia, Statistical Data Analysis for Energy Communities, TECNICA ITALIANA-Italian Journal of Engineering Science 64 (2020) 385–397. https://doi.org/10.18280/ti-ijes.642-438.

[60] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374 (2016) 20150202. https://doi.org/10.1098/rsta.2015.0202.

[61] H.F. Kaiser, An index of factorial simplicity, Psychometrika 39 (1974) 31–36. https://doi.org/10.1007/BF02291575/METRICS.

[62] M.S. Bartlett, The Effect of Standardization on a χ 2 Approximation in Factor Analysis, Biometrika 38 (1951) 337. https://doi.org/10.2307/2332580.

[63] J.H. Ward, Hierarchical Grouping to Optimize an Objective Function, J. Am. Stat. Assoc. 58 (1963) 236. https://doi.org/10.2307/2282967.

[64] F. Murtagh, P. Legendre, Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?, J. Classif. 31 (2014) 274–295. https://doi.org/10.1007/s00357-014-9161-z.

[65] A. Likas, N. Vlassis, J. J. Verbeek, The global k-means clustering algorithm, Pattern Recognit. 36 (2003) 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2.

[66] Leonard. Kaufman, P.J.. Rousseeuw, Finding Groups in Data : an Introduction to Cluster Analysis, (1990) 368. https://doi.org/10.1002/9780470316801.

[67] F. Bin, A. Abid, A Novel Approach for PAM Clustering Method, Int. J. Comput. Appl. 86 (2014) 1–5. https://doi.org/10.5120/15074-3039.

[68] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

[69] G. McLachlan, D. Peel, Finite Mixture Models, Wiley, 2000. https://doi.org/10.1002/0471721182.

[70] J.C. Gower, A General Coefficient of Similarity and Some of Its Properties, Biometrics 27 (1971) 857. https://doi.org/10.2307/2528823.

[71] J. Hendrickson Bachelor, Methods for Clustering Mixed Data, 2014. https://scholarcommons.sc.edu/etd/2590 (accessed June 7, 2026).

[72] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Communications in Statistics 3 (1974) 1–27. https://doi.org/10.1080/03610927408827101.

[73] D.L. Davies, D.W. Bouldin, A Cluster Separation Measure, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1 (1979) 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.

[74] J.C. Dunn†, Well-Separated Clusters and Optimal Fuzzy Partitions, Journal of Cybernetics 4 (1974) 95–104. https://doi.org/10.1080/01969727408546059.

[75] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1985) 193–218. https://doi.org/10.1007/BF01908075/METRICS.

[76] A. Strehl, J. Ghosh, Cluster Ensembles-A Knowledge Reuse Framework for Combining Multiple Partitions, Journal of Machine Learning Research 3 (2002) 583–617.

[77] M. Shamsaiee, U. Eicker, Socio-Demographic insights on urban building energy consumption, Energy Build. 352 (2026) 116793. https://doi.org/10.1016/j.enbuild.2025.116793.

[78] Y. Xie, A.I. Mohd Noor, Factors Affecting Residential End-Use Energy: Multiple Regression Analysis Based on Buildings, Households, Lifestyles, and Equipment, Buildings 2022, Vol. 12, Page 538 12 (2022) 538. https://doi.org/10.3390/BUILDINGS12050538.

Journal Pre-proof

[79] S.J. Lee, S.Y. Song, Determinants of residential end-use energy: Effects of buildings, sociodemographics, and household appliances, Energy Build. 257 (2022). https://doi.org/10.1016/j.enbuild.2021.111782.

[80] E.L. Ofetotse, E.A. Essah, R. Yao, Evaluating the determinants of household electricity consumption using cluster analysis, Journal of Building Engineering 43 (2021) 102487. https://doi.org/10.1016/j.jobe.2021.102487.

[81] Y. Zhang, X. Bai, F.P. Mills, Characterizing energy-related occupant behavior in residential buildings: Evidence from a survey in Beijing, China, Energy Build. 214 (2020) 109823. https://doi.org/10.1016/j.enbuild.2020.109823.

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

[83] S. Cong, D. Nock, Y.L. Qiu, B. Xing, Unveiling hidden energy poverty using the energy equity gap, Nature Communications 2022 13:1 13 (2022) 2456-. https://doi.org/10.1038/S41467-022-30146-5.

[84] K. Eisfeld, S. Seebauer, The energy austerity pitfall: Linking hidden energy poverty with self-restriction in household use in Austria, Energy Res. Soc. Sci. 84 (2022) 102427. https://doi.org/10.1016/J.ERSS.2021.102427.

[85] B. Menyhért, Energy poverty in the European Union. The art of kaleidoscopic measurement, Energy Policy 190 (2024) 114160. https://doi.org/10.1016/J.ENPOL.2024.114160.

[86] G. Carrus, L. Tiberio, S. Mastandrea, P. Chokrai, I. Fritsche, C.A. Klöckner, T. Masson, S. Vesely, A. Panno, Psychological Predictors of Energy Saving Behavior: A Meta-Analytic Approach, Front. Psychol. 12 (2021) 648221. https://doi.org/10.3389/FPSYG.2021.648221/FULL.

[87] Y. Zhang, X. Bai, F.P. Mills, J.C.V. Pezzey, Examining the attitude-behavior gap in residential energy use: Empirical evidence from a large-scale survey in Beijing, China, J. Clean. Prod. 295 (2021) 126510. https://doi.org/10.1016/J.JCLEPRO.2021.126510.

[88] F. Belaïd, Decarbonizing the Residential Sector: How Prominent is Household Energy-Saving Behavior in Decision Making?, Energy Journal 45 (2024) 125–148. https://doi.org/10.5547/01956574.45.1.FBEL;SUBPAGE:STRING:ABSTRACT;JOURNAL:JOURNAL:ENJA;WEBSITE:WEBSITE:SAGE;WGROUP:STRING:PUBLICATION.

[89] M. Pelenur, Household energy use: a study investigating viewpoints towards energy efficiency technologies and behaviour, Energy Efficiency 2018 11:7 11 (2018) 1825–1846. https://doi.org/10.1007/S12053-018-9624-X.

[90] X. Piao, S. Managi, Household energy-saving behavior, its consumption, and life satisfaction in 37 countries, Scientific Reports 2023 13:1 13 (2023) 1382-. https://doi.org/10.1038/s41598-023-28368-8.

[91] J. Shim, D. Song, Unveiling energy inefficiencies: A study on building energy consumption in single-person households, Renewable and Sustainable Energy Reviews 214 (2025) 115546. https://doi.org/10.1016/j.rser.2025.115546.

[92] R.K. Moussa, D.K. Kanga, Elicitation of electricity consumption habit formation among new subscribers, Heliyon 10 (2024). https://doi.org/10.1016/j.heliyon.2024.e28360.

[93] M.M. Jaber, D. Qasim, A. Ghaith, What Drives Sustainable Energy Use? A Case Study from Jordanian Households, Environments 2025, Vol. 12, 12 (2025). https://doi.org/10.3390/environments12120452.

[94] K. Chadwick, R. Russell-Bennett, N. Biddle, The role of human influences on adoption and rejection of energy technology: A systematised critical review of the literature on household energy transitions, Energy Res. Soc. Sci. 89 (2022). https://doi.org/10.1016/j.erss.2022.102528.

Journal Pre-proof