LBNL-

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

# 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

November, 2014

# Big-Data for Building Energy Performance: Lessons from Assembling a Very Large National Database of Building Energy Use

Paul A. Mathew[1] , 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

Revised November 18, 2014

[1] Corresponding author, pamathew@lbl.gov ; + 1 510 496 5116

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

Acknowledgements

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, [Building Technologies Program] or [Federal Energy Management Program], of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

**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 700,000 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.

*Keywords: Buildings Performance Database; building performance; big data; building data collection; data-driven decision support*

## 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

## 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 (Williams et al, 2012; McKinsey & Co., 2009; Pacala and Socolow, 2004). 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 (Lovins, 2011). Other studies find that engineering-based analyses may overestimate potential energy savings (Alcott and Greenstone, 2012), and more generally inaccurately predict energy use in real buildings (NBI, 2008). Discrepancies between modeled and measured energy use and savings have been attributed to difficulties in accounting for occupant behavior (Ryan and Sanquist, 2012), interactive effects between building systems (Chidiac et al, 2011), uncertainty in model inputs (Eisenhower et al, 2012), and inefficiencies in operational buildings due to improper maintenance and operation of building systems (O'Neill et al, 2011; Mills, 2009).

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 (ENERGY STAR, 2012) has increased the amount of available building energy data.

Empirical data analysis using large-scale data sets has been transformational in fields such as crime-fighting (U.S. Depts. of Transportation & Justice, 2009), political

campaigns (Issenberg, 2012), and commerce (Bryant et al, 2008). 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 (Deng et al, 2014; Chidiac et al, 2011). 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 (Nicholls, 2014; NRCCNS, 2012; Gold and Elliot, 2010; Pérez-Lombard et al, 2008).

The DOE funded Buildings Performance Database 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 (USEIA, 2003; USEIA, 2009). These datasets are collected for energy planning and forecasting purposes, but also provide summary statistics of the national

115 buildings stock (NBI, 2014; USEERE, 2011; ENERGY STAR, 2013). The EIA's Annual
116 Energy Outlook relies heavily on CBECS and RECS to evaluate energy use trends in
117 the buildings sector (USEIA, 2014). A similar database compiled by the California
118 Energy Commission, CEUS, is analyzed to understand energy use by the California
119 commercial buildings stock (Itron, Inc., 2006). Both CBECS and RECS are extremely
120 costly to collect, resulting in relatively small sample sizes and infrequent data
121 updates (NRCCNS, 2012). The BPD was developed, in part, to explore low-cost data
122 collection methods in response to an industry need for bigger and more up-to-date
123 data than RECS and CBECS can provide. Additionally, the sampling of buildings
124 within CBECS and RECS was structured, in part, to gain a national-scale
125 representative view of the building stock. While significant, specifically for national-
126 scale energy analyses, such databases may not provide fine detail or resolution at
127 regional spatial scales.

128 Other databases target certain subsets of the building stock or specific use-cases.
129 These datasets include collections by Labs 21, ENERGY STAR Portfolio Manager, and
130 the New Buildings Institute (NBI), among others. Labs 21 collects benchmarking
131 data for laboratories across the country, focusing on laboratory-specific energy
132 drivers (Mathew et al 2010). Portfolio Manager collects energy benchmarking data
133 and assigns EPA ENERGY STAR Scores for several building types. Both Labs 21 and
134 Portfolio Manager collect data submitted by users online, resulting in low data
135 collection costs. The NBI collects energy use and design data for LEED certified
136 buildings (NBI, 2014). The database has been used to compare performance of LEED
137 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 (NBI, 2008). 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 (NBI, 2008). 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) (Satchwell et al., 2010; Mills, 2009), 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 (Deng et al., 2014).

### 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 level summary of the intended use cases of the BPD for different stakeholders.

*Table 1: Summary of use cases for the BPD.*

One of the ongoing challenges with the BPD is reconciling the scope of the use cases with the data availability and data collection effort. Each use case presents its own data collection requirements and priorities, as indicated in the examples below:

- Simple peer group benchmarking based on whole building energy use intensity (energy use per unit area) to screen and prioritize buildings for overall efficiency potential: For most building types, this can be done

reasonably effectively with whole building annual energy use, building size, climate zone and optionally two to three additional characteristics such as occupancy schedule.

