Journal of Cleaner Production 467 (2024) 142870

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

# Driving forces and typologies behind household energy consumption disparities in China: A machine learning-based approach

Yi Wu [a], Yixuan Zhang [b,*], Yifan Li [c], Chenrui Xu [d], Shixing Yang [e], Xi Liang [a,e]

[a] The Bartlett School of Sustainable Construction, University College London, London WC1E 7HB, UK
[b] Cardiff Business School, Cardiff University, Cardiff CF10 3EU, UK
[c] Graduate School of Engineering, University of Tokyo, Tokyo 1138654, Japan
[d] School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, UK
[e] UK-China (Guangdong) CCUS Centre, Guangzhou 510440, China

ARTICLE INFO

Handling Editor: Cecilia Maria Villas Bôas de Almeida

*Keywords:*
Household energy consumption
Energy consumption inequality
Machine learning approach
Household typology

ABSTRACT

Establishing an intuitive link between driving factors of household energy consumption activities and inequalities is important for the understanding of household heterogeneity in energy consumption behaviours. This paper proposes a novel typology framework based on machine learning approaches and data from 3637 Chinese households in 2014 from 85 cities. Activity-based energy consumption was measured, highlighting inequalities across activities, regions and household types. The results showed significant energy consumption disparities between urban/rural and north/south households, especially in cooking, space heating and vehicle activities. By identifying driving factors of energy consumption, a new household typology classified samples into 6 (all), 6 (urban) and 7 (rural) types. Within these types, households with similar demographic structures, lifestyles and energy consumption habits were clustered. Demographic structure, region, and primary energy demand were used as the basis for the typology. The findings demonstrated how household lifestyle differences explained the cause and underlying driving factors of urban-rural energy consumption inequalities and provided suggestions for city-by-city and type-by-type measurements to support effective low-carbon transformation in cities.

## 1. Introduction

Cities contain 55% of the world's population and contribute to two-thirds of global energy consumption and 70% of the world's greenhouse gas emissions (Seto et al., 2014; World Bank, United Nations Development Programme and Global Infrastructure Facility, 2020). The rapid urbanisation of cities has resulted in significant increases in energy consumption in both the economic sector and household sector (Huang et al., 2023; Luqman et al., 2023). The household sector is a major contributor to greenhouse gas (GHG) emissions, accounting for more than 60% of global emissions (Hertwich and Peters, 2009; Ivanova et al., 2016; Long et al., 2022; Nejat et al., 2015). In recent years, energy consumption and GHG emissions in China have been rising rapidly, among which the share of the household sector has been increasing significantly (Fan et al., 2017; Miao, 2017). As estimated, urbanisation contributed 15.4% to the increase in residential energy consumption during 1996–2012 (Fan et al., 2017). The household sector is the second largest energy consumption sector, following the industrial sector, accounting for about 10% of total energy consumption (Lu and Liu, 2014; Miao, 2017). Meanwhile, China has long been pushing energy-efficiency and emission reduction policies in both industrial and household sectors, for instance, setting up targets for energy conservation in the national Five-Year Plan, launching the "Top 1000 Enterprises Energy Conservation Action Program" (Guilhot, 2022), and has provided energy-saving guide and education in the household sector (Kuai et al., 2022). Especially, China has implemented energy efficiency label policies for essential electric appliances, aiming to encourage household energy-saving purchases and usage (Lei et al., 2022). In rural areas, it is vital to pay more attention to household energy use activities as rural households are expected to have larger potential for energy saving (Zhang et al., 2022). Compared to urban residents, rural households face multiple choices of traditional energy sources, such as woods, briquettes, and other biomass products. However, they have limited access to cleaner energy due to the wealth gap (Ma et al., 2022). Fortunately, the rise of modernisation and rural-targeted policy implementation has greatly improved energy access in rural households and reduced energy

* Corresponding author.
*E-mail address:* zhangy352@cardiff.ac.uk (Y. Zhang).

https://doi.org/10.1016/j.jclepro.2024.142870
Received 25 March 2024; Received in revised form 17 May 2024; Accepted 10 June 2024
Available online 10 June 2024

inequality, such as the “Home Appliances Go to Countryside (HAGC)" program (Ji et al., 2019) and “Automobiles Go to Countryside (AGC)" program (Hoken and Sato, 2022; Zheng, 2020).

Above all, given the emphasised potential in household energy conservation and emission mitigation (Du et al., 2021), it is of great significance to figure out what drives households to lower energy consumption and to gain a deeper understanding of the relationship between the drivers as well as household energy consumption (HEC) and per capita household energy consumption (PHEC) in the context of China’s carbon peaking and carbon neutrality targets. Furthermore, analysing factors that can depict specific household behaviours and corresponding lifestyles is valuable for formulating refined energy conservation and climate change mitigation policies because the effectiveness of interventions is highly conditional on the prior knowledge of household’s energy consumption behaviours (Long et al., 2021, 2024).

Currently, there is an increasing body of literature that explores various driving factors of HEC and PHEC in China (Lei et al., 2022; Shi et al., 2021; Wu et al., 2017; Wu et al., 2019b; Zheng et al., 2014; Zhou and Yang, 2016), and demonstrates inequality problems between urban and rural areas (Dou et al., 2021; Ma et al., 2021; Shi, 2019; Wu et al., 2017) or even regional inequality regarding energy consumption (Shi et al., 2020b; Zhang and Zhou, 2020). However, most studies adopted top-down approaches to analyse the contribution of various consumption domains to HEC. For example, Liu et al. (2021) combined input-output tables with provincial data to decompose HEC by domains in which housing, food, and transport were recognised as the main drivers. Dou et al. (2021) found a positive correlation between energy inequality and household $CO_2$ emissions in China’s 30 provinces for the period 2000–2017 and lowering energy inequality can thus reduce household $CO_2$ emissions.

In terms of bottom-up approaches, a majority of studies have focused on measuring HEC and energy consumption inequality, while research on the underlying drivers behind them remains underdeveloped. Zheng et al. (2014) and Zheng and Wei (2019) designed and implemented the Chinese Residential Energy Consumption Survey (CRECS) 2013 and 2014, measuring energy consumption inequality rural HEC in China at the household level. Moreover, Shi (2019) examined inequality of opportunity in energy consumption expenditure with samples of 1912 households in China using data from the Chinese General Social Survey (CGSS) 2015 and the heterogeneity analysis in this study identified disadvantaged groups who were facing unequal opportunities for energy consumption in China, especially in developing areas. Mao et al. (2020) created a typology with four clustered groups for Tujia and Miao rural households in Chongqing city, China, identifying household energy consumption patterns based on lifestyles and production demand, e.g. ownership of appliances and pig-raising activities. Zou and Luo (2019) examined a set of energy-related determinants of rural households based on a Tobit model using rural household samples from CGSS 2015, showing that health, off-farm occupation, education, and economic condition significantly affected energy consumption among rural households. Wu et al. (2019a) offered an overview of household energy consumption in rural China in 2013 and concluded with five end-use activities (cooking, home appliance, space heating, water heating, and space cooling) and six energy types (coal, gas, liquefied petroleum gas, solar power, biomass, grid power, and district heating). Additionally, using household samples in Japan, Chen et al. (2022) applied a machine learning approach to deepen the analysis of driving factors for household activity-based emissions by exploring the importance of socioeconomic characteristics and household lifestyles. Nevertheless, household heterogeneity and heterogeneity-induced energy consumption inequality are often ignored in the existing studies. Only a few common features at the household level are frequently discussed in the existing studies, such as geographic and socioeconomic attributes, as well as demographic structure, while complicated features regarding household energy habits and lifestyles were often skipped. For example, Lei et al. (2022) explored HEC patterns by classifying based on heterogeneity, but the criteria were limited in features of income level, age span and family size. In this case, there is a gap about how detailed HEC behaviours affect household energy consumption and inequality and no existing studies, to the best of our knowledge, have attempted to classify households into reasonable groups based on lifestyle differences, on which refined energy policies can be implemented. To address the issue, new approaches should be applied to maximise the utilisation of granularity data, such as supervised and unsupervised learning approaches (Chen et al., 2022; Long et al., 2024). It is noteworthy that some studies have started to apply unsupervised learning approaches in analysis of HEC patterns (Tang et al., 2022) and carbon footprint patterns (Froemelt and Wiedmann, 2020). However, little understanding was developed about the casual relation between household behaviours and HEC patterns; instead, they focused on the carbon footprint of household consumption expenditures and the temporal patterns of HEC.

In response to the above gaps, firstly this study calibrates the estimation of HEC based on the household-level detailed energy consumption activities, including the duration and frequency of appliance use, as well as the power of appliances; and then we propose a new method that classifies households into fine-grained clusters with similar features. This method consists of four steps and applies multiple machine learning models. Secondly, compared to the existing literature, we innovatively develop a new household typology to characterize the inherent relationship between HEC and energy consumption drivers, which can help infer the sources of HEC differences. The advantage of this typology is that it circumvents potential endogeneity issues and selection bias that may arise from modelling HEC directly or subjectively selecting features; instead, it can explain variations in HEC independently based on differences in household habits and lifestyles. Thirdly, this paper complements the understanding of inequality in energy consumption by revealing the convergence and divergence in the specific energy demands of clustered households. This provides novel implications into tailored and targeted energy policies in China and the other developing regions, and more valuable insights in comparative studies between urban and rural areas from the perspective of household lifestyle.

The remainder of this paper is structured as follows. Section 2 provides detailed methods and data applied in the paper. Section 3 presents the results regarding city-level and household-level analyses. The last section concludes the findings in the paper and discusses policy implications and limitations.

