Applied Energy 140 (2015) 85–93

Contents lists available at ScienceDirect

Applied Energy

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

Big-data for building energy performance: Lessons from assembling a very large national database of building energy use

Paul A. Mathew *, Laurel N. Dunn, Michael D. Sohn, Andrea Mercado, Claudine Custudio, Travis Walter

Environmental Energy Technologies Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Mail Stop: 90R2000, Berkeley, CA 94720, USA

HIGHLIGHTS

- The largest compilation of building energy data in the US, with over 750,000 buildings.
- Most of the effort lies in data cleansing and mapping to a common data schema.
- Paper includes comparisons to data in CBECS and RECS – the US national statistical datasets.
- The database supports empirical comparison of energy use and data-driven savings analysis.

ARTICLE INFO

Article history:
Received 14 March 2014
Received in revised form 19 November 2014
Accepted 20 November 2014

Keywords:
Buildings Performance Database
Building performance
Big data
Building data collection
Data-driven decision support

ABSTRACT

Building energy data has been used for decades to understand energy flows in buildings and plan for future energy demand. Recent market, technology and policy drivers have resulted in widespread data collection by stakeholders across the buildings industry. Consolidation of independently collected and maintained datasets presents a cost-effective opportunity to build a database of unprecedented size. Applications of the data include peer group analysis to evaluate building performance, and data-driven algorithms that use empirical data to estimate energy savings associated with building retrofits. This paper discusses technical considerations in compiling such a database using the DOE Buildings Performance Database (BPD) as a case study. We gathered data on over 750,000 residential and commercial buildings. We describe the process and challenges of mapping and cleansing data from disparate sources. We analyze the distributions of buildings in the BPD relative to the Commercial Building Energy Consumption Survey (CBECS) and Residential Energy Consumption Survey (RECS), evaluating peer groups of buildings that are well or poorly represented, and discussing how differences in the distributions of the three datasets impact use-cases of the data. Finally, we discuss the usefulness and limitations of the current dataset and the outlook for increasing its size and applications.

1. Introduction

Energy efficiency is a cost-effective resource for curbing energy use and carbon emissions from buildings. Engineering-based studies forecast large energy and economic savings potential over time from modest investments in efficiency across the building stock [1–3]. One study by the Rocky Mountain Institute estimates that a $0.5 trillion investment in efficiency across the buildings sector could return $1.4 trillion in savings by 2050 [4]. Other studies find that engineering-based analyses may overestimate potential energy savings [5], and more generally inaccurately predict energy use in real buildings [6,7]. Discrepancies between modeled and measured energy use and savings have been attributed to difficulties in accounting for occupant behavior [8], interactive effects between building systems [9], uncertainty in model inputs [10], and inefficiencies in operational buildings due to improper maintenance and operation of building systems [11,12].

A historic lack of empirical energy data has limited our ability to validate engineering-based predictions of energy savings potential in buildings. However, a recent surge in the number of buildings benchmarking energy use [13] has increased the amount of available building energy data.

Abbreviations: CBECS, Commercial Buildings Energy Consumption Survey; RECS, Residential Energy Consumption Survey; CEUS, Commercial End-Use Survey; DOE, Department of Energy; BPD, Buildings Performance Database; USEIA, United States Energy Information Administration; EUI, Energy Use Intensity; BEDES, Building Energy Data Exchange Specification; ESCO, Energy Service Company; NBI, New Buildings Institute; USEERE, United States Office of Energy Efficiency & Renewable Energy.

* Corresponding author. Tel.: +1 510 496 5116.
E-mail address: pamathew@lbl.gov (P.A. Mathew).

http://dx.doi.org/10.1016/j.apenergy.2014.11.042
0306-2619/

Empirical data analysis using large-scale data sets has been transformational in fields such as crime-fighting [14], political campaigns [15], and commerce [16]. Large-scale empirical building energy data may prove beneficial to stakeholders throughout the industry including policymakers, building owners, and investors in energy efficiency. Several technology, market, and policy drivers, such as smart meters and energy disclosure laws, have led to unprecedented data collection throughout the buildings sector, which has spurred several efforts to bring data-driven decision-making to stakeholders in building performance.

Data-driven algorithms offer low-cost alternatives to energy models for predicting energy savings and estimating financial return on energy efficiency investments [9,17]. A building energy database could also improve analyses currently driven by small or outdated datasets, such as informing energy efficiency policy, or planning for future energy demand [18–21].