- Comparison of energy efficiency scores for different building types and geographic regions: This has been of particular interest in cities and states with energy disclosure laws, e.g. How does the distribution of energy efficiency scores for office buildings in New York City compare to those in San Francisco? This type of analysis also only requires whole building annual energy use data and building characteristics.
- Portfolio-level analysis of the impacts of energy technologies: For example, is there a statistically observable "shift" in the distribution of energy use intensities for buildings with variable air volume systems vs. constant volume systems? This type of analysis will require data on building system characteristics in addition to whole building energy use and characteristics.
- Energy savings from specific retrofit measures: This type of analysis will require pre- and post-retrofit energy use data as well as data on the type of retrofit and related building system characteristics. These types of data are much more difficult to acquire in a consistent format for large numbers of buildings.

## 3. Data Acquisition, Mapping and Cleansing

Data collectors throughout the buildings industry voluntarily submit data for inclusion in the BPD. Widespread data collection is a relatively recent phenomenon in the buildings sector, which means there are no widely used standards for formatting data or quality control. A critical research effort is how to bring together these disparate sources of data and how to develop an architecture that facilitates the aggregation, and mapping of the data to ensure that incomplete, erroneous or otherwise suspect data does not compromise integrity in database entries or analysis results. The following sections discuss considerations in collecting, mapping and cleansing building energy data for the BPD.

### 3.1. Data Acquisition

The BPD contains data for over 750,000 buildings from nearly 30 data sources, listed in Table 2. Source data sets range in size from 10 to 650,000 buildings, and vary substantially in the level of detail provided for each building. All datasets acquired by the BPD are mapped to a common data format to facilitate import into the database, a process detailed in section 3.2. As an incentive to submit data to the BPD, mapped and cleansed data is returned to each data contributor, along with a statistical overview of the dataset.

In order to develop useful decision-support tools for analyzing building performance, a database must contain sufficient data to conduct robust statistical analysis. As noted earlier, the criteria for data sufficiency will vary based on the intended analysis. Indeed, “big data” does not in and of itself guarantee more

insightful analysis, as documented even in the mainstream media (Ogas 2013, Marcus and Davis 2014) . In general, however, the quality of analysis is expected to improve as the database increases in size, as it will allow better assessment of uncertainty and variability, and inform the analysis of data sufficiency for various use cases. For this reason, ongoing acquisition of new data is key to the success of the BPD. Three general categories of data sources are targeted for outreach: existing databases, entities monitoring building performance, and building portfolio owners or managers.

*Table 2: Data contributors by sector as of February 2014*

Existing databases in the BPD include CBECS, RECS and CEUS. Combined, these datasets account for over 15,000 buildings, or about 2% of the database. As discussed previously, these datasets are sampled so as to statistically represent the underlying distribution of buildings in the U.S. commercial, U.S. residential, and California commercial building stocks, respectively. Data is gathered using surveys administered by the U.S. EIA and the California Energy Commission; the surveys collect high-level details about building assets and operational characteristics for every building. All three datasets are publicly available, heavily analyzed, and maintain very high data quality standards. For CBECS and RECS, the BPD includes postal code and monthly energy use data that is not publicly available. In cases where complete energy use data was unavailable from surveyed buildings, CBECS and RECS use statistical methods to extrapolate energy use. In contrast, a key dictum of the BPD is to restrict the database to empirical data.  This decision, in part,

reduces the potential conflicts with interpreting energy records in BPD. We thus decided to exclude buildings from CBECS and RECS with extrapolated energy use data.

Data contributors monitoring performance of specific portfolios of buildings include cities, public utilities and energy efficiency programs. The interests motivating these parties to collect data are diverse, resulting in high variation in the depth and quality of the data they provide. Cities, for example, collect primarily benchmarking data from local buildings that they can use to inform energy efficiency policies. One study evaluating the level of detail needed to analyze a building stock for this purpose found that collecting highly detailed data from energy audits added little value to models for predicting energy use in the New York building stock, while New York's requirement that certain buildings undergo energy audits substantially increased data collection costs (Hsu, 2014). Energy efficiency programs, on the other hand, often conduct energy audits to identify opportunities for reducing energy use. These data sources typically include either the results of an energy audit in the submitted data, or other details about energy efficiency measures taken in each building. The BPD does not refuse data that is missing asset and equipment data, however, buildings missing key data fields that describe location, size, building type and energy consumption are excluded from the database.