## 2. Methods and data description

### 2.1. Survey samples

In this study, the survey CRECS 2014, which also forms the energy data section of CGSS 2015, has been applied for the following analysis. Although the survey was implemented in 2014, its consistent research value, particularly in the field of household-level energy activities, consumption and urban-rural comparative studies, merits emphasis. The advantage of the survey data is its coverage of detailed socioeconomic characteristics of households, description of energy activities and habits, as well as the bottom-up data collection at the appliance level, making it the optimal choice for investigating the issues addressed in this study. In other relevant research areas, the data still holds undeniable value, such as in studies on energy poverty (Cheng et al., 2023), internet behaviour and energy conservation practices (Guo et al., 2023), the effect of environmental regulations on household emissions (Cheng et al., 2024) and the emission mitigation potential of residential heating (Wang and Wei, 2024).

There are 3863 households from 85 cities and 28 provinces in China including relatively even samples from urban and rural areas (55% for rural and 45% for urban) in the data. The data covers multiple dimensions of household characteristics, including the habit of appliance use and details (type, quantity, purchase year, purchase subsidies, fuel type, energy efficiency labelling, power, time, frequency, effective area, etc.), family members' personal information (number of family members, age, family relations, house ownership, etc.), household economic conditions including annual income, expenditure, etc. Eventually, in this study, variables are categorised into six major types covering the household characteristics of our interest to be included in the modelling procedures where geographic, demographic, economic, living, family-relationship and energy-consumption variables are included.

### *2.2. Household energy consumption estimation and inequality analysis*

Measuring household energy activities is the key part of the paper, which is also the key basis for the accurate estimation of HEC. We refer to the processing and parameters applied by Wu et al. (2017), Wu (2019) and Li et al. (2022), and carefully translate the detailed information from the raw survey. Throughout the updated estimation strategy, household energy activities are transformed into direct energy consumption with the unit of kilogram standard coal equivalent (kgce). Notably, as the survey does not cover indirect energy consumption or embedded energy consumption in household's nonenergy consumption, the scope of HEC in this study is limited to direct energy consumption only. Specifically, the estimation consists of three steps. First, we estimate the unit of energy consumption or power of appliances. Taking cookers as an example, due to the great difference in appliance and energy types between the urban and rural, we assume a fixed heat load of cookers, which is 2 kW on average. Regarding electric appliances, it is relatively easy to retrieve the power of appliances and for those without detailed information, we infer the possible power of appliances based on energy efficiency labels, otherwise, we skip the specific appliance in estimation if both information is missing. Second, we estimate the actual volume of energy activity. Except for space heating, the other activities can simply be estimated by frequency and service time, while for space heating, an average energy consumption of space heating will be calculated based on the modes of space heating for a specific user, e.g. district and distributed space heating. And then, heating days can be collected in the survey. Note that for frequently used home appliances, such as lighting bulbs, fridges, and televisions, we assume the use frequency equals the stay-in days in a household. Third, we identify energy types and corresponding standard coal equivalent coefficients in each energy activity. The coefficients are collected from National Yearbooks (NBS, 2023) and published studies (Li et al., 2022; Wu et al., 2019). Part of parameters are updated in the estimation because the correct parameters should be published before the survey year. The estimation equation is described as follows:

$$E_i = \sum\nolimits_m \sum\nolimits_j e_{ijm} \times f_j \tag{1}$$

where $E_i$ represents the household $i$'s total energy consumption (unit: kgce). $e_{ijm}$ denotes household $i$'s energy consumption activity of energy type $j$ from activity $m$ and $f_j$ represents the coal equivalent coefficient of energy type $j$. Additionally, taking the example of electric appliances, the estimation is simplified to be the product of unit power (energy consumption), service time, frequency and coal equivalent coefficient of electricity, which is as follows:

$$E_{appliance} = \sum\nolimits_n U_n \times T_n \times F_n \times f_n \tag{2}$$

where $E_{appliance}$ is the sum of electric appliances' energy consumption. $U_n, T_n, F_n, f_n$ denote unit power or unit energy consumption of appliance $n$, service time, frequency and coal equivalent coefficient of electricity, respectively. Nevertheless, the study tried hard to retain data accuracy in estimation by including every detailed difference inherent in each energy consumption activity. The full description of the method for household energy consumption estimation is available in Supplementary Material S1.

Notably, to simplify the analysis of category-specific energy consumption modes and merge data with relatively small values, we compress the original ten categories of energy consumption activities into six categories, which means energy consumption from (1) air conditioners, (2) freezers, (3) laundry machines, (4) televisions, (5) computers, and (6) lighting bulbs are summed up and the new category is labelled as "Appliance".

In addition, inequalities related to energy consumption have long been a concern (Dou et al., 2021; Ma et al., 2021; Shi, 2019). We are interested in the energy consumption inequality between urban and rural and in different energy consumption activities. Following the idea of using the Gini coefficient and Lorenz curve to measure energy consumption inequality (Wu et al., 2017), we calculate the Gini coefficient of energy consumption by taking household as the unit and the coefficient is defined as follows:

$$Gini = 1 - \left| \sum_{i=1}^{N} (X_{i+1} - X_i)(Y_{i+1} + Y_i) \right| \tag{3}$$

where $X$ denotes the cumulative proportion of households in samples, and $Y$ is the cumulative proportion of energy consumption. $X_i$ represents household $i$ in samples with an ascending index, and accordingly, $Y_i$ is the energy consumption of household $i$ reordered ascendingly. Besides, we also calculate the Gini coefficient separately for different energy consumption activities to illustrate the potential inequalities associated with lifestyles.

### *2.3. Machine learning approaches*

The study proposes a synthetic machine learning modelling framework to establish a new typology for households, which consists of three steps: (1) data preprocessing, (2) optimal feature selection, (3) optimised clustering process.

#### *2.3.1. Data preprocessing*

Different formats of raw questionnaire data will be transformed into categorical or numeric variables to meet the requirements of the machine learning modelling processing.

First, we exclude samples with zero total energy consumption, resulting in a final dataset comprising 3637 households. Then, to mitigate possible biases stemming from units and issues of skewness, particularly to resolve heteroscedasticity (Akbari and Haghighat, 2021; Huebner et al., 2016), all variables except dummy variables are normalised with z-score normalisation. Besides, dependent variables, i.e. HEC, PHEC, and activity-specific HEC are not included in modelling in order to avoid biases.

#### *2.3.2. Optimal feature selection*

As shown in Table 1, due to the aim of accurately describing potential driving factors for disparities in HEC, we have included a plethora of variables to characterise various household features. However, modelling high-dimensional data with low-dimensional methods may lead to unexpected biases (Chernozhukov et al., 2018). Generally, many studies employ linear regression models to reduce data dimensions and implement feature selection such as Ridge Regression and Least Absolute Shrinkage and Selection Operator (LASSO) Regression (Hoerl and Kennard, 1970; Tibshirani, 1996). The LASSO model has been widely applied in studies focusing on the characteristics of household energy consumption and emissions (Chen et al., 2022; Huebner et al., 2016; Maruejols et al., 2022; Mashhadi and Behdad, 2018; Shi et al. 2020). Compared to traditional methods, this form of feature selection demonstrates high efficiency and offers a high level of predictive accuracy (Tibshirani, 1996; Zhang et al., 2019). Therefore, to address the requirement for handling high-dimensional data and preserving estimation of confidence intervals, we follow the ideas from the more recent literature (Chernozhukov et al., 2018; van de Geer et al., 2014; Zhang and Zhang, 2014), employing the debiased LASSO model to fit the data and obtain estimated coefficients for feature selection. This

**Table 1**
Variables applied in the modelling processing.

| Variables | Count | Mean | Min | Max | Converted |
|---|---|---|---|---|---|
| **Geographic** | | | | | |
| Prefecture | 3637 | 38.660 | 0 | 84 | × |
| Region (north/south) | 3637 | 0.550 | 0 | 1 | × |
| **Demographic** | | | | | |
| Mean age | 3637 | 49.771 | 17 | 92 | ✓ |
| House area | 3637 | 2.264 | 0 | 5 | ✓ |
| Family size | 3637 | 2.865 | 1 | 13 | ✓ |
| No. of children | 3637 | 0.499 | 0 | 6 | ✓ |
| No. of elderly | 3637 | 0.973 | 0 | 9 | ✓ |
| **Economic** | | | | | |
| Annual expenditure | 3062 | 36,032 | 0 | 999999 | ✓ |
| Annual income | 3330 | 72,053 | 0 | 999999 | ✓ |
| **Living** | | | | | |
| Stay-out days | 3637 | 1.851 | 0 | 6 | ✓ |
| Stay-in days | 3637 | 6.707 | 0 | 7 | ✓ |
| **Family-structure** | | | | | |
| Is single elderly family | 3637 | 0.042 | 0 | 1 | × |
| Is single adult with elderly family | 3637 | 0.072 | 0 | 1 | × |
| Is single adult only family | 3637 | 0.030 | 0 | 1 | × |
| Is couple elderly family | 3637 | 0.146 | 0 | 1 | × |
| Is couple adult family | 3637 | 0.063 | 0 | 1 | × |
| Is single family with children | 3637 | 0.022 | 0 | 1 | × |
| Is couple family with children | 3637 | 0.132 | 0 | 1 | × |
| Is grandparenting family | 3637 | 0.016 | 0 | 1 | × |
| Is big family | 3637 | 0.222 | 0 | 1 | × |
| Is family with elderly | 3637 | 0.615 | 0 | 1 | × |
| **Energy-consumption** | | | | | |
| No. of cookers | 3637 | 2.669 | 0 | 10 | ✓ |
| Mean power of cookers | 2379 | 872 | 150 | 1500 | ✓ |
| Mean use freq. of cookers | 3606 | 2.021 | 0.033 | 3 | ✓ |
| Mean use time of cookers | 3603 | 37.203 | 15 | 360 | ✓ |
| No. of water heaters | 3637 | 0.567 | 0 | 3 | ✓ |
| Mean use freq. of water heaters | 1969 | 1.419 | 0.033 | 3 | ✓ |
| Mean use time of water heaters | 1969 | 38.532 | 15 | 360 | ✓ |
| Mean energy efficiency of water heaters | 1278 | 1.511 | 1 | 5 | ✓ |
| No. of air conditioners | 3637 | 0.567 | 0 | 3 | ✓ |
| Mean use frequency of air conditioners | 1556 | 2.409 | 1 | 6 | ✓ |
| Mean power of air conditioners | 1351 | 3063 | 2600 | 4000 | ✓ |
| Mean use time of air conditioners | 1570 | 218 | 60 | 480 | ✓ |
| Mean energy efficiency of air conditioners | 1051 | 1.931 | 1 | 5 | ✓ |
| Type of space heating | 3637 | 2.173 | 0 | 3 | ✓ |
| Mean use time of space heating | 664 | 4.151 | 1 | 6 | ✓ |
| Space heating area | 667 | 65.465 | 10 | 120 | ✓ |
| Annual cost of space heating | 567 | 1657 | 0 | 20000 | ✓ |
| No. of vehicles | 3637 | 0.156 | 0 | 1 | ✓ |
| Fuel price for vehicle | 434 | 6.906 | 6 | 11 | ✓ |
| Cost of vehicle | 383 | 8631 | 0 | 60000 | ✓ |
| Annual driving distance | 402 | 0.532 | 0 | 7 | ✓ |
| Vehicle fuel type | 402 | 7.455 | 0 | 20 | ✓ |
| Actual vehicle displacement | 3637 | 2.669 | 0 | 10 | ✓ |
| Total energy consumption | 3637 | 1051 | 0.011 | 17756 | ✓ |
| Per capita energy consumption | 3637 | 465 | 0.005 | 13173 | ✓ |