The DOE funded Buildings Performance Database [42] seeks to fill the identified need for "big data" in the buildings sector. In this paper, we discuss our amassing of energy use data from nearly 750,000 commercial and residential buildings aggregated from smaller datasets collected by organizations such as cities, utilities, energy efficiency programs and building portfolio owners. The paper addresses technical considerations in generating a large-scale database for building performance analysis. We first evaluate the need for a building energy database, discussing existing databases, their applications and shortcomings, and opportunities for analysis afforded by a larger more comprehensive database (Section 2). We then discuss the process of compiling the BPD, including data outreach, aggregation, and quality assurance (Section 3). We then assess the quantity and depth of data contained in the BPD (Section 4: "how big is the data?"), comparing the BPD to the national building stock, and discuss how the distribution of buildings in the database either helps or hinders data analysis prospects (Section 4: "how useful is the data?"). We conclude by reflecting on the current state of the BPD, considering its effectiveness as a decision-support tool and identifying opportunities to improve the quality and depth of building data analysis (Section 5).

2. Background: The need for a comprehensive database of building energy

2.1. The current state of empirical building data

Empirical building data holds widespread potential in buildings management, energy efficiency, policy assessment, and energy planning. This section discusses existing building energy databases and their applications. We highlight data collection methods and salient characteristics of each dataset, and how these impact use-cases for the data. Based on our review of other databases, we identify the need for a comprehensive database to consolidate data from throughout the industry in order to reduce data collection costs, create new opportunities for analyzing building data, and reach a broader audience within the building sector.

Databases including CBECS and RECS contain in-depth energy use and asset data for representative samples of the national commercial and residential building stocks [22,23]. These datasets are collected for energy planning and forecasting purposes, but also provide summary statistics of the national buildings stock (NBI, 2014) [24–26]. The EIA's Annual Energy Outlook relies heavily on CBECS and RECS to evaluate energy use trends in the buildings sector [40]. A similar database compiled by the California Energy Commission, CEUS, is analyzed to understand energy use by the California commercial buildings stock [27,39]. Both CBECS and RECS are extremely costly to collect, resulting in relatively small sample sizes and infrequent data updates [19]. The BPD was developed, in part, to explore low-cost data collection methods in response to an industry need for bigger and more up-to-date data than RECS and CBECS can provide. Additionally, the sampling of buildings within CBECS and RECS was structured, in part, to gain a national-scale representative view of the building stock. While significant, specifically for national-scale energy analyses, such databases may not provide fine detail or resolution at regional spatial scales.

Other databases target certain subsets of the building stock or specific use-cases. These datasets include collections by Labs 21, ENERGY STAR Portfolio Manager, and the New Buildings Institute (NBI), among others. Labs 21 collects benchmarking data for laboratories across the country, focusing on laboratory-specific energy drivers [28]. Portfolio Manager collects energy benchmarking data and assigns EPA ENERGY STAR Scores for several building types. Both Labs 21 and Portfolio Manager collect data submitted by users online, resulting in low data collection costs. The NBI collects energy use and design data for LEED certified buildings (NBI, 2014). The database has been used to compare performance of LEED certified buildings to the national building stock and to evaluate their performance relative to design stage simulations conducted as part of the LEED certification process [6,7]. Numerous other databases collect building data for applications unrelated to energy performance. For example, CoStar and Zillow are private companies that collect data on U.S. real-estate markets for commercial and residential buildings, respectively, monitoring market prices based on building size, characteristics, and location. The BPD draws on performance-related data collected throughout the industry including but not limited to data collected for other databases; these diverse datasets are then aggregated into one database. The successful use of regional-scale, or market-specific, databases implies the need for a database that can provide an overview at multiple scales of the building stock. Even an incomplete national database, like the BPD in its current form, is nonetheless useful for various local-scale energy analyses. In other words, we do not have to wait for the BPD to be "complete" before important energy analysis can be explored.

One anticipated use of BPD data is to power new data-driven algorithms for estimating the energy savings associated with building retrofits, augmenting modeling-based energy savings predictions. One study comparing design-stage energy simulations to measured energy use in operational buildings using the NBI's database of LEED Certified buildings, found that actual energy use deviated from simulated energy use by 25% or more in over half the buildings in the database [6,7]. Using empirical data to compute energy savings rather than engineering-based estimates may account for factors such as occupant behavior, operational inefficiencies and interactive effects that are difficult or costly to account for in building energy models.

One benefit to data-driven approaches to energy savings prediction is that results are given as probabilistic distributions of energy savings. Understanding uncertainty in energy savings is becoming increasingly important with the rise of Energy Service Companies (ESCOs) [12,29], who finance investments in energy efficiency using utility bill energy savings. Calculating the probability of achieving a particular level of energy savings may boost investor confidence in energy efficiency by quantifying uncertainty in estimated return on investment based on empirical data. Understanding uncertainty in energy savings is key to evaluating investment risk, which has thus far been largely limited to simulated energy savings analysis [17].

2.2. Intended use cases for the BPD

The BPD is intended to be a broad data collection effort to support a range of different analysis use cases. Table 1 provides a high