Other data sources include property managers and entities that own and operate portfolios of buildings such as school districts, local governments, federal agencies, college campuses, and retail chain stores. These sources are likely to monitor

complete energy consumption and may provide some equipment data, but rarely with the level of detail present in CBECS, RECS and CEUS.

#### 3.1.1. Data Privacy

Options for preserving data privacy in public databases containing sensitive information are well established. To preserve the BPD's status as a repository for real building data, we chose not to employ techniques in which the actual data is modified, such as data swapping (Dalenius and Reiss, 1982) and randomization (Kargupta et al., 2005). Instead, the BPD shows only aggregated data to users, and suppresses energy use data for peer groups of fewer than ten buildings. These techniques minimize the likelihood that users will be able to single out consumption data for any building in particular.

### 3.2. Data Mapping

The BPD stores data in the BEDES format, discussed in detail below. BEDES provides a common language for storing data with clear guidelines regarding fields, data types, and permitted values. Translating source data to BEDES facilitates aggregation into the database and, if adopted by data collectors throughout the industry, may simplify data sharing among data collectors. Source data often contains fields and data types that loosely equate to those specified in BEDES, although some degree of interpretation is usually required to translate differences in formatting and naming conventions. Some data contributors, however, maintain equipment data primarily for internal use by facilities managers and this data typically requires extensive mapping to be translated into the BEDES specification.

In many cases, data that is not explicitly included in a source dataset can be extrapolated using either the data provided, or outside knowledge about the data contributor. In one example, EPA ENERGY STAR provided data for buildings that have achieved the ENERGY STAR Label. The data did not specify that each building achieved the certification, but knowledge about the dataset allowed us to extrapolate information not explicitly stated in the data. Although many aspects of the mapping process can be automated, these types of situations require that mapping involve a fair degree of human interaction with the data.

#### 3.2.1. Data Specification

Energy-related data collection in the buildings industry is a relatively recent phenomenon, and a uniform format for collecting data has yet to be established. ENERGY STAR Portfolio Manager, a benchmarking tool commonly used throughout the industry, allows users to download data in their standard format. The data contained in Portfolio Manager is collected primarily for benchmarking purposes and to calculate ENERGY STAR Ratings (ENERGY STAR, 2014), neither of which require collection of detailed asset data. BEDES recently emerged as a standard format for storing comprehensive data regarding building assets, characteristics, and use patterns. Developed in conjunction with the BPD, BEDES includes over 600 fields, and accommodates information about hundreds of factors that influence building energy consumption. BEDES is designed to preserve as much detail as is provided by data contributors. Wide deployment of BEDES is expected to facilitate

collection, exchange, and aggregation of high-level building characteristic data throughout the industry.

BEDES fields are subdivided into several categories including site, residential facility, commercial facility, building systems, energy efficiency measures, and energy use. The relationships between these categories are detailed in Figure 1. The "one to many" relationship indicates that one building entry may contain more than one value for a particular field. For example, a single building may contain multiple types of lighting, but can only be in one location. Therefore the Site table contains only one entry for each building, while the Lighting table may contain many.

Site fields store location data such as postal code, climate zone, and elevation; these fields apply to all buildings. The residential and commercial facility fields describe facility-level characteristics, such as floor area and vintage, as well as building and operational characteristics specific to residential or commercial buildings. For example, residential facility fields record the number of residents, ownership status, or education level of residents, among other fields relevant to residential but not commercial buildings. The Measures fields collect data about energy efficiency measures, retrofits, and other changes to building systems or components that may account for changes in energy use over time. Activity Area fields store data about the different activities that occur within a mixed-use commercial building, such as the floor area occupied and operational characteristics specific to each activity type. These fields allow us to identify a dominant facility type, but also enable analysis of building performance using more detailed information about activities within a building. For example, a building that is 90%

offices and 10% data center will be classified as primarily an office building, but may use more energy than a similar building occupied by 100% offices.