**Note**: For households without vehicles, variables including driving distance, vehicle fuel type, actual vehicle displacement, etc., will be filled with zero values to be aligned with the distribution of vehicle ownership (No. of vehicles).
**Source**: the authors.

step is implemented using tools provided by EconML (Battocchi et al., 2019) in the Python 3.9 programming environment. In our study, the dependent variable $y_i$ is per capita energy consumption; independent variables include all variables in Table 1, except total energy consumption and per capita energy consumption.

### 2.3.3. *Optimised clustering process*

The third stage involves using an unsupervised learning model, specifically a clustering model, to identify households with similarities and differences in features of energy consumption activities and lifestyles; based on the optimal clustering results, a new typology is defined (Abu-Bakar et al., 2021; Alhussein et al., 2020; Froemelt et al., 2018; Yang et al., 2018; Zhou et al., 2017). Clustering is the process of dividing a dataset into several subsets based on given features, with the core logic of ensuring that objects within a cluster are similar to each other while objects between clusters are dissimilar. Representative clustering methods include centroid-based, density-based, distribution-based, partitioning clustering and hierarchical clustering approaches (Mahdi et al., 2021; Xu and Tian, 2015).

The most common method in the study of household energy consumption and climate-related patterns is K-means (Al-Wakeel et al., 2017; Chévez et al., 2017; Gianniou et al., 2018; Hincks et al., 2023; McLoughlin et al., 2015; Ofetotse et al., 2021). The advantages of the K-means method include ease of interpretation, simplicity of implementation, fast convergence, adaptability to sparse data, and good cluster recovery quality (Han et al., 2022; Maimon and Rokach, 2005; Milligan, 1980; Steinley, 2003). As a result, it is widely used in research on household energy consumption pattern recognition.

This paper also chooses the K-means clustering algorithm because households often exhibit grouping characteristics (or similar behaviours) in energy consumption activities. These characteristics may be determined by features including household economic conditions, demographic structure, energy-related activities, and lifestyles. These key features are identified and selected by the LASSO model. Therefore, in the sample, there are k representative households (centroids) and the remaining samples that are similar to these representative households. As a representative method of centroid-based clustering, the K-Means algorithm partitions the samples based on the similarity or distance of their features, aiming to minimise the distance between all points and their cluster centroids (Bandyopadhyay and Saha, 2013). This minimisation ensures accurate classification of points with other similar points, and all clusters have homogeneous subsets.

The steps for K-means clustering are as follows: (1) select $k$ initial samples as the initial cluster centres $a = a_1, a_2, \cdots, a_k$; (2) for each sample $x_i$ in the dataset, calculate its distance to the $k$ cluster centres and assign it to the class corresponding to the cluster centre with the minimum distance; (3) for each cluster $a_j$, recalculate its cluster centre $a_j = \frac{1}{|c_i|}\sum_{x \in c_i} x$ (i.e., the centroid of all samples belonging to that class); (4) repeat the above steps 2 and 3 until a termination condition is met (such as a maximum number of iterations or minimal change in error). The study explores the $k$ values ranging from 2 to 12 and the clustering results become completely random when $k$ exceeds 12. Metrics applicable to the K-means method include Euclidean distance, Manhattan distance, and cosine distance. In this study, given the multi-dimensional complexity of the samples, cosine distance is employed to capture relative differences in direction.

Clustering evaluation is usually performed using the cluster validity index, which includes the following popular metrics: Davies-Bouldin (Davies and Bouldin, 1979), Calinski-Harabasz (Caliński and Harabasz, 1974), and Silhouette (Rousseeuw, 1987). Among these, the Silhouette coefficient is particularly effective in assessing clustering performance (Shahapure and Nicholas, 2020). It measures the closeness of each point in a cluster to the points in the neighbouring clusters, ranging [−1, 1]. A coefficient close to −1 indicates objects assigned to the wrong cluster, close to zero suggests points not clearly in one cluster, and a coefficient close to 1 indicates objects very far from neighbouring clusters (Ofetotse et al., 2021). Therefore, this study uses the Silhouette coefficient ($s$) to determine the optimal number of clusters.

$$s = \frac{b - a}{\max(a, b)} \tag{4}$$

where $a$ represents the average distance from a sample point to all other points within the same cluster, indicating the similarity of the sample point to other points in the same cluster. $b$ represents the average distance from the sample point to all points in the nearest neighbouring cluster, reflecting the similarity of the sample point to other points in the nearest neighbouring cluster.

### 2.4. Energy consumption-based household typology

Based on the clustering results, we propose an unsupervised household typology that can describe distinct HEC patterns among different clusters. Specifically, based on the results of feature selection and by summarising the inter-cluster HEC disparities for different energy activities, we follow the approach outlined by Long et al. (2024) to propose a household typology based on underlying household characteristics, aiming to describe variations in energy-related behaviours and habits among different households.

## 3. Results

### 3.1. Household energy consumption disparity across cities

This section presents the geographic distribution of PHEC and city-level PHEC, further divided into urban and rural groups. Fig. 1 describes the distribution of PHEC in cities and the distributional disparities in PHEC in different groups. In general, the PHEC of the north is about 30% higher than that of the south in China, with the former being about 421.18 kgce and the latter being 323.89 kgce. Regarding the urban and rural disparities, the PHEC of the north (urban = 560.72 kgce, rural = 316.96 kgce) are 19% and 38.8% higher than those of the south (urban = 403.96 kgce, rural = 266.36 kgce) respectively.

These findings are higher than the results reported in the existing literature, where samples from 2013 are applied for estimation (Wu et al., 2019). In the meantime, deviations of PHEC in the northern households are much bigger and exist both in urban and rural areas. The largest deviation is in urban households of Hohhot city and rural households of Shenyang city. Furthermore, according to the PHEC distribution in the southern cities, the further south the city, the lower the PHEC. The lowest PHEC is found in Yuxi city, Yunnan province. However, some developed southern cities, such as Shanghai municipality and Hangzhou city, have relatively higher PHEC than the other northern cities. Huludao city, Shenyang city and Beijing municipality are ranked the top three cities with the highest PHEC on average.

To ensure the timeliness of the data does not significantly impact the results, we compare findings related to the HEC and PHEC in the latest publications, showing that the distribution of the HEC in 2018 closely resembles that of 2015, despite the national average HEC growing higher in 2018 than in 2015 (Wu et al., 2022). Similarly, Jiang et al. (2022) found that in 2021, in Guangzhou city (located in the south of China), the HEC in the central urban area and outskirts urban area (including energy consumption from home appliances, cooking, heating/cooling, and vehicles) are 177.5 kgce and 203.47 kgce respectively, which approximately align with our estimations.

Distributions of activity-based PHEC and appliance ownership are illustrated in Fig. 2. It holds the same city order as that in Fig. 1 to make both comparative. First of all, according to the distributions of five types of activities, the most significant difference exists in the distribution of space heating energy consumption. The northern households commonly have a huge share of space heating energy consumption across all types, while only a few of the southern cities have that. This divergence is caused by climate differences and the way of supplying space heating in the north, which is usually administrated centrally by the government, but no central space heating is available in the south. In contrast, it can be found that PHEC in water heating in the southern cities overpass those in the northern and PHEC in vehicles in some of the southern cities are relatively high in both regions. PHEC in cooking and appliances are evenly distributed. Regarding the appliance ownership distribution, no significant divergence is observed except that the southern cities have slightly more appliances than the northern ones, and it means that owning more appliances does not always indicate higher PHEC at the city level.

### 3.2. Household energy consumption inequality

Inequality has been a persistent concern in many studies, and it is also a multi-dimensional problem involving discussions in the energy consumption area. Numerous studies have attempted to illustrate the HEC inequality (Dou et al., 2021; Shi, 2019; Wu et al., 2017; Zhang and Zhou, 2020), but the findings regarding HEC remain limited due to the poor availability of data and inconsistent estimation methods. In addition, it is important to not only focus on the macro-level energy consumption inequality but understanding the intrinsic reason why HEC inequality emerges is also valuable. In this case, Lorenz curves in Fig. 3 present the inequality analyses from both the regional and activity-based perspectives and further emphasise the importance of exploring inequalities that exist in energy consumption activities. Firstly, our estimation shows that the inequality measure regarding HEC in urban and rural areas yields Gini coefficients of 0.504 and 0.527 and those in the south and north are 0.533 and 0.507, which is approximately consistent with the findings in the existing literature (Zheng et al., 2014). Secondly, it shows that the highest HEC inequality is observed in water heating-related energy consumption (the Gini coefficient = 0.755). The inequalities in space heating (the Gini coefficient = 0.546) confirms that households are faced with unevenly distributed and limited space heating resources, generally between the north and the south. In terms of appliance and cooking, as the most essential demand, the Gini coefficients of appliance and cooking are 0.548 and 0.592, where the inequality observed in cooking-related energy consumption suggests significant disparities among households, particularly in terms of cookers and cooking habits between rural and urban areas.

