Applied Energy 226 (2018) 240–251

Contents lists available at ScienceDirect

Applied Energy

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

Distribution patterns of energy consumed in classified public buildings through the data mining process

Yibo Chen[a,b,*], Jianzhong Wu[c]

[a] Department of Civil Engineering, Zhengzhou University, Zhengzhou 450000, China
[b] School of Mechanical Engineering, Tongji University, Shanghai 200092, China
[c] School of Engineering, Cardiff University, Cardiff, Wales CF24 3AA, UK

HIGHLIGHTS

- Regional energy distribution patterns were explored based on the data mining process.
- The pre-preprocessing was conducted stepwise aiming at missing data and abnormal data.
- The Lorenz curve was introduced to quantify energy distribution inequality.
- Various empirical formulae addressing regional distribution patterns were provided.
- Application potential based on the identified principles was discussed for planning stage.

ARTICLE INFO

Keywords:
Building energy consumption
Data mining
Lorenz curve
Pre-processing
Regional distribution patterns

ABSTRACT

Reliable spatio-temporal distribution analysis of building energy consumption is a crucial basis of bottom-up regional energy models, especially when faced with uncertain information at the planning stage. In existing statistical models, the regional consumption levels were mostly identified based on a large amount of samples with multi-dimensional parameters, which are usually not available for developing countries. Pointing to this, the distribution features of regional energy consumption are explored in this paper based on the whole procedure of data mining, which consists of three parts namely pre-processing, information mining, and validation & application. In this process, 212 samples of classified public buildings in Beijing and 66 samples in Hangzhou are included. Firstly, the pre-processing is conducted stepwise aiming at processing the missing data and the abnormal data. Afterwards, the interdisciplinary Lorenz curve is introduced to transfer the scatters into regular curves with satisfied fitting goodness. Thus, empirical formulae are extracted to quantify the nonlinear distribution principles of individual EUIs along with the accumulative building area. Finally, the achieved empirical formulae of different building types are validated, and the application potential of the identified patterns is discussed aiming at the planning stage. Through data mining of the limited datasets, this paper attempts to identify the hidden distribution patterns of regional energy consumption, which enables the regional modeling.

1. Introduction

As a major contributor to CO_2 emissions, the building sector accounts for around 28% of the total energy consumption in China and even higher proportions in many other regions [1]. It has been estimated that the building energy consumption in China has an annual increasing rate of 7% since 2001 [2]. Due to the pressures from global warming and resource shortage, the focus of energy strategies in China has transformed gradually from supply-side planning into demand-side management [3]. Consequently, the simultaneous control of both energy efficiency and energy quantity is proposed, in order to realize the targets of CO_2 emission peaks and the occupied proportion of non-fossil fuels, as requested by the U.S. – China Joint Announcement on Climate Change [4]. Under such circumstances, the district energy system is proposed as an integrated supplier of power, heating and cooling for multi-consumers, and it can achieve the gradient utilization of different energy grades. Regional energy is believed to be more applicable for districts with higher load intensity and longer operation hours [5]. At the same time, statistical analysis of spatio-temporal features of regional energy consumption is a crucial basis of reliable regional planning strategies [6]. Otherwise, the energy conservation assessed from subtly-designed energy systems could be overridden by the traditional

* Corresponding author at: Department of Civil Engineering, Zhengzhou University, Zhengzhou 450000, China.
E-mail address: yb_chen77@163.com (Y. Chen).

https://doi.org/10.1016/j.apenergy.2018.05.123
Received 31 December 2017; Received in revised form 28 May 2018; Accepted 31 May 2018

Nomenclature

EUI	normalized electrical energy usage intensity per square meter, kWh/m^2
annual EUI	energy consumed by a unit of building area within a year, kWh/(m^2·a)
monthly EUI	energy consumed by a unit of building area within a month, kWh/(m^2·mon)
Ea_{win}	average EUI of winter season, kWh/m^2
R^o	relative range,%
X_{ab}	abnormal data point
Q3, Q1	upper quartile and lower quartile of samples
IQR	inter-quartile range of samples
β	included angle between horizontal line and the tangent of Lorenz curve
$L(p)$	fitting Lorenz curve function of income distribution
$L(A)$	fitting Lorenz curve function of energy distribution
Gini	Gini index of the Lorenz curve
C_i	fitted coefficients of Lorenz curves
s	cumulative building area, m^2
q	cumulative energy consumption, kWh
St	total building area of samples, m^2
Qt	total building energy consumption, kWh
K	statistical average annual EUI of all samples, kWh/(m^2·a)
EUI_i	a variable of annual EUI changing along with building area, kWh/(m^2·a)

extensive estimation of regional energy consumption.