*Figure 1: Building Energy Data Exchange Specification schema of data fields.*

BEDES is designed to collect detailed information about building systems and components such as lighting, HVAC, and envelope. System information accounts for the majority of fields in BEDES, including system type, quantity, fuel, efficiency and other information for 23 different building systems and components. Fields relating to energy use can accommodate annual, monthly, or interval consumption data for various fuel streams. These fields include data such as the fuel type, units, metering configuration, rate structure, and emissions factors, as well as the time, duration, reading, and peak energy use for each interval.

### 3.3. Data Cleansing

Cleansing ensures the integrity of the database, and is intended to remove incomplete, erroneous, or otherwise suspect data. We decided that the cleansing process must involve a series of checks to verify that data conforms to a range, list, or equation-describing values permitted in every BEDES field. Several examples of permitted values and checks are described in Table 3. Checks may be as simple as comparing a given elevation against the minimum and maximum elevations in a region or as complex as comparing energy use intensities against distributions of similar buildings in the database to identify outliers and otherwise suspect data.

*Table 3: Examples of BPD cleansing rules by data field including data types, permitted values (out-of-range checks), and in-range checks for each field.*

In-range and out-of-range checks are employed to confirm that values in the data are within reasonable or researched limits. In many cases, these checks are examples where knowledge of building energy is applied to improve the quality of data in the BPD; however, engineering-based judgments are avoided wherever possible. Out-of-range checks compare data entries against a range of permitted values. For example, only California and Louisiana contain elevations below sea level, which means negative elevations are only permissible if a building is in one of those two states. In-range checks confirm that values are not unrealistically high or low. For example, electricity readings less than zero are deleted during cleansing unless the building also generates electricity on site. Ranges and equations for in- and out-of-range checks are determined not only by researching expected values, but also by using data in other fields to identify inconsistencies within a building entry. For example, the heated floor area of a building cannot exceed its gross floor area.

Other more manual checks involve analyzing the distribution of buildings by energy use to identify unlikely values or distributions. In one example, shown in Figure 2, an unlikely peak in the number of buildings with roughly 30 kBtu/ft$^2$–year prompted an inquiry, which revealed that energy use for many buildings in the dataset had been estimated rather than measured. These buildings were removed

during cleansing because the BPD includes only buildings with measured energy use.

In most cases when data integrity issues are encountered, the field in question is removed from the data entry but the entry itself is not deleted. However, if the cleansing process results in a building's failure to meet the BPD's minimum data requirements, then the entire building is removed. Minimum data requirements include [1] floor area, [2] climate zone, [3] facility type and [4] at least one year of measured energy use data.

*Figure 2: Distribution of residential buildings in Pennsylvania by energy use intensity. Upon investigation, the peak at 30 kBtu/ft$^2$¬year was attributed to estimated, not measured, energy use values.*

## 4. Results

This section describes the size and distribution of the database, as well as its potential usefulness as a decision-support tool. Results included here are based on the database as of January 2014, but the database is constantly growing.

### 4.1. How "big" is the database?

The database contains 44,000 commercial buildings and 700,000 residential buildings. All buildings report floor area and ASHRAE climate zone, and every building contains at least one year of energy use data. 98% of the buildings report sufficient data to calculate site and source energy use. Location and electricity

consumption data are required for all buildings, but beyond these minimum requirements, most data is relatively sparse. Building systems with the most data include roof and heating systems, which are reported for roughly 20% of buildings in the database. Other systems—including lighting, HVAC, windows and walls—include data for 2% or less of database records. While CBECS, RECS and CEUS include a number of buildings that report data for all of the systems listed, most data contributors do not provide any asset data.

The distribution of buildings in the database is influenced by the size and depth of new datasets. While CBECS, CEUS and RECS select buildings to survey based on the underlying distribution of the building stocks that those datasets represent, the BPD selects buildings solely on data availability and completeness. As a result, the distribution of buildings can shift with the inclusion of large source datasets that are focused on a specific region or market. For example, 650,000 buildings in the database are located in one of two California counties, comprising 92% of residential buildings and 87% of the entire database.

