



Exploratory Analysis of Energy Use Across Building Types and Geographic Regions in the United States

Michael D. Sohn^{1*} and Laurel N. Dunn²

¹ Lawrence Berkeley National Laboratory, Berkeley, CA, United States, ² Department of Civil and Environmental Engineering, University of California, Berkeley, Berkeley, CA, United States

Collecting energy use data is becoming common practice in the buildings sector. Current applications include understanding regional energy flows in the building stock, and tracking energy performance of individual buildings. Beyond these, research and commercial applications of building energy data are as yet unexplored. Research is needed to provide insight into the data being collected, to identify appropriate applications of these data, and opportunities to improve data collection efforts. To that end, we present an exploratory analysis of the existing public large-scale building energy datasets, focusing on the two largest datasets: the Commercial Building Energy Consumption Survey and the Building Performance Database. We provide background information on both datasets, present an overview of the detail and sparsity of information in each, and report on the relationships we observe between data fields included in the two datasets, and compare our findings with results from the literature. We discuss how these results could be applied to support energy efficiency investments, and opportunities to improve data collection efforts to ensure that the data collected are adequate to provide insight into building energy consumption and support novel applications of building energy data.

OPEN ACCESS

Edited by:

Nahid Mohajeri,
University of Oxford, United Kingdom

Reviewed by:

Graziano Salvalai,
Politecnico di Milano, Italy
Hui Yan,
South China University of Technology,
China

*Correspondence:

Michael D. Sohn
mdsohn@lbl.gov

Specialty section:

This article was submitted to
Urban Science,
a section of the journal
Frontiers in Built Environment

Received: 13 April 2019

Accepted: 26 August 2019

Published: 12 September 2019

Citation:

Sohn MD and Dunn LN (2019)
Exploratory Analysis of Energy Use
Across Building Types and
Geographic Regions in the
United States.
Front. Built Environ. 5:105.
doi: 10.3389/fbuil.2019.00105

Keywords: building performance database, commercial buildings energy consumption survey, building energy data, building energy benchmarking, exploratory data analysis

1. INTRODUCTION

In recent years, empirical data and data-driven decision support tools have been transformational in numerous industries including marketing (Bryand et al., 2008), crime-fighting (U.S. Departments of Transportation and Justice, 2009), and political campaigning (Issenberg, 2012). The success of these tools in other industries has led to speculation about the role big data can play in buildings, for example to scale up investments in energy efficiency.

Numerous studies (e.g., Pacala and Socolow, 2004; McKinsey & Co., 2009; Williams et al., 2011) suggest that relatively modest investments in energy efficiency at scale can yield large energy and economic savings for the buildings sector. These studies rely on engineering-based models to predict energy savings. Uncertainty in these predictions combined with high retrofit costs are thought to present a major barrier to eliciting the scale of investment needed to unlock deep energy savings in the building stock (Mills et al., 2004).

Unlike engineering-based models, data-driven algorithms can quantify uncertainty in energy use and energy savings predictions, with or without detailed information about building

characteristics. By quantifying uncertainty, these algorithms can help stakeholders to identify low-risk and/or high-savings building retrofits. However, data-driven algorithms are thought to be limited by the quantity, quality, and scope of the supporting data.

A recent increase in building data collection has led to parallel efforts to aggregate (Mathew et al., 2015) and summarize (Kontokosta, 2012a; Hsu, 2014a,b) the available data, and to develop data-driven algorithms for modeling building energy consumption (Kontokosta, 2012b; Hsu, 2014a, 2015; Walter et al., 2014). Mathew et al. (2015) and Brown et al. (2014) identify the Building Performance Database (BPD) (U.S. Department of Energy, 2015a) as a candidate for supporting data-driven algorithms to inform investments in energy efficiency. Another candidate is the Commercial Buildings Energy Consumption Survey (CBECS) (U.S. Energy Information Administration, 2003).