### 3.3. A machine learning-based HEC typology

#### 3.3.1. Results of feature selection

The driving factors with relatively high importance coefficients to PHEC are obtained through a feature selection process. Among them, energy activity-related and demographic features are dominant, as shown in Table 2. There are 8 variables with an absolute coefficient larger than 0.093 (the optimal threshold validated by the optimised clustering results) identified by the debiased LASSO model, and the optimal model is fitted with the best alpha value 0.0085.

Energy activity-related features, such as space heating, cooking and vehicles, are key drivers for PHEC. Specifically, utilisation of household appliances has positive impacts on PHEC, indicating that higher demand for appliances will significantly increase PHEC and relatively intensive energy consumption behaviours, like longer heating/cooking time, larger space heating area, more frequent use of water heaters, and farther driving distance are positively correlated with PHEC. Notably, it is estimated that the mean power of cookers is negatively correlated with PHEC, showing that less energy-efficient cookers (though they might have lower power) can cost more energy. Significantly, the penetration rate of private cars in the surveyed samples is relatively low (about 15.6%) but the actual vehicle displacement becomes a core factor in explaining PHEC. In this case, it can be argued that differences in

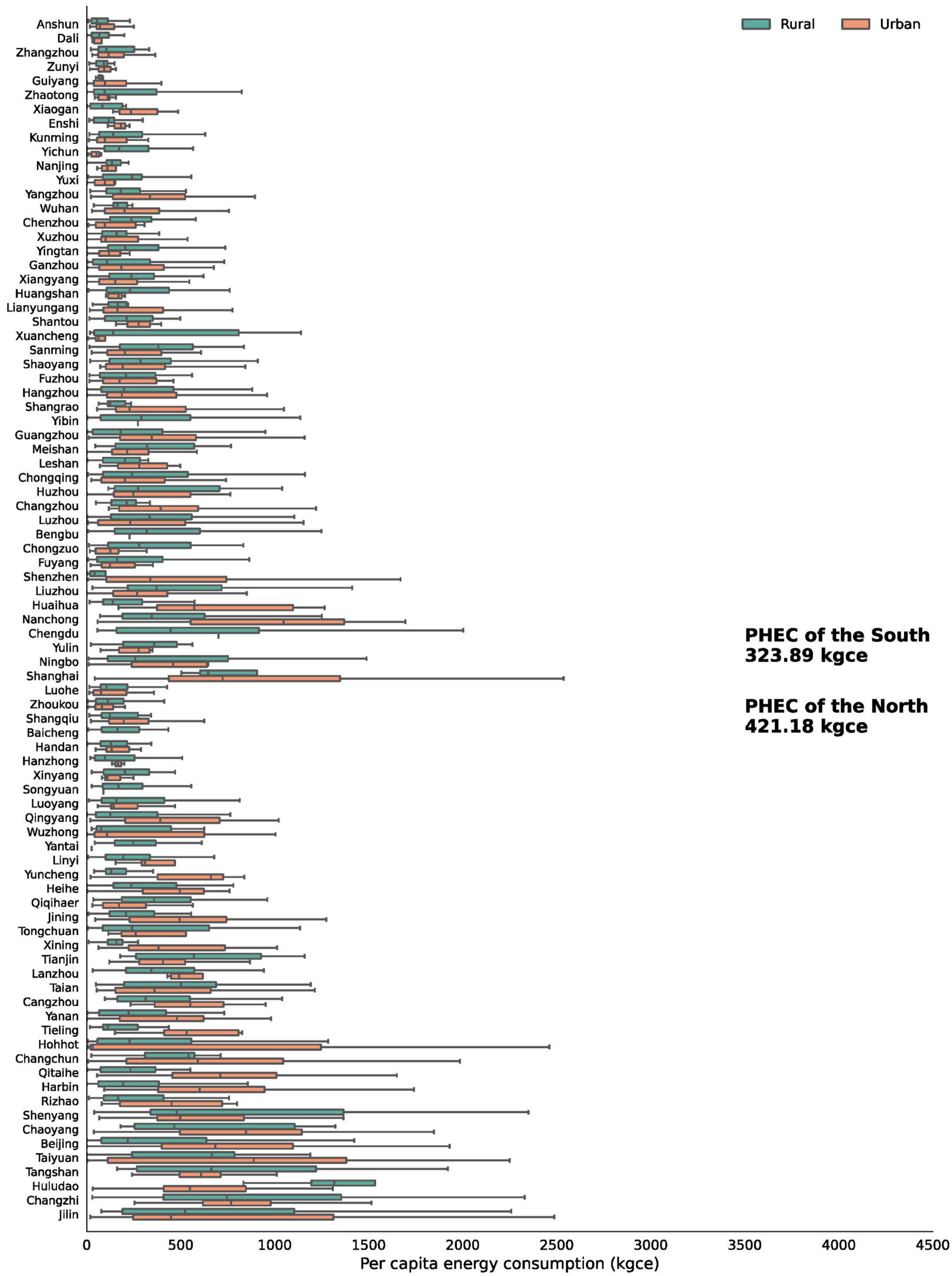


**Fig. 1. Per capita energy consumption across cities ordered by regions.** The boxplot shows the statistics within cities. Cities in the yellow area are southern cities, while those in the blue area are northern cities. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)
**Source**: the authors.


**Fig. 2. Category-specific per capita energy consumption and average ownership of appliances across cities.** Colours in the chart represent a specific category of energy consumption in a household's daily life, and there are six categories in total (appliance for light red, cooking for light green, space heating for blue, vehicle for yellow, water heating for dark red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)
**Source**: the authors.

household energy activities, appliance ownership, and disparities in energy use habits are likely to exacerbate inequality. As emerging technologies are iteratively developed and promoted, the context of HEC has become more complicated. It is vital to make decarbonisation technologies more accessible, affordable and inclusive while improving energy utilisation technologies.

In terms of demographic structures, PHEC is negatively correlated with family size, because the more people living in the same house, the more energy activities can be shared, such as air-conditioning, cooking, lighting, etc. In particular, in our samples, we find extended families,

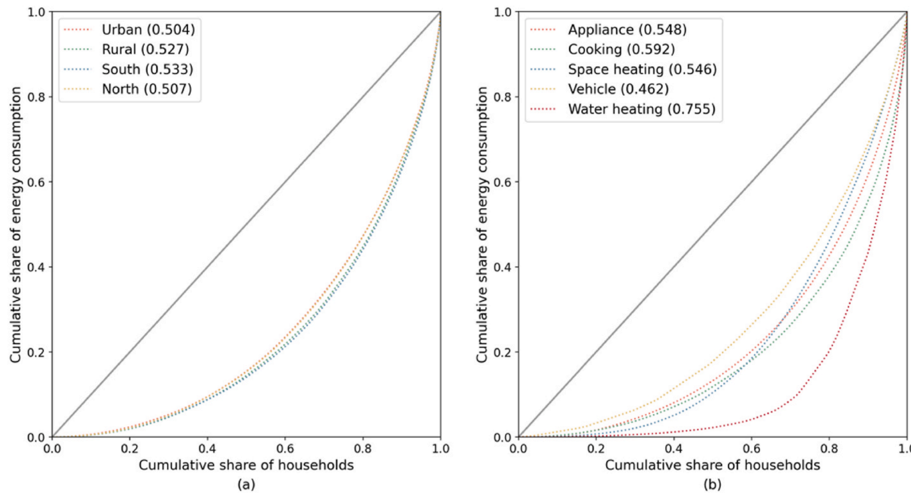


**Fig. 3. Lorenz curves by regions and energy consumption activities. Fig. (a)**, Lorenz curves of HEC by urban/rural and north/south regions. **Fig. (b)**, Lorenz curves of HEC by energy consumption activities including appliance, cooking, space heating, vehicle and water heating. The diagonal is the line of perfect equality. The numbers in the parentheses are the Gini coefficients.

**Table 2**
Coefficients, standard errors, and confidence intervals for optimally selected variables through the debiased LASSO model.

| Variables | Coefficient | Std. Errors | CI (95%) | |
|---|---|---|---|---|
| | | | Lower | Upper |
| Annual driving distance | 0.149 | 0.022 | 0.105 | 0.193 |
| Mean use frequency of water heaters | 0.148 | 0.021 | 0.107 | 0.189 |
| Mean use time of cookers | 0.140 | 0.015 | 0.112 | 0.169 |
| Mean use time of space heating | 0.110 | 0.035 | 0.041 | 0.180 |
| Space heating area | 0.103 | 0.027 | 0.049 | 0.156 |
| Actual vehicle displacement | 0.095 | 0.025 | 0.047 | 0.143 |
| Mean power of cookers | −0.104 | 0.015 | −0.134 | −0.073 |
| Family size | −0.284 | 0.018 | −0.319 | −0.248 |

**Source**: the authors.

defined as families with more than five members, are featured by relatively lower PHEC but unexpectedly higher energy demand in cooking. The reason is that extended families usually consist of adults, children and the elderly (i.e., three generations living together), and thus houses may be fully occupied during the day and the elderly at home usually play the role of caring for children. Extended families are commonly observed in China (especially in rural areas), and it represents a sustained energy consumption mode that occurs in houses. Note that the trend may change in the long term, as the typical demographic structure in China has been in transition, and familieshave been shrinking due to declining fertility rates (Long et al., 2024; Yu et al., 2018). Compared to the HEC activity-related features, factors such as household income and expenditure (economic) or house area and the average age of families are insignificant as their coefficients fall under the threshold for the optimal clustering.