Figure 3 shows the relative frequency distribution of commercial and residential buildings in the BPD by building type compared to CBECS and RECS, respectively. The figure reveals that relative to the national building stock, the BPD has greater representation of office, retail and education buildings, but includes a fairly consistent representation of the residential building stock by building type. The greater representation of certain building types is unsurprising because many of the BPD's data contributors manage or monitor portfolios that consist of only one type of building. In one example, Kohl's Department stores submitted data for a number

of retail stores in its own portfolio. As a result, the database may contain a higher proportion of department stores than does the national building stock; another similar source dataset would skew the data further towards that building type. Although bias in the data affects the distribution of buildings relative to the national building stock, it means that the database may be particularly valuable to users interested in analyzing performance in specific markets or regions that are well represented in the data. The question of whether the BPD is "large enough" cannot be answered in general but only for specific research questions that are being explored using the BPD.

*Figure 3: Relative frequency distribution of BPD commercial and residential buildings by major building type compared to statistics of the national released by CBECS 2003 and RECS 2009.*

Figure 4 shows the relative frequency distributions of commercial and residential buildings in the BPD by census region, relative to statistics of the national building stock released by CBECS and RECS. The West census region is currently well represented among commercial buildings, and very well represented among residential buildings in the database. In the BPD, the West census region is heavily dominated by California. The largest residential source dataset, comprising 90% of residential buildings in the database, is located entirely in California. The distribution of commercial buildings may be attributable to the CEUS dataset, which is also located entirely in California, or due to high market penetration of

benchmarking programs (ENERGY STAR, 2012), and building certification programs in California (Simons et al., 2009).

*Figure 4: Relative frequency distribution of BPD commercial and residential buildings by census region compared to statistics of the national released by CBECS 2003 and RECS 2009.*

Figure 5 shows cumulative frequency distributions of annual energy use intensity for all retail and all office buildings in the BPD and in the CBECS dataset. The figure illustrates that retail buildings in the two databases follow similar distributions, while the distributions of office buildings differ substantially. Future research will further explore the causes of these differences and their implications for different use cases.

*Figure 5: Cumulative frequency distributions of site energy use intensity (kBtu/ft$^2$ - year) for retail and office buildings in BPD and CBECS.*

Although comparing distributions of BPD data to CBECS is useful for evaluating the national or regional representativeness of peer groups, the BPD also contains data-rich regions that are unavailable within a national-scale overview database like CBECS. For example, the BPD contains about 14,000 ENERGY STAR Labeled buildings. The DOE Buildings Data Energy Book estimates market penetration of the ENERGY STAR Label at 3.7% of the commercial building stock, or 22,000 buildings (Buildings Energy Data Book 2011, CBECS 2003), 65% of which are included in the

BPD. In another example, the BPD contains data collected under mandatory benchmarking ordinances in San Francisco, Seattle, Washington D.C., and New York. If compliance with these ordinances is high, then the BPD could contain a large fraction of the building stocks to which each ordinance applies. In data-rich regions of the database, the BPD may contain a large fraction of the corresponding subset of the building stock.

## 5. How useful is the data?

The database contains extensive low granularity data including location, size and building type. These fields are useful for benchmarking data and evaluating performance relative to a diverse peer group of buildings. The BPD presents data in histograms, showing quartiles by energy use for a selected peer group of buildings, allowing users to compare the performance of their own building or portfolio to its peers in the BPD (Figure 6). While this level of detail is sufficient for evaluating performance in very diverse portfolios of buildings (Hsu, 2014), more detailed data can be useful for other types of analysis, such as data-driven algorithms for estimating energy savings.

*Figure 6: Screen image of the BPD user interface, showing a histogram of energy use intensity for a peer group of office buildings. Source: BPD website (bpd.lbl.gov) designed by Building Energy Inc.*

One data-driven algorithm being vetted for release to BPD users fits a multiple linear regression model to physical & operational characteristics and equipment data to predict energy savings due to building retrofits (Walter et. al., 2014). Such an algorithm could provide a low-cost alternative to energy auditing, and add value to engineering and modeling-based estimates of energy savings by quantifying uncertainty in energy savings predictions. Uncertainty estimates can help potential investors to identify retrofits that not only maximize return, but also minimize risk. The accuracy of predictions made by such an algorithm would rely heavily on availability of building asset data. Currently, of the seven building systems included in models being developed for the BPD, the only datasets with complete or near-complete asset data are CEUS, CBECS and RECS. More than 18 of the BPD's 25 data contributors include entries with no asset data, totaling 67% of residential buildings and 87% of commercial buildings. The remaining 33% of residential buildings and 13% of commercial buildings, however, do have some level of asset data that can be used to fit models for estimating the energy use impact of different types of equipment. One opportunity for further research is to attempt to quantify the amount of data needed to fit models that will generate accurate energy savings predictions.