The current work presents an exploratory analysis summarizing the building data available in the BPD and in CBECS, which are among the largest and most detailed sources of data about building energy consumption in the United States. This analysis presents just one of many possible studies these data could support, and is the first of several publications that rely these data. With this work, we aim to understand: (1) what information are available, (2) what are the limitations of these datasets, (3) what insights do these datasets provide into similarities/differences in energy use patterns among buildings, (4) are these insights consistent across the two datasets and with results in the literature, and (5) what additional data (if any) are needed to support the potential big-data applications discussed above. The key contributions of this work are to determine whether we can leverage the empirical building data available today confirm or refute conventional knowledge about building energy use and identify/quantify energy efficiency opportunities, or whether more detailed data are needed to support these applications.

The current paper is structured as follows: section 2 reports on the emphasis and content of the two datasets. Section 3 outlines our analysis approach. Section 4 highlights the range and distribution of reported values for various data fields in different subsets of each database, and identifies relationships between these variables. Section 5 provides an interpretation of the results in section 4, and discusses the implications of these results on current and future applications of the data.

2. DATA

2.1. Building Performance Database (BPD)

When we performed this analysis (March 2015), the BPD contained 45,000 commercial and 650,000 residential building

records. To our knowledge, the BPD contains more buildings than any other public building dataset. BPD data are collected by independent agencies and volunteered for inclusion in the database. To date, more than 50 contributors have provided data (U.S. Department of Energy, 2015b). Contributors include researchers, building owners, local governments, electric utilities, and federal agencies. CBECS is among the BPD's source datasets.

Unlike CBECS, representativeness is not considered in compiling data for the BPD. Because BPD relies on volunteered data, the dataset has an inherent bias toward buildings with benchmarked energy use. In practice, policy and market influences make benchmarking more common among certain subsets of the building stock than among others. Furthermore, individual contributors typically provide data for portfolios of buildings they own, manage or track for other reasons. More often than not, these portfolios share some attribute such as: geographic proximity, common ownership, or participation in an energy efficiency or building certification program. For example, because energy regulation typically happens at the local or state (rather than federal) level, certain regions are very well represented in the BPD, while others may be unrepresented altogether.

The BPD contains only the information provided by data contributors. Thus the level of detail in the dataset is constrained by the interests and expertise of the organizations/individuals that collect and/or volunteer the data. **Table 1** lists reporting frequencies for the 24 most populated data fields in BPD; CBECS reporting frequencies for the same fields are also provided. The most commonly reported data fields (e.g., gross floor area) constitute the minimum requirements for inclusion in the database, as detailed in Custodio et al. (2014). The completeness and quality of other data fields is varies among data contributors. However, the BPD reports only measured (not predicted) values. Refer to Mathew et al. (2015) for further detail about BPD contents, data sources, and procedures for aggregation and quality control.

2.2. Commercial Buildings Energy Consumption Survey (CBECS)

To our knowledge, CBECS is the most detailed public repository of commercial building energy and characteristics data available today. Because CBECS is collected through an in-person survey that covers all reported data fields, reporting frequencies are high for all surveyed data fields.

Through CBECS, the Energy Information Administration (EIA) aims to provide a representative snapshot of the U.S. building stock. The present analysis uses the 2003 survey data. Buildings in CBECS are sampled so as to capture a diverse cross-section of the building stock, and weighted based on the number of similar buildings in the larger building stock.

Because BPD is not representative of the larger building stock, the current analysis aims to understand relationships between variables in the two datasets. We do not intend to make inferences about relationships in the underlying building stock. Thus our analysis reports on the un-weighted values provided in CBECS

Abbreviations: BPD, Buildings Performance Database; CBECS, Commercial Buildings Energy Consumption Survey; CDD, Cooling Degree Days; DOE, Department of Energy; EIA, Energy Information Administration; EUI, Energy Use Intensity; GIS, Geographic Information System; HDD, Heating Degree Days; ISD, Integrated Surface Database; NOAA, National Oceanic and Atmospheric Administration; RECS, Residential Energy Consumption Survey; kBtu, Thousand British Thermal Units; MJ, Megajoules (million joules); m^2 , Square meters.

TABLE 1 | Reporting frequencies in BPD and CBECS for the 24 most frequently reported data fields in the Buildings Performance Database.