### 3.3.2. *Results of clustering*

As described in Section 2.4, we chose the value of *k* corresponding to the highest Silhouette score as the optimal value for K-Means. Through experimentation within a range of importance weights for the variables and their coefficients obtained from the feature selection process, we aimed to find the minimum threshold and *k* value that would generate the highest Silhouette score and robust outcomes across three groups (all, urban, and rural samples). Detailed results are provided in the Supplementary Material S2. Eventually, the optimal threshold is 0.093, and the optimal values of *k* are 6, 6, and 7 for all, urban, and rural households, respectively.

Although the selected factors identified by the debiased LASSO model for PHEC are mainly related to energy activities, clustered households exhibit noticeable and superior inter-group differences in many features. Therefore, we propose a new household typology based on the clustering results and the underlying household features such as activity-based energy demand, age, region and demographic structure. According to the clustering results, we categorise the full samples into six groups and further divide urban and rural households into six and seven groups. Fig. 4 displays the new household typology sorted by PHEC.

In general, the typology includes seven clusters: Less-Energy-Demanded family (LED), Less-Energy-Demanded Extended family (LED-E), Less-Energy-Demanded Grandparenting family (LED-G), Water-heating-Demanded Southern family (WDS), Cooking-Demanded Elderly family (CDE), Heating-Demanded Northern family (HDN), and Driving-demanded High-income Younger family (DHY). It is evident that significant differences in PHEC can be seen in clusters of households, as shown in Fig. 4. Fig. 4(a) shows that LED-E has the lowest PHEC followed by LED-G while the PHEC of LED-G is almost twice as high as the former's. CDE and WDS have similar PHEC but there is a significant difference in energy demand between them where the former consumes more energy in cooking and the latter's PHEC is highly dependent on water heating. HDN ranks the second highest across all groups, and it is featured by its high consumption in space heating as most of the households clustered in this category live in the north of China. DHY has the highest PHEC, up to about 1000 kgce, featured by its significantly high vehicle demand and young population. In Fig. 4(b)–(c), groups are decomposed or reaggregated in either urban or rural samples based on the clustering results of full samples. Six out of seven clusters (LED-E, LED-G, WDS, CDE, HDN, and DHY) are shared between urban and rural samples from the full samples. For instance, LED-E and LED-G both remain in the urban and rural context because the extended and grandparenting families both account for a significant proportion of population. Furthermore, in terms of HDN and DHY, it is surprising to find that the two groups present opposite distributions of PHEC in urban and rural samples. In the rural context, high-income younger families (DHY) do not exhibit a significantly higher PHEC (569.8 kgce), which is much lower than that of DHY in urban samples (956.7 kgce) on average. It is noteworthy that the DHY with high vehicle demand, may see a future increase in HEC, driven by the growth in private transportation demand and the improvement of transportation infrastructure, especially in rural areas (Chen et al., 2019). HDN clusters exhibit robust distributions of PHEC across all, urban, and rural samples, as their

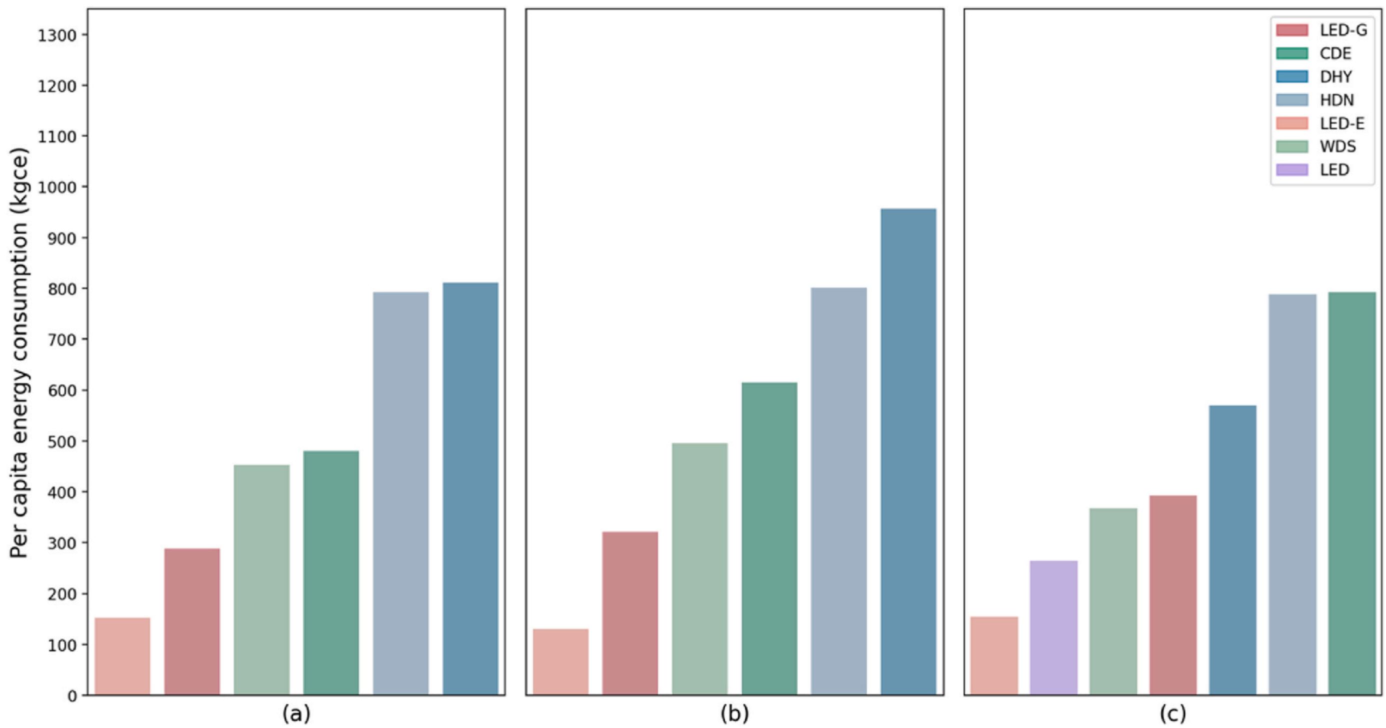


**Fig. 4. The new typology for clustered households in all samples and subsamples sorted by PHEC. Fig. (a)**, PHEC of six household types based on all samples. **Fig. (b)**, PHEC of six urban household types. **Fig. (c)**, PHEC of seven rural household types.
**Source**: the authors.

energy consumption is mainly driven by the demand for space heating.

*3.3.3. Inequalities in demand-specific clusters*

It is noteworthy that clustering aims to minimise intra-cluster differences based on the common features in samples. In this case, we might expect that through clustering, clustered households might have highly similar energy demands, especially in specific activities such as cooking or vehicle. By computing Gini coefficients for different clusters, we analyse how the inequality of HEC evolves in different samples. Fig. 5 shows that the disparities in HEC between different clusters manifest both convergence and divergence in specific energy demands. Specifically, the Gini coefficients of HEC converge in clusters such as HDN, DHY, and WDS while they diverge in clusters such as CDE, LED, LED-E and LED-G.

In Fig. 5(a)–(c), we emphasise that in both urban and rural samples, CDE exhibits the most uneven distribution of energy consumption because the energy consumption of CDE is dominated by cooking, and there are significant differences in the choice and use of cooking

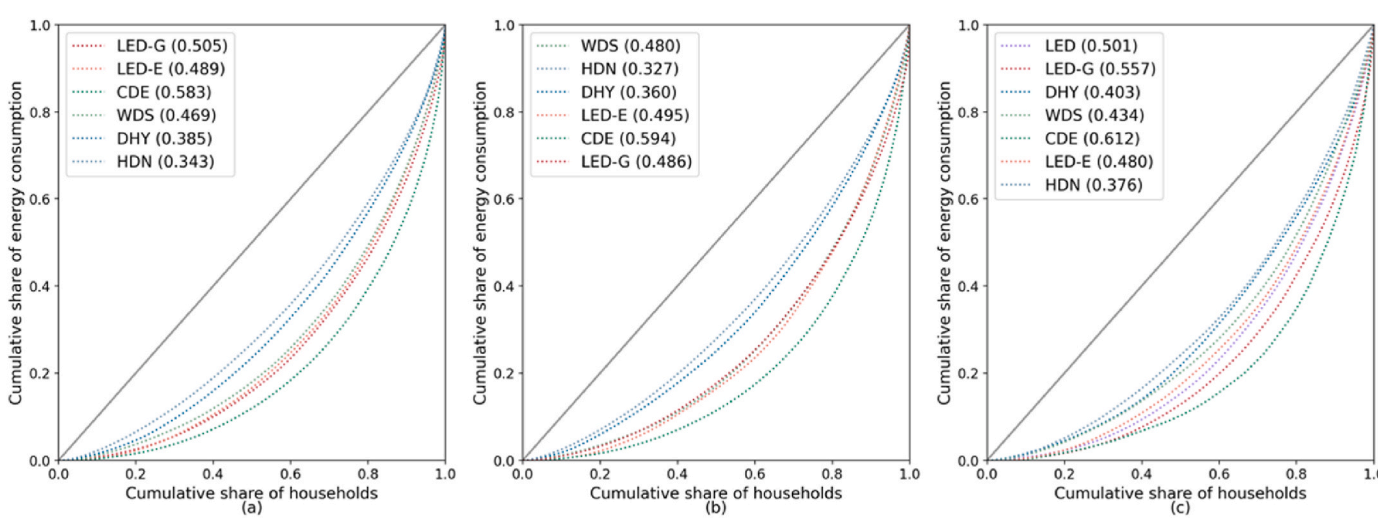


**Fig. 5. Lorenz curves by clustered groups. Fig. (a)**, Lorenz curves of HEC by clustered groups in all samples. **Fig. (b)**, Lorenz curves of HEC by clustered groups in rural samples. **Fig. (c)**, Lorenz curves of HEC by clustered groups in urban samples. The diagonal is the line of perfect equality. The numbers in the parentheses are the Gini coefficients. The legends represent the clustered groups including, Less-Energy-Demanded family (LED), Less-Energy-Demanded Extended family (LED-E), Less-Energy-Demanded Grandparenting family (LED-G), Cooking-Demanded Elderly family (CDE), Water-heating-Demanded Southern family (WDS), Heating-Demanded Northern family (HDN), and Driving-demanded High-income Younger family (DHY).
**Source**: the authors.