The "usefulness" of a database like the BPD can also be evaluated relative to alternative options for estimating energy savings and for evaluating energy use relative to a peer group of buildings. Although energy savings prediction accuracy has yet to be tested, statistical algorithms provide a promising, low-cost means of estimating energy savings. Energy audits and whole building modeling are labor and

skill-intensive, requiring investments that are prohibitive to some stakeholders—data-driven algorithms and peer group comparisons may be effective low-cost options for small-scale investors. In particular, homeowners may find the database to be a valuable tool both due to its low cost and because single-family homes are well represented in the database. Large-scale commercial investors, however, should still consider more targeted decision support tools such as auditing and simulation-based analysis, as these can more accurately account for building-specific conditions.

Despite the current data limitations of the BPD, particularly with respect to asset data, the inherent strength of the BPD is that it contains actual data from real buildings, which can be used to confirm results from simulated data. Many building decision-makers are concerned about savings analysis based on simulated data and validation against empirical data can help build confidence in energy savings estimates.

In discussing "usefulness", we have identified a number of specific questions that could be answered using the BPD. However, an application programming interface (API) to the BPD is publicly available to encourage development of commercial software tools that utilize the data in novel ways. The database was initially developed to satisfy an identified need for empirical energy consumption data, and as such current data collection and cleaning efforts, as well as presentation of data in the API and user interface, are geared towards applications in energy performance. However, the database also provides a wealth of information about physical and operational building characteristics, and analysis opportunities are by

no means limited to energy performance. The database is designed such that outreach efforts, database structure, data cleaning, and analysis tools can evolve as novel applications for the data emerge.

## 6. Conclusions and Outlook

This paper described a broad and concerted effort to collect and analyze existing data on the energy use and building characteristics of commercial and residential buildings in the United States. The effort resulted in the DOE Buildings Performance Database—the largest public domain database of commercial and residential buildings in the United States to date. The BPD provides a comparison of energy use intensities for user-customizable peer groups of buildings. It also allows analysis of energy impacts of various technologies, to the extent that such data are available for the buildings in the BPD.

The value of large databases like the BPD relative to existing databases lies in the large number of building records available for specific data-rich regions or markets in the database. Confirming that these peer groups adequately represent the underlying building stock is key to deriving actionable information from the data for many use-cases. A typical test for determining representativeness of building data is to compare the data against CBECS. However, CBECS is not representative for local or narrowly defined peer groups. As a result, comparisons with CBECS are less relevant in datasets with high penetration in certain markets but not others, as we demonstrated to be the case in the BPD. Further research is needed to develop a test for evaluating representativeness at the peer group level. Reporting on the

representativeness of every peer group in the BPD will not be feasible. As such, database users are tasked with verifying that relevant peer groups are adequately representative in regions or markets of interest.

Several conclusions can be drawn from the experience to date:

- The availability of building data on a large scale remains a challenge, especially data on building system characteristics. In theory, there is plenty of such data available—in drawings, specifications, maintenance records, etc. However, much of this data is effectively inaccessible for broader application because it is widely distributed, poorly archived, in custom formats, and lacks clarity on who owns the data and whether it can be shared.
- There is a major need to standardize building data. Literally every dataset imported into the BPD to date had its own unique data format and data field definitions. It has become clear that the lack of standard data formats, terms and definitions is a significant ongoing barrier to realizing the full potential of big energy data.
- While empirical data is valuable in what it can say about actual performance, it also tends to have a lot of "noise" that limits the ability to extract decision-grade information, especially for savings analysis. In the near term the primary application of such data is in peer-comparison and "sanity checking" of savings estimates.