	Data field	Reporting Frequency	
		BPD	CBECS
1	ASHRAE climate zone	1.00	0.00
2	Gross floor area	1.00	1.00
3	Building classification	1.00	1.00
4	Building type	1.00	1.00
5	Postal code	0.94	0.00
6	City	0.94	0.00
7	Energy measurement interval	0.85	1.00
8	Site energy consumption	0.85	0.98
9	Electricity consumption	0.80	0.98
10	Source energy consumption	0.80	0.98
11	Fuel consumption	0.61	0.98
12	Year built	0.34	1.00
13	Weekly operating hours	0.33	0.92
14	Number of occupants	0.32	0.92
15	Heating type	0.10	0.85
16	Cooling type	0.10	0.81
17	Heating fuel	0.09	0.92
18	Lighting type	0.09	0.89
19	Roof material	0.09	0.92
20	Wall material	0.09	0.92
21	Window glass layers	0.09	0.92
22	Window glass type	0.07	0.89
23	Wall R-value	0.03	0.00
24	HVAC flow control type	0.03	0.85

building records and does not take into account the number of buildings each record represents in the larger building stock.

3. METHODS

3.1. Exploratory Data Analysis

The current work aims to identify relationships between energy use and building characteristics using exploratory analysis (Behrens, 1997). We report on the range, distribution, Pearson correlation coefficient, and the slope of a least squares regression line fitted to the data. From these results, we identify similarities/differences in energy use among peer groups of buildings in the two datasets, and relationships between numeric data fields.

We select fields for analysis from each database based on data availability. Although CBECS contains many more data fields than we report on, the scope of the current work is limited to fields that are well populated in both datasets. We evaluate trends between energy use and each of five data fields: (1) gross floor area, (2) year built, (3) average weekly operating hours, (4) HDD, and (5) CDD.

3.2. Units of Measurement

We report energy use in terms of site (or delivered) energy consumption. Site energy consumption includes all electricity

and other fuels (e.g., natural gas, fuel oil) used in a building (e.g., for heating and/or cooking). We follow the convention of reporting site energy consumption in SI units (MJ) to aggregate fuels streams characteristically reported in different units of measure, for example in kWh or by volume. Energy use is also commonly reported in terms of source (or primary) energy consumption. Source energy consumption is typically used to understand regional energy flows, while site energy consumption is used to understand energy flows and expenditures within a building. The two values can differ substantially due to losses during heat conversion and electricity transmission. Because the focus of this work is to report similarities/differences between individual buildings, rather than on drawing inferences about the larger building stock, we report building energy consumption in terms of annual site energy use (*MJ*) and annual site energy use intensity (*MJ/m²*).

3.3. Peer Group Analysis

Differences in energy use among buildings are related not only to physical and operational characteristics, but also to differences in energy useage patterns in various building types and climates. To highlight correlations in the data, we control for extraneous differences among buildings in two ways: (1) normalizing by observed correlates, and (2) filtering the database to compare buildings only with their peers with respect to building type or climate.