appliances between urban and rural households, for example, in rural areas, there are still households that use traditional firewood stoves fuelled by wood and crop residues. These stoves are less energy-efficient and more polluting than gas stoves. Moreover, LED-G groups have divergent scores in urban (the Gini coefficient = 0.486) and rural samples (the Gini coefficient = 0.557), and the divergence may stem from two aspects: (1) although LED-G households in urban and rural areas present highly similar energy-related habits and household characteristics, their energy demand may still be subject to potential resource constraints, such as the relatively lower level of electrification in rural areas compared to urban areas. (2) HEC in rural areas is generally prioritised to meet basic living needs, such as cooking, whereas households in urban areas may incur more energy expenses associated with water heating and home entertainment. To sum up, the clustering results indicate that although common household features can partly account for convergence in HEC, there are also complicated driving factors contributing to inequality in HEC that need further exploration.

### 3.3.4. Decomposition of driving factors in HEC disparities

In Fig. 6, we demonstrate the distributions of characteristics in different groups represented by the new typology through a heatmap with standardised household features ranging from 0 to 1. The new typology systematically classifies households by aspects including demographic structure, living habits, economic conditions, and energy demands (presented in rows in Fig. 6). Firstly, WDS and HDN families show strong region-specific characteristics, with over 86% of WDS families originating from southern China, while over 96% of HDN families are from northern China.

Secondly, regarding the demographic structure, LED-E clusters the extended families in all, urban and rural samples with sizes over 4.90, 4.95 and 4.92 on average per household. Besides, a clear divergence in the ages of DHY and CDE groups can be observed, with 41 years old for DHY and 56 years old for CDE on average. LED families are only observed in rural samples, and unlike LED-E and LED-G, these households have relatively lower levels of energy demand but do not exhibit a


**Fig. 6. A heatmap for household features among clusters within the new typology. Fig. (a)**, Features of six household types from all samples. **Fig. (b)**, Features of six urban household types. **Fig. (c)**, Features of seven rural household types.
**Source**: the authors.

prominent proportion in any energy activity. LED households are mostly nuclear families, and although they originate from rural areas, 6.4% of them own vehicles.

Thirdly, in DHY groups, samples are strongly clustered based on living habits (stay-out days) and economic conditions (annual income and annual expenditure). However, it is important to highlight that although we identify DHY groups in both urban and rural samples, the urban group has a significant economic advantage with its average annual income (¥175k) 177% higher than that of the rural group (¥99k) and average annual expenditure (¥78k) 163% higher than that of the rural group (¥48k).

Furthermore, there is strong heterogeneity between clusters in terms of living habits and energy activity-based demand. The observed differences can be concluded in five aspects: (1) appliance demand, which distinguishes LED-E and LED-G families in urban samples, accounting for 10% and 11% of HEC respectively, (2) cooking demand, which distinguishes CDE by its 57%–83% of energy consumption from cooking. In the urban context, CDE represents a typical southern elderly family, characterised by minimal outdoor activities, low expenditure, a strong tendency to stay at home, and a relatively strong demand for cooking and water heating than space heating; (3) water heating demand, which clusters a majority of samples from the south in WDS with more than 45% of its energy consumption coming from water heating. Additionally, WDS is also featured by its intensive uses of water heaters and air conditioners as both two types of appliances are the main temperature-controlling appliances applied in the south of China; (4) space heating demand, which yields great disparities between the north and the south in HDN groups due to the prevalent use of heating systems and central heating supply in the north. In contrast, southern households are excluded from the central heating supply and use air conditioners in summer and winter. Besides, though the northern households also use air conditioners for cooling in summer, the short duration and relatively low average temperature of summer in the north make the use of air conditioners less frequent than in the south; (5) vehicle demand, which is mainly found in younger high-income households (DHY) and associated with the demand in outdoor activities. Compared to other types of groups, DHY delineates distinctive features in energy consumption among young generations and underlying inequalities generated from consumption-oriented energy activities.

Overall, the DHY group projects potential inequalities in energy consumption in the context of a high-quality lifestyle compared to the other types of groups in 2014, when most households do not own private cars and participate in outdoor recreation. Comparing the shared groups of urban and rural samples (excluding LED), we see a significant difference in household demand, especially in cooking. Specifically, urban households have relatively stronger heating demand and may be more sensitive to temperature changes, while rural households exhibit higher energy consumption in cooking. These disparities can be explained not only by the urban-rural economic gap but also by the gap in electrification and household appliance ownership. It also reveals nonnegligible inequalities in the quality of urban and rural life. Particularly, the average power of cooking appliances and air conditioners in urban groups is 13.8% and 176% higher than that in rural groups, and the average energy efficiency rates of water heaters and air conditioners in urban groups are 114.9% and 149% higher than that in rural groups. In this case, the energy efficiency gap between the urban and rural indicates that compared to urban households' access to various appliances with higher energy efficiency, rural households are constrained by limited access to energy-efficient and cleaner appliances.

## 4. Discussion

This paper estimates HEC using granular survey data and recalibrates energy consumption activity-related features at the household level for a novel methodology proposed to create a typology for household classification based on their characteristics and energy consumption behaviours. The methodology consists of a three-step hybrid machine-learning approach for fine-grained clusters of households. The typology can explicitly illustrate common characteristics within clusters and differences between clusters. It also complements the understanding of driving factors for HEC and inequality dynamics in energy consumption by revealing that inequalities may be narrowed or broadened through clustering.

Based on the clustering results of households, we recommend that energy-saving policies and energy transition policies targeted at households should be more specific and tailored. This involves considering not only traditional socioeconomic and demographic characteristics but also lifestyle characteristics of different household types, urban/rural distinctions, and regional differences in demand between the southern and northern parts of China. This approach is crucial for achieving tailored policymaking that addresses different households' unique needs and circumstances. The corresponding policy suggestions are provided as follows.

Firstly, we recommend that policies should pay more attention to the young generation featured by advantageous economic conditions and high demand for mobility in both urban and rural areas since households in this group have the highest PHEC and present a huge space for energy conservation. By decomposing their lifestyles, we suggest that tailored policy measures can consider increasing the penetration of electric vehicles and providing incentives for energy-efficient appliance replacement; besides, the government can deepen the perception of green consumption among young families and foster environmental awareness.

Secondly, energy-saving policies between urban and rural areas should exhibit more flexible differentiations, focusing specifically on cooking, water heating, and space heating demands. For young or single individuals in urban areas, reducing their carbon footprint is necessary. In rural areas, policies should prioritise narrowing energy consumption inequality by improving energy access to various energy-efficient home appliances. Specifically, for southern households in rural areas, it is recommended to provide energy-saving incentives to lower the per capita energy consumption in water heating demand and space heating and to implement trade-in programmes for appliances with low energy efficiency rates. Additionally, intensify research and innovation for high-energy efficiency appliances are also recommended.

Thirdly, we recommend placing a greater emphasis on temperature adjustment-related energy demands. On the one hand, it involves improving the cleanliness of heating in northern regions. On the other hand, there is a need for better management of heating and cooling demands in southern regions, as inefficient and unregulated heating and cooling can lead to low energy efficiency and intensive energy consumption. Therefore, we suggest increasing the technological innovation and coverage of clean heating in the northern regions from a centralised management perspective. Furthermore, we propose expanding the use of energy-efficient appliances in the southern regions, including potential consumer subsidies and green consumption education.

Regarding inequality, based on the changes in the Gini coefficients through clustering, it can be concluded that the logic of grouping based on household lifestyles and the inherent energy consumption characteristics proves effective. This also implies that tailored policies should be designed for effective implementations. Taking into account the differentiated policies, it is important to not only focus on household types with similar trends but also to pay attention to those with outlier trends. Focusing on households with outlier trends might reduce their potential exposure risks to energy poverty (Lei et al., 2023; Xu et al., 2023).

Notably, this paper has limitation in the following areas. The accuracy of estimation is conditional on a few assumptions about the power of home appliances and household energy use habits, which may possibly lead to a slight overestimation or underestimation of HEC. For instance, the heat load of cookers might vary according to family size, as in rural areas, households may own bigger cookers than average to feed an extended family. Besides, the rest of the limitation is threefold. Firstly, there are missing values in the survey, which might result in underestimation or overestimation in HEC. For example, missing values in terms of the power of electric appliances, which are replaced by the average power of those with the same energy efficiency label, pose a challenge to estimate energy-based consumption. In particular, regarding the cookers that consume wood or briquettes, only if the exact capacity and heat load of them are given, the energy-based consumption can be estimated without biases. Thirdly, the gaps between the south and the north of China are determined by various natural, economic, and lifestyle factors. Although some of the deviations caused by the above gaps are well included in the modelling, there might be a few missing features that would affect the clustering. Lastly, since the CRECS 2014 ( CGSS 2015) data were collected approximately a decade ago, rapid modernisation in urban and rural China, coupled with the introduction of new technologies may have altered household's energy consumption demands and patterns. However, some studies have found that the impact of changes in household's demand and energy-related behaviours is more pronounced than that of equipment upgrades (Chen et al., 2023). This finding aligns with the goal of this study, highlighting that the development of typology offers better compatibility with structural changes in households than the analysis only based on household socioeconomic features. Additionally, climate change may also influence household's energy consumption demand. For instance, the increasing number of extreme hot days significantly raises household's demand for summer cooling (Jiang and Wei, 2024). However, the above bias also opens a gate for further exploration of household typology by incorporating more behavioural factors and technological changes.