The next phase of this effort will focus on increasing data breadth and depth, by exploring novel cost-effective ways of crowd-sourcing asset data. Additionally,

research efforts will focus on methods that are better suited to extract meaningful decision-grade information from sparse datasets.

**Acknowledgements**

This research was supported in part by the Assistant Secretary for Energy Efficiency and Renewable Energy of the U.S. Department of Energy (DOE), and performed under U.S. DOE Contract No. DE-AC02-05CH11231. In particular, the authors thank Elena Alschuler at the DOE for her support and management of the BPD Project. The authors thank Building Energy Inc. for contributing artwork for Figure 6.

**References**

Allcott, H., Greenstone, M., 2012. Is there an energy efficiency gap? Journal of Economic Perspectives 26, 1, 3-28.

Bryant, R.E., Katz, R.H., Lazowska, E.D., 2008. Big-data computing: creating revolutionary breakthroughs in commerce, science and society. Computational Research Association.

California Energy Commission, 2002. California Commercial End-Use Survey.

Chidiac, S.E., Catania, E.J.C., Morofsky, E., Foo, S., 2011. Effectiveness of single and multiple energy retrofit measures on the energy consumption of office buildings. Energy 36, 8, 5037-5052. http://dx.doi.org/10.1016/j.energy.2011.05.050

Dalenius, T., Reiss, S.P., 1982. Data-swapping: a technique for disclosure control. Journal of Statistical Planning and Inference 6, 1, 73-85. http://dx.doi.org/10.1016/0378-3758(82)90058-1

Deng, Q., Zhang, L., Cui, Q., Jiang, X., 2014. A simulation-based decision model for designing contract period in building energy performance contracting. Building and Environment 71, 71-80. http://dx.doi.org/10.1016/j.buildenv.2013.09.010

Eisenhower, B., O'Neill, Z., Fonoberov, V.A., Mezic, I., 2012. Uncertainty and sensitivity decomposition of building energy models. Journal of Building Performance Simulation 5, 3, 171-184. 10.1080/19401493.2010.549964

ENERGY STAR, 2014. Portfolio Manager Website. Accessed January 10, 2014. www.portfoliomanager.energystar.gov

ENERGY STAR, 2013. Portfolio Manager Technical Reference: ENERGY STAR Score. Accessed January 10, 2014. http://www.energystar.gov/buildings/facility-owners-and-managers/existing-buildings/use-portfolio-manager

ENERGY STAR, 2013. Portfolio Manager Data Trends: ENERGY STAR Certification. Accessed January 10, 2014. http://www.energystar.gov/buildings/sites/default/uploads/tools/DataTrends_Energy_20121002.pdf?0e8f-2275

ENERGY STAR, 2012. Portfolio Manager Data Trends: Energy Use Benchmarking. Accessed January 10, 2014. https://portfoliomanager.energystar.gov/pdf/reference/ENERGY%20STAR%20Score.pdf?ec2b-0568

Gold, R., Elliot, R.N., 2010. Where have all the data gone? The crisis of missing energy efficiency data. ACEEE Report No. E101.

Hsu, D., 2014. How much information disclosure of building energy performance is necessary?. Energy Policy 64, 263-272.

Issenberg, S., 2012. The Victory Lab: The secret science of winning campaigns. Crown.

Itron, Inc., 2006. California Commercial End-Use Survey. Technical report prepared for the California Energy Commission.

Kargupta, H., Datta, S., Want, Q., Sivakumar, K., 2005. Random-data perturbation techniques and privacy-preserving data mining. Knowledge and Information Systems 7, 4, 387-414.

Lovins, A., Rocky Mountain Institute, 2011. Reinventing Fire: Bold Business Solutions for the New Energy Era. Chelsea Green Publishing.

Marcus, Gary, and Ernest Davis. 2014. "Eight (No, Nine!) Problems With Big Data." The New York Times, April 6. http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html.

Mathew, P., Clear, R., Kircher, K., Webster, T., Lee, K.H., Hoyt, T. 2010. "Advanced Benchmarking for Complex Building Types: Laboratories as an Exemplar," Proceedings of the 2010 ACEEE Summer Study of Energy Efficiency in Buildings. ACEEE, Washington, D.C.