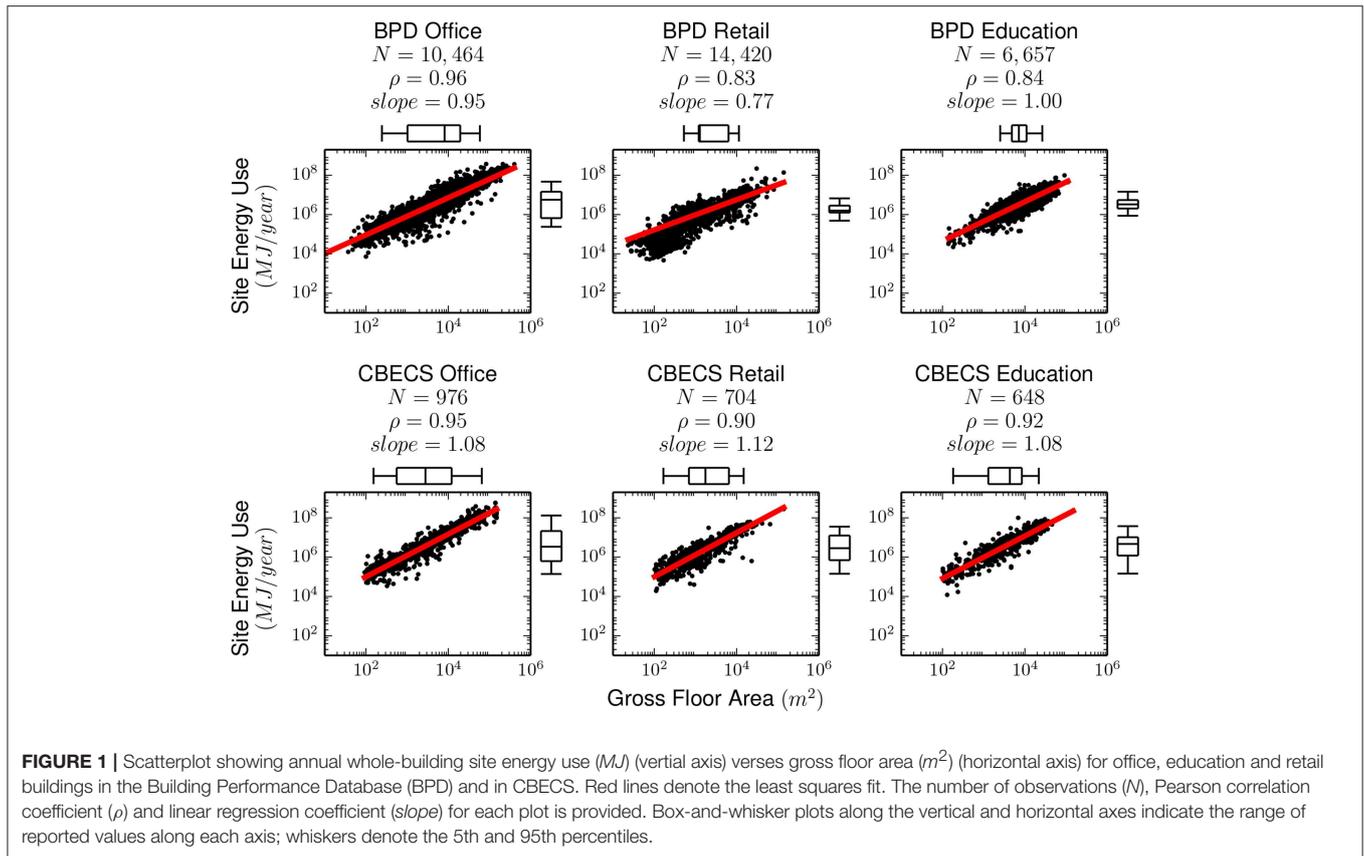
By comparing buildings with their peers, we aim to control for unreported or extraneous characteristics/consumption patterns common to certain subsets of the building stock. For example, defining regional peer groups controls for regional differences in building codes and heating/cooling requirements related to climate.

We identify trends in peer groups defined first by building type, then by building location. The building types we examine include: education, office, and retail. Combined, these account for 75% of commercial buildings in the BPD, and 45% of buildings in CBECS. The locations we examine include: San Francisco and New York City. Although the two cities make up only a small fraction of buildings in BPD, they are illustrative examples of disparate climates.

CBECS building records report locations by census division, but provide no more granular location information. Therefore, we compare local BPD peer groups with CBECS peer groups defined by the census divisions that contain each local peer group. Although the local and regional building stocks differ, we expect differences in climate to be evident both locally (in BPD) and regionally (in CBECS).

3.4. Heating and Cooling Degree Days Calculation

Each building record in CBECS includes annual heating and cooling degree days (HDD and CDD) for the survey year, in this case 2003. Although HDD and CDD are not reported in the BPD, they can be computed using building location (postal code) and the energy measurement interval, given by start and end time stamps for energy use readings. We compute HDD and CDD by linking BPD data with orthogonal temperature data



obtained from NOAA's Integrated Surface Database (ISD) (U.S. National Oceanic and Atmospheric Administration, 2014). We use reported building locations to identify the nearest weather station, and hourly temperature data for that weather station to compute HDD and CDD. We use base $65^\circ F$ ($18.3^\circ C$), as detailed in ASHRAE (2004). Only about 5% of building records in BPD report insufficient data to compute HDD and CDD using this approach.