## CRediT authorship contribution statement

**Yi Wu:** Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yixuan Zhang:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Yifan Li:** Visualization, Validation, Formal analysis. **Chenrui Xu:** Validation, Resources, Data curation. **Shixing Yang:** Validation, Data curation. **Xi Liang:** Validation, Resources.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The data that has been used is publicly available on the websites of Chinese Residential Energy Consumption Survey (http://crecs.ruc.edu.cn/sjxz/sjsysm/index.htm) and Chinese General Social Survey (http://cgss.ruc.edu.cn/info/1014/1015.htm) after regsitration and application processes.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclepro.2024.142870.

## References

Abu-Bakar, H., Williams, L., Hallett, S.H., 2021. Quantifying the impact of the COVID-19 lockdown on household water consumption patterns in England. npj Clean Water 4 (1), 13. https://doi.org/10.1038/s41545-021-00103-8.

Akbari, S., Haghighat, F., 2021. Occupancy and occupant activity drivers of energy consumption in residential buildings. Energy Build. 250, 111303.

Al-Wakeel, A., Wu, J., Jenkins, N., 2017. K-means based load estimation of domestic smart meter measurements. Appl. Energy 194, 333–342.

Alhussein, M., Aurangzeb, K., Haider, S.I., 2020. Hybrid CNN-lstm model for short-term individual household load forecasting. IEEE Access 8, 180544–180557. https://doi.org/10.1109/ACCESS.2020.3028281.

Bandyopadhyay, S., Saha, S., 2013. Clustering algorithms. In: Bandyopadhyay, S., Saha, S. (Eds.), Unsupervised Classification: Similarity Measures, Classical and Metaheuristic Approaches, and Applications. Springer Berlin Heidelberg, pp. 75–92. https://doi.org/10.1007/978-3-642-32451-2_4.

Battocchi, Keith, Dillon, Eleanor, Hei, Maggie, Lewis, Greg, Paul, Oka, Oprescu, Miruna, Syrgkanis, V., 2019. EconML: a Python package for ML-based heterogeneous treatment effects estimation (Version 0.x). https://github.com/py-why/EconML.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat. Theor. Methods 3 (1), 1–27.

Chen, G., Zhu, Y., Wiedmann, T., Yao, L., Xu, L., Wang, Y., 2019. Urban-rural disparities of household energy requirements and influence factors in China: classification tree models. Appl. Energy 250, 1321–1335. https://doi.org/10.1016/j.apenergy.2019.04.170.

Chen, P., Wu, Y., Zhong, H., Long, Y., Meng, J., 2022. Exploring household emission patterns and driving factors in Japan using machine learning methods. Appl. Energy 307, 118251. https://doi.org/10.1016/j.apenergy.2021.118251.

Chen, S., Huang, Y., Hu, J., Yang, S., Lin, C., Mao, K., Rao, Z., Chen, Y., 2023. Prediction of urban residential energy consumption intensity in China toward 2060 under regional development scenarios. Sustain. Cities Soc. 99, 104924 https://doi.org/10.1016/j.scs.2023.104924.

Cheng, S., Wei, T., Wang, F., Zhuang, L., 2023. Does financial market participation eradicate household energy poverty? Energy Econ. 122, 106687 https://doi.org/10.1016/j.eneco.2023.106687.

Cheng, S., Wang, K., Meng, F., Liu, G., An, J., 2024. The unanticipated role of fiscal environmental expenditure in accelerating household carbon emissions: evidence from China. Energy Pol. 185, 113962 https://doi.org/10.1016/j.enpol.2023.113962.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. Econom. J. 21 (1), C1–C68. https://doi.org/10.1111/ectj.12097.

Chévez, P., Barbero, D., Martini, I., Discoli, C., 2017. Application of the k-means clustering method for the detection and analysis of areas of homogeneous residential electricity consumption at the Great La Plata region, Buenos Aires, Argentina. Sustain. Cities Soc. 32, 115–129.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. (2), 224–227.

Dou, Y., Zhao, J., Dong, X., Dong, K., 2021. Quantifying the impacts of energy inequality on carbon emissions in China: a household-level analysis. Energy Econ. 102 https://doi.org/10.1016/j.eneco.2021.105502.

Du, K., Shao, S., Yan, Z., 2021. Urban residential energy demand and rebound effect in China: a stochastic energy demand frontier approach. Energy J. 42 (4) https://doi.org/10.5547/01956574.42.4.kdu.

Fan, J.-L., Zhang, Y.-J., Wang, B., 2017. The impact of urbanization on residential energy consumption in China: an aggregated and disaggregated analysis. Renew. Sustain. Energy Rev. 75, 220–233. https://doi.org/10.1016/j.rser.2016.10.066.

Froemelt, A., Wiedmann, T., 2020. A two-stage clustering approach to investigate lifestyle carbon footprints in two Australian cities. Environ. Res. Lett. 15 (10), 104096 https://doi.org/10.1088/1748-9326/abb502.

Froemelt, A., Dürrenmatt, D.J., Hellweg, S., 2018. Using data mining to assess environmental impacts of household consumption behaviors. Environ. Sci. Technol. 52 (15), 8467–8478. https://doi.org/10.1021/acs.est.8b01452.

Gianniou, P., Liu, X., Heller, A., Nielsen, P.S., Rode, C., 2018. Clustering-based analysis for residential district heating data. Energy Convers. Manag. 165, 840–850.

Guilhot, L., 2022. An analysis of China's energy policy from 1981 to 2020: transitioning towards to a diversified and low-carbon energy system. Energy Pol. 162 https://doi.org/10.1016/j.enpol.2022.112806.

Guo, J., Xu, Y., Qu, Y., Wang, Y., Wu, X., 2023. Exploring factors affecting household energy consumption in the internet era: empirical evidence from Chinese households. Energy Pol. 183, 113810.

Han, J., Pei, J., Tong, H., 2022. Data Mining: Concepts and Techniques. Morgan kaufmann.

Hertwich, E.G., Peters, G.P., 2009. Carbon footprint of nations: a global, trade-linked analysis. Environ. Sci. Technol. 43 (16), 6414–6420. https://doi.org/10.1021/es803496a.

Hincks, S., Carter, J., Connelly, A., 2023. A new typology of climate change risk for European cities and regions: principles and applications. Global Environ. Change 83, 102767. https://doi.org/10.1016/j.gloenvcha.2023.102767.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

Hoken, H., Sato, H., 2022. Effects of public transfers on income inequality and poverty in rural China. China World Econ. 30 (5), 29–48. https://doi.org/10.1111/cwe.12436.

Huang, L., Long, Y., Chen, J., Yoshida, Y., 2023. Sustainable lifestyle: urban household carbon footprint accounting and policy implications for lifestyle-based decarbonization. Energy Pol. 181, 113696 https://doi.org/10.1016/j.enpol.2023.113696.

Huebner, G., Shipworth, D., Hamilton, I., Chalabi, Z., Oreszczyn, T., 2016. Understanding electricity consumption: a comparative contribution of building factors, socio-demographics, appliances, behaviours and attitudes. Appl. Energy 177, 692–702.

Ivanova, D., Stadler, K., Steen-Olsen, K., Wood, R., Vita, G., Tukker, A., Hertwich, E.G., 2016. Environmental impact assessment of household consumption. J. Ind. Ecol. 20 (3), 526–536. https://doi.org/10.1111/jiec.12371.

Ji, T., Jia, N., Lin, F., Wu, H., 2019. Fiscal subsidy policy on home appliances: its effects on domestic consumption and exports in China. China World Econ. 27 (4), 53–73. https://doi.org/10.1111/cwe.12286.

Jiang, S., Wei, Z., 2024. Urbanization exacerbated the rapid growth of summer cooling demands in China from 1980 to 2023. Sustain. Cities Soc. 106, 105382 https://doi.org/10.1016/j.scs.2024.105382.

Jiang, L., Shi, X., Wu, S., Ding, B., Chen, Y., 2022. What factors affect household energy consumption in mega-cities? A case study of Guangzhou, China. J. Clean. Prod. 363, 132388 https://doi.org/10.1016/j.jclepro.2022.132388.

Kuai, P., Zhang, X., Zhang, S., Li, J., 2022. Environmental awareness and household energy saving of Chinese residents: unity of knowing and doing or easier said than done? J. Asian Econ. 82, 101534 https://doi.org/10.1016/j.asieco.2022.101534.

Lei, M., Cai, W., Liu, W., Wang, C., 2022. The heterogeneity in energy consumption patterns and home appliance purchasing preferences across urban households in China. Energy 253. https://doi.org/10.1016/j.energy.2022.124079.

Lei, T., Wang, D., Yu, X., Ma, S., Zhao, W., Cui, C., Meng, J., Tao, S., Guan, D., 2023. Global iron and steel plant CO2 emissions and carbon-neutrality pathways. Nature 622 (7983), 514–520. https://doi.org/10.1038/s41586-023-06486-7.

Li, Z., Liu, Y., Zhang, Z., 2022. Carbon emission characteristics and reduction pathways of urban household in China. Front. Environ. Sci. 10 https://doi.org/10.3389/fenvs.2022.896765.

Liu, M., Huang, X., Chen, Z., Zhang, L., Qin, Y., Liu, L., Zhang, S., Zhang, M., Lv, X., Zhang, Y., 2021. The transmission mechanism of household lifestyle to energy consumption from the input-output subsystem perspective: China as an example. Ecol. Indicat. 122 https://doi.org/10.1016/j.ecolind.2020.107234.