McKinsey & Co., 2009. Pathways to a low-carbon economy: Version 2 of the global greenhouse gas abatement cost curve.

Mills, E., 2009. Building Commissioning: A golden opportunity for reducing energy costs and greenhouse gas emissions. Technical report prepared for the California Energy Commission.

Mills, E., Mathew, P., Piette, M.A., Bourassa, N., Brook, M., 2008. Action-Oriented Benchmarking: Concepts and Tools. Energy Engineering 105, 4, 21-40.

New Buildings Institute, 2008. LEED Case Study Database. Access February 10, 2014. http://buildings.newbuildings.org/

New Buildings Institute, 2008. Energy Performance of LEED for New Construction Buildings. Technical report prepared for the U.S. Green Building Council.

National Research Council Committee on National Statistics, 2012. Effective Tracking of Building Energy Use: Improving the Commercial Buildings and Residential Energy Consumption Surveys. National Academies Press.

Nicholls, C., 2014. Energy use in non-domestic buildings: the UK government's new evidence base. Building Research & Information 42, 1, 109-117. http://dx.doi.org/10.1080/09613218.2014.832484.

Ryan, E.M., Sanquist, T.F., 2012. Validation of building energy modeling tools under idealized and realistic conditions. Energy and Buildings 47, 375-382. 10.1016/j.enbuild.2011.12.020

O'Neill, Z., Shashanka, M., Pang, X., Bhattacharya, P., Bailey, T., Haves, P., 2011. Real time mode-based energy diagnostics in buildings. Proceedings of Building Simulation, 2011.

Ogas, Ogi. 2013. "Beware the Big Errors of 'Big Data.'" WIRED. February 8. http://www.wired.com/2013/02/big-data-means-big-errors-people/.

Pacala, S., Socolow, R., 2004. Stabilization Wedges: Solving the Climate Problem for the Next 50 Years with Current Technologies. Science 305, 5686, 968-972.

Pérez-Lombard, L., Ortiz, J., Pout, C., 2008. A review on buildings energy consumption information. Energy and Buildings 30, 394-398.

Satchwell, A., Goldman, C., Larsen, P., Gilligan, D., Singer, T., 2010. A Survey of the U.S. ESCO Industry: Market growth and development from 2008 to 2011. Technical Report LBNL-3479E, Lawrence Berkeley National Laboratory.

Simons, R.A., Choi, E., Simons, D.M., 2009. The effect of state and city green policies on the market penetration of green commercial buildings. Journal of Sustainable Real Estate 1, 1, 139-166.

Walter, T., Price, P.N., Sohn, M.D., 2014. Uncertainty estimation improves energy measurement and verification procedures. Applied Energy 130, 230-236. 10.1016/j.apenergy.2014.05.030.

Williams, J.H., DeBenedictis, A., Ghanadan, R., Mahone, A., Moore, J., Morrow, W.R., Price, S., Torn, M.S., 2011. The Technology Path to Deep Greenhouse Gas Emissions Cuts by 2050: The Pivotal Role of Electricity. Science 333, 6064, 53-59. 10.1126/science.1208365.

U.S. Departments of Transportation and Justice, 2009. Data-Driven Approaches to Crime and Traffic Safety (DDACTS): Operational Guidelines. Technical Report.

U.S. Energy Information Administration, 2003. Commercial Building Energy Consumption Survey (CBECS). Technical Report.

U.S. Energy Information Administration, 2009. Residential Energy Consumption Survey (RECS). Technical Report.

U.S. Energy Information Administration, 2011. Annual Energy Outlook. Technical Report.

U.S. DOE Office of Energy Efficiency and Renewable Energy, 2008. Energy Efficiency Trends in Residential and Commercial Buildings. Technical Report.

U.S. DOE Office of Energy Efficiency and Renewable Energy, 2011. Buildings Energy Data Book. Technical Report.

U.S. DOE, 2014. Building Energy Data Exchange Specification (BEDES). Accessed February 10, 2014. http://energy.gov/eere/buildings/building-energy-data-exchange-specification-bedes

U.S. DOE, Lawrence Berkeley National Laboratory, 2014. Buildings Performance Database. Accessed February 10, 2014. www.bpd.lbl.gov