4. RESULTS

In the following sections, we identify and compare relationships between data fields for various peer groups of buildings in BPD and in CB ECS.

4.1. Peer Groups by Building Type

Figure 1 shows a strong correlation between whole-building (site) energy use and gross floor area. Pearson correlation coefficients range from $\rho = 0.83$ to $\rho = 0.96$ among peer groups. Regression line slopes range from $m = 0.77$ to $m = 1.12$; the narrow range suggests that energy use scales similarly with respect to floor area for all three building types and in both datasets. Box-and-whisker plots illustrate the range and distribution of reported values along each axis.

BPD peer groups in **Figure 1** contain between 10 and 20 times more buildings than the corresponding peer groups in CB ECS. As we examine other data fields with lower reporting frequencies,

BPD peer group size declines by an order of magnitude due to data sparsity. Although CB ECS peer group sizes do not change, BPD peer groups consistently outnumber their counterparts in CB ECS.

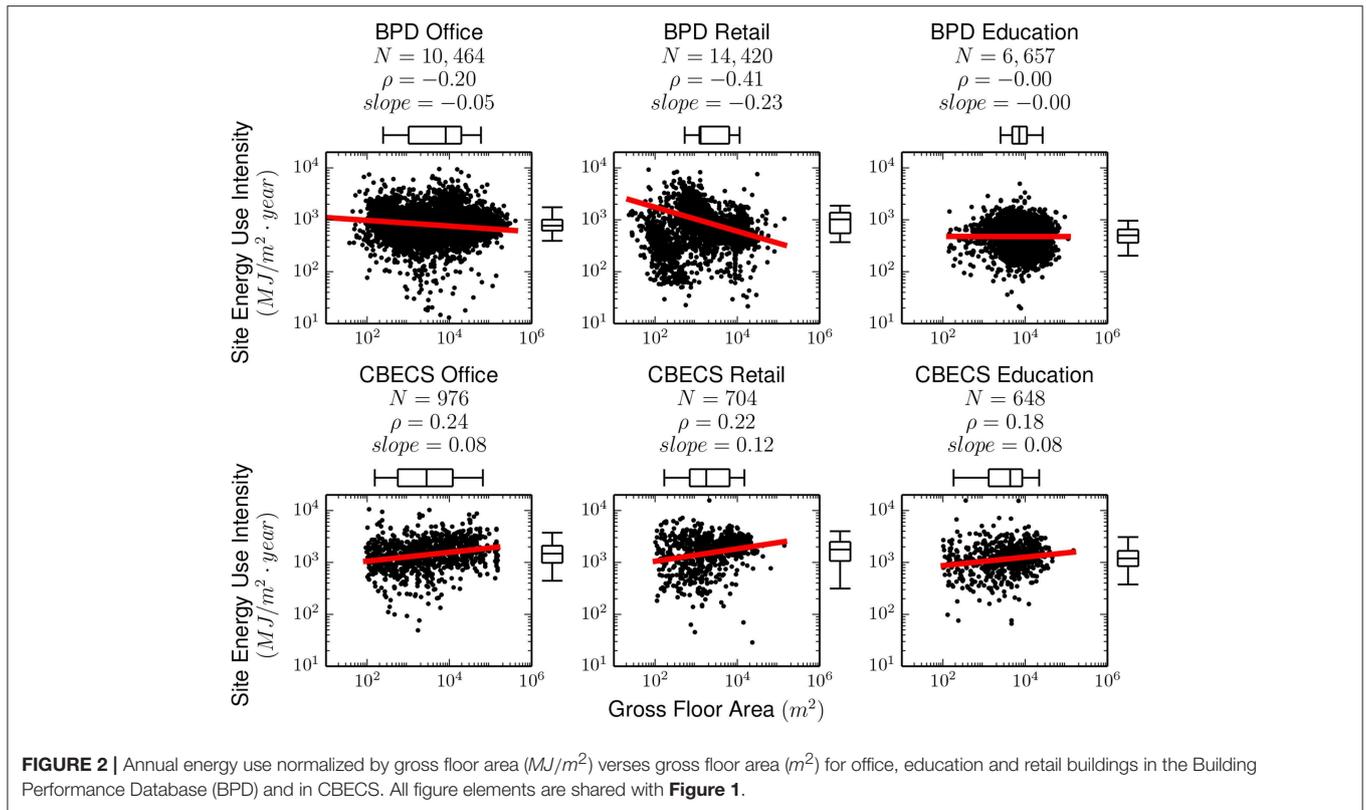
The median energy use is similar across building types and datasets. The similarity is surprising, as some education buildings are only operational during the school year. The range and distribution of floor areas and energy use are also similar across building types and in both datasets. Retail and education buildings in BPD do report a narrower distribution of energy use than their counterparts in CB ECS.