The bottom-up regional energy consumption is the superposition of energy consumed in individual buildings within a region [7], and it can be obtained via physical simulation, statistical analysis and hybrid models [8]. However, influenced by various internal and external factors, the energy usage intensity (EUI) of individual buildings within a building type is diverse with a range of 30–300 kWh/(m^2·a) in China [2]. In other words, the practical annual EUIs of individual buildings are usually unequally distributed along with the building area. Unlike the traditional descriptive distribution analysis such as frequency analysis, this kind of distribution pattern is hidden behind the changing principles of EUIs along with the building area within a region. This kind of hidden distribution pattern is crucial for the bottom-up regional modeling, in which the regional energy consumption is accumulated by multiplying individual EUIs by corresponding building areas [9,10]. In traditional static planning or linear programming based on refined building classifications, the city-scale building energy consumption is estimated by linear extrapolation through building area, static indicators and coincidence factors [11,12]. Kontokosta [13,14] concluded that the ignorance of such distribution inequality and its following impacts largely hinders the energy prediction and the estimation of energy saving potential. Consequently, the hidden consumption patterns of individual EUIs among different buildings should be identified firstly, in order to facilitate the nonlinear bottom-up modeling instead of the traditional static indicators and empirical coincidence factors.

Currently, besides the traditional distribution tests [15], there are already some studies about the identification of regional distribution inequality among individual EUIs.

One solution is to use the hybrid physical model. Specifically, the distribution characteristics of samples are firstly quantified, serving as the input or calibration scales of prototype physical building models [16]. Among all these, the Bayesian algorithm and Monte Carlo simulation are the most popular methods used. For instance, Yu et al. [7] carried out a two-step city-wide estimation by making use of the Bayesian analysis, in order to further correct the priori engineering estimates. Here the priori beliefs were extracted from initial engineering typical buildings, and were later transformed into the posteriori distributions of regional energy consumption patterns. Similarly, Monte Carlo simulation can be utilized to establish various physical models by considering the uncertain distribution features of physical parameters, thus the individual EUIs among different buildings can be distinguished from the simulated results [17,18]. This kind of combination can obtain dynamic results based on different groups of selected parameters, yet it takes large quantities of calculation time. Moreover, the identified distribution is still limited to the assumption of typical distribution patterns of physical parameters. At the same time, this procedure can be even more difficult when faced with the uncertain information during planning and design stages [19].

Along with the growing energy datasets and the development of data mining techniques, the statistical data-driven solution has attracted increasing attentions, and it is especially necessary for the planning stage with incomplete information [20]. Data mining (DM) has shown great potential, and it is defined as a process of discovering hidden principles from ambiguous datasets at the intersection of database systems, machine learning, pattern recognition and traditional statistics etc. [21,22].

Actually, the DM process has been applied in various interdisciplinary areas including the identification of energy consumption patterns. Relative studies mainly focus on the recognition of load shapes and consumption patterns of different consumers, based on various machine learning methods. For example, based on the combination of liner regression, random forest and support vector regression, Kontokosta and Tull [23] developed a data-driven model to identify the disparity of energy consumption for 23,000 buildings. The datasets of this study include property and zoning information of each building sample. Similarly, Howard et al. [24] attributed the extracted end-use energy consumption to each land block via zoning information, yet no reliable validations were conducted for the attributed EUIs of individual buildings. Besides, based on the smart metering data of over 2000 commercial buildings in California, clustering and frequency distribution were adopted by Luo et al. [25] to identify the shapes of electric loads for different building types. The derived load shapes were later inserted in a public tool called Commercial Building Energy Saver. Through this kind of method, the distribution inequality of energy consumption among different buildings can be assessed directly from the datasets, yet there are no quantified relationships between EUIs and the related parameters. As a specific evaluation tool of distribution inequality, the interdisciplinary Lorenz curve was introduced to assess the inequality of energy consumption among different buildings [26]. The Lorenz curve has been proved to be an efficient overall comparative metric, yet its traditional simplex form and functions are limited for the regional modeling.

It can be concluded that the data-driven statistical method is useful when faced with the practical datasets. However, it should be noticed that the above models were mostly established based on complete datasets with multi-dimensional parameters in developed countries. However, issues of confidentiality and unwillingness to share data in developing countries have seriously hindered the progress of data mining applications in this field. On the other hand, the uneven developing levels of different cities in China directly lead to the large regional disparity of building energy distribution, the identification of which is indispensable for the proactive decision-making at the planning stage with uncertain information. Actually, reliable regional-scale empirical models driven by actual datasets are still absent in China.

A data-driven approach with limited datasets in China was developed in this paper based on the whole process of data-mining. The remaining parts are organized as follows. Section 2 firstly introduces available datasets and the proposed methodology. Section 3 presents the results of pre-processing, followed by the results of Lorenz curve