Long, Y., Yoshida, Y., Zeng, I.Y., Xue, J., Li, Y., 2021. Fuel-specific carbon footprint embodied in Japanese household lifestyles. Earth's Future 9 (9), e2021EF002213. https://doi.org/10.1029/2021EF002213.

Long, Y., Yoshida, Y., Huang, L., Gasparatos, A., 2022. Carbon footprint differentiation in the Japanese residential sector due to income-driven divergences in consumption and time allocation. Earth's Future 10 (10), e2022EF002954. https://doi.org/10.1029/2022EF002954.

Long, Y., Yoshida, Y., Huang, L., Chen, P., Wu, Y., Gasparatos, A., 2024. Demographic transitions hinder climate change mitigation for Japan's shrinking and aging households. Cell Reports Sustainability 1 (3).

Lu, H., Liu, G., 2014. Spatial effects of carbon dioxide emissions from residential energy consumption: a county-level study using enhanced nocturnal lighting. Appl. Energy 131, 297–306. https://doi.org/10.1016/j.apenergy.2014.06.036.

Luqman, M., Rayner, P.J., Gurney, K.R., 2023. On the impact of urbanisation on CO2 emissions. npj Urban Sustainability 3 (1), 6. https://doi.org/10.1038/s42949-023-00084-2.

Ma, S., Xu, X., Li, C., Zhang, L., Sun, M., 2021. Energy consumption inequality decrease with energy consumption increase: evidence from rural China at micro scale. Energy Pol. 159 https://doi.org/10.1016/j.enpol.2021.112638.

Ma, R., Deng, L., Ji, Q., Zhai, P., 2022. Environmental regulations, clean energy access, and household energy poverty: evidence from China. Technol. Forecast. Soc. Change 182. https://doi.org/10.1016/j.techfore.2022.121862.

Mahdi, M.A., Hosny, K.M., Elhenawy, I., 2021. Scalable clustering algorithms for big data: a review. IEEE Access 9, 80015–80027. https://doi.org/10.1109/ACCESS.2021.3084057.

Maimon, O., Rokach, L., 2005. Data Mining and Knowledge Discovery Handbook, 2. Springer.

Mao, S., Qiu, S., Li, T., Tang, M., Deng, H., Zheng, H., 2020. Using characteristic energy to study rural ethnic minorities' household energy consumption and its impact factors in chongqing, China. Sustainability 12 (17). https://doi.org/10.3390/su12176898.

Maruejols, L., Höschle, L., Yu, X., 2022. Vietnam between economic growth and ethnic divergence: a LASSO examination of income-mediated energy consumption. Energy Econ. 114, 106222.

Mashhadi, A.R., Behdad, S., 2018. Discriminant effects of consumer electronics use-phase attributes on household energy prediction. Energy Pol. 118, 346–355.

McLoughlin, F., Duffy, A., Conlon, M., 2015. A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl. Energy 141, 190–199.

Miao, L., 2017. Examining the impact factors of urban residential energy consumption and CO2 emissions in China – evidence from city-level data. Ecol. Indicat. 73, 29–37. https://doi.org/10.1016/j.ecolind.2016.09.031.

Milligan, G.W., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45, 325–342.

NBS, 2023. China Energy Statistical Yearbook.

Nejat, P., Jomehzadeh, F., Taheri, M.M., Gohari, M., Abd Majid, M.Z., 2015. A global review of energy consumption, CO2 emissions and policy in the residential sector (with an overview of the top ten CO2 emitting countries). Renew. Sustain. Energy Rev. 43, 843–862. https://doi.org/10.1016/j.rser.2014.11.066.

Ofetotse, E.L., Essah, E.A., Yao, R., 2021. Evaluating the determinants of household electricity consumption using cluster analysis. J. Build. Eng. 43, 102487.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Seto, K.C., Dhakal, S., Bigio, A., Blanco, H., Carlo Delgado, G., Dewar, D., Huang, L., Inaba, A., Kansal, A., Lwasa, S., 2014. Human Settlements, Infrastructure, and Spatial Planning.

Shahapure, K.R., Nicholas, C., 2020. Cluster quality analysis using silhouette score. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA).

Shi, X., 2019. Inequality of opportunity in energy consumption in China. Energy Pol. 124, 371–382. https://doi.org/10.1016/j.enpol.2018.09.029.

Shi, X., Wang, K., Cheong, T.S., Zhang, H., 2020a. Prioritizing driving factors of household carbon emissions: an application of the LASSO model with survey data. Energy Econ. 92, 104942.

Shi, X., Yu, J., Cheong, T.S., 2020b. Convergence and distribution dynamics of energy consumption among China's households. Energy Pol. 142 https://doi.org/10.1016/j.enpol.2020.111496.

Shi, X., Cheong, T.S., Yu, J., Liu, X., 2021. Quality of life and relative household energy consumption in China. China World Econ. 29 (5), 127–147. https://doi.org/10.1111/cwe.12390.

Steinley, D., 2003. Local optima in K-means clustering: what you don't know may hurt you. Psychol. Methods 8 (3), 294.

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

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B Stat. Methodol. 58 (1), 267–288.

van de Geer, S., Buhlmann, P., Ritov, Y.A., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Stat. 42 (3), 1166–1202. https://doi.org/10.1214/14-AOS1221.

Wang, M., Wei, C., 2024. Toward sustainable heating: assessment of the carbon mitigation potential from residential heating in northern rural China. Energy Pol. 190, 114141.

World Bank, United Nations Development Programme, & Global Infrastructure Facility, 2020. Catalyzing Private Sector Investment in Climate Smart Cities.

Wu, S., 2019. Methods for estimating residential energy consumption. In: Zheng, X., Wei, C. (Eds.), Household Energy Consumption in China: 2016 Report. Springer, Singapore, pp. 95–111. https://doi.org/10.1007/978-981-13-7523-1_3.

Wu, S., Zheng, X., Wei, C., 2017. Measurement of inequality using household energy consumption data in rural China. Nat. Energy 2 (10), 795–803. https://doi.org/10.1038/s41560-017-0003-1.

Wu, S., Wang, D., Hu, J., Wei, C., 2019a. Research background and main conclusions. In: Zheng, X., Wei, C. (Eds.), Household Energy Consumption in China: 2016 Report. Springer, Singapore, pp. 1–37. https://doi.org/10.1007/978-981-13-7523-1_1.

Wu, S., Zheng, X., You, C., Wei, C., 2019b. Household energy consumption in rural China: historical development, present pattern and policy implication. J. Clean. Prod. 211, 981–991. https://doi.org/10.1016/j.jclepro.2018.11.265.

Wu, D., Geng, Y., Zhang, Y., Wei, W., 2022. Features and drivers of China's urban-rural household electricity consumption: evidence from residential survey. J. Clean. Prod. 365, 132837 https://doi.org/10.1016/j.jclepro.2022.132837.

Xu, D., Tian, Y., 2015. A comprehensive survey of clustering algorithms. Annals of Data Science 2 (2), 165–193. https://doi.org/10.1007/s40745-015-0040-1.

Xu, R., Tong, D., Davis, S.J., Qin, X., Cheng, J., Shi, Q., Liu, Y., Chen, C., Yan, L., Yan, X., Wang, H., Zheng, D., He, K., Zhang, Q., 2023. Plant-by-plant decarbonization strategies for the global steel industry. Nat. Clim. Change 13 (10), 1067–1074. https://doi.org/10.1038/s41558-023-01808-z.

Yang, T., Ren, M., Zhou, K., 2018. Identifying household electricity consumption patterns: a case study of Kunshan, China. Renew. Sustain. Energy Rev. 91, 861–868. https://doi.org/10.1016/j.rser.2018.04.037.

Yu, B., Wei, Y.-M., Gomi, K., Matsuoka, Y., 2018. Future scenarios for energy consumption and carbon emissions due to demographic transitions in Chinese households. Nat. Energy 3 (2), 109–118.

Zhang, C.-H., Zhang, S.S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. J. Roy. Stat. Soc. B Stat. Methodol. 76 (1), 217–242. https://doi.org/10.1111/rssb.12026.

Zhang, N., Zhou, M., 2020. The inequality of city-level energy efficiency for China. J. Environ. Manag. 255, 109843 https://doi.org/10.1016/j.jenvman.2019.109843.

Zhang, T., Shi, X., Zhang, D., Xiao, J., 2019. Socio-economic development and electricity access in developing economies: a long-run model averaging approach. Energy Pol. 132, 223–231.

Zhang, X., Xu, K., He, M., Wang, J., 2022. A review on the rural household energy in China from 1990s—transition, regional heterogeneity, emissions, energy-saving, and policy. Front. Energy Res. 10 https://doi.org/10.3389/fenrg.2022.907803.

Zheng, X., 2020. Analysis of the policy effect of "automobile go to countryside" and relative suggestions. In: 2020 2nd International Conference on Economic Management and Cultural Industry (ICEMCI 2020).

Zheng, X., Wei, C., 2019. Household Energy Consumption in China: 2016 Report. Springer.

Zheng, X., Wei, C., Qin, P., Guo, J., Yu, Y., Song, F., Chen, Z., 2014. Characteristics of residential energy consumption in China: findings from a household survey. Energy Pol. 75, 126–135. https://doi.org/10.1016/j.enpol.2014.07.016.

Zhou, K., Yang, S., 2016. Understanding household energy consumption behavior: the contribution of energy big data analytics. Renew. Sustain. Energy Rev. 56, 810–819. https://doi.org/10.1016/j.rser.2015.12.001.

Zhou, K., Yang, S., Shao, Z., 2017. Household monthly electricity consumption pattern mining: a fuzzy clustering-based model and a case study. J. Clean. Prod. 141, 900–908. https://doi.org/10.1016/j.jclepro.2016.09.165.

Zou, B., Luo, B., 2019. Rural household energy consumption characteristics and determinants in China. Energy 182, 814–823. https://doi.org/10.1016/j.energy.2019.06.048.