To control for the effect of floor area on site energy use, we examine subsequent trends in terms of annual energy use intensity (EUI) with units MJ/m^2 . **Figure 2** reveals no correlation between EUI and floor area, with correlation coefficients ranging from $\rho = -0.41$ to $\rho = 0.24$. The range and distribution of values along the EUI axis are the same across all six peer groups.

Table 2 reports summary statistics by peer group for the most frequently reported data fields. We also list correlation and linear regression coefficients relating each data field to site EUI. We observe no strong correlations between site EUI and any other variables in the database.

4.2. Peer Groups by Building Location

Table 3 lists summary statistics for data fields by location. Again, we observe no strong correlations between site EUI and any other variables in either location. Given the differences in climate



between San Francisco and New York, we expect that heating and cooling loads will be higher in New York than in San Francisco, and that these differences will be evident in site EUI. Although the distribution of site EUI for buildings in New York is shifted slightly to the right of the distribution for buildings in San Francisco, the difference is minimal, and the range of values is roughly the same in the two regions. These observations pertain to both BPD and CBECS peer groups.

We further investigate these differences in EUI by examining the cumulative distribution of EUI for various fuel types in each local peer group, as shown in **Figure 3**. Surprisingly, comparing buildings in San Francisco and New York, neither peer group is clearly more energy intensive than the other. Similarly, electricity use intensities among the two peer groups in CBECS are identically distributed. We do find that buildings in the regions with more extreme winters (New York and the Middle Atlantic) use slightly more fuel than buildings in the regions with milder climates (San Francisco and the Pacific). However, the observed differences in EUI are surprisingly small considering the vast differences in climate.

5. DISCUSSION

5.1. Distributions of Values

Comparing the range and distribution of values in each data field listed in **Tables 2, 3**, we observe only slight differences among buildings in the two datasets and among peer groups, including all three building types and both locations. We find that BPD peer groups are consistently more narrowly

distributed than their counterparts in CBECS. Two factors contribute to the narrow distribution of values in BPD relative to CBECS. First, the range of values reported in CBECS is likely wide because CBECS captures a diverse cross section of the building stock (Hsu, 2015); the distributions of values in the building stock and in the weighted CBECS dataset are likely narrower than what we observe in the unweighted dataset. Second, BPD may represent only a narrow cross section of the building stock due to self-selection of buildings represented in the dataset.

5.2. Energy Use Correlates

The only clear correlation we observe is between floor area and whole-building energy use (**Figure 1**) with correlation coefficients (ρ) ranging from 0.83 to 0.96. The strong correlation is not surprising, and confirms conventional knowledge that energy use scales with floor area. Floor area is a key input in physics-based building models (Kavgic et al., 2010; Deru et al., 2011; U.S. Department of Energy, 2014).

Both Hsu (2014a) and Kahn et al. (2014) report a negative correlation between EUI and floor area in local subsets of the commercial building stock. We observe a weak negative correlation ($\rho = -0.43$) among BPD buildings in New York, which is consistent with results in Hsu (2014a) for a comparable peer group of buildings in New York. However, we observe no such correlation in any other peer group.

Kontokosta (2012b) reports a positive correlation between EUI and operating hours among buildings in New York City; we also observe a weak correlation among BPD buildings in New

TABLE 2 | Table showing range and distribution of reported values in BPD and CBECS for office, education and retail buildings.

Type	Field	Source	N	Min	Q1	Q2	Q3	Max	Avg	ρ	m	
Retail	Site	BPD	14,420	0.00	1.26	1.59	2.85	219.83	2.42	-0.05	-8.17	
	Energy	CBCECS	704	0.02	0.70	2.86	12.63	279.40	9.21	0.27	21.90	
	Site	BPD	14,420	22	548	1,018	1,371	9,249	1,019	1.00	1.00	
	EUI	CBCECS	704	29	1,068	1,771	2,488	15,650	1,912	1.00	1.00	
	Floor	BPD	14,420	0.02	1.20	1.28	6.37	139.35	3.59	-0.44	-56.77	
	Area	CBCECS	704	0.10	0.70	1.75	6.50	139.35	4.71	0.07	11.47	
	Year	BPD	3,593	1880	1998	2002	2006	2012	1999	-0.09	-1.78	
	Built	CBCECS	704	1800	1964	1980	1993	2003	1974	0.22	10.05	
	HDD	BPD	9,684	0	1,220	2,320	2,983	7,408	2,162	-0.11	-0.05	
		CBCECS	704	0	1,215	2,291	3,287	5,379	2,256	-0.00	-0.00	
		CDD	BPD	9,684	0	593	816	1,249	2,925	996	0.22	0.20
			CBCECS	704	19	375	642	1,094	3,086	867	0.09	0.19
Office	Operating	BPD	2,654	28	75	85	97	168	86	0.13	2.62	
	Hours	CBCECS	355	4	50	62	80	168	69	0.38	18.91	
	Site	BPD	10,464	0.01	0.66	5.67	14.50	374.34	12.27	0.15	3.75	
	Energy	CBCECS	976	0.03	0.63	3.40	21.32	585.46	25.27	0.35	7.44	
	Site	BPD	10,464	13	618	771	1,005	9,495	889	1.00	1.00	
	EUI	CBCECS	976	49	981	1,480	2,096	10,576	1,726	1.00	1.00	
	Floor	BPD	10,464	0.01	1.02	8.15	19.51	402.43	15.70	-0.08	-1.76	
	Area	CBCECS	976	0.09	0.56	2.79	12.08	148.64	13.56	0.10	4.33	
	Year	BPD	2,359	1776	1963	1981	1992	2011	1972	0.13	3.04	
	Built	CBCECS	976	1826	1960	1977	1988	2003	1969	0.09	3.63	
	HDD	BPD	7,021	0	1,602	2,332	2,872	5,966	2,218	0.14	0.07	
		CBCECS	976	0	1,440	2,683	3,455	6,144	2,495	0.06	0.06	
Education		BPD	7,021	9	561	784	1,157	2,925	927	-0.06	-0.05	
		CBCECS	976	11	372	521	842	3,213	723	-0.02	-0.05	
	Operating	BPD	6,526	8	49	58	67	168	61	0.17	4.05	
	Hours	CBCECS	976	2	45	54	65	168	66	0.29	9.16	
	Site	BPD	6,657	0.02	2.03	3.30	5.47	98.70	4.85	0.43	21.04	
	Energy	CBCECS	648	0.01	1.20	4.70	10.35	240.56	9.54	0.36	26.50	
	Site	BPD	6,657	20	360	502	654	4,949	531	1.00	1.00	
	EUI	CBCECS	648	66	848	1,197	1,671	15,597	1,414	1.00	1.00	
	Floor	BPD	6,657	0.14	4.84	7.16	11.01	110.61	9.62	0.01	0.17	
	Area	CBCECS	648	0.10	1.29	4.32	8.45	148.64	6.67	0.06	7.67	
	Year	BPD	1,933	1853	1929	1956	1973	2013	1953	-0.01	-0.17	
	Built	CBCECS	648	1848	1956	1969	1987	2003	1968	0.01	0.45	
HDD	BPD	5,177	111	1,928	2,484	3,060	5,966	2,485	0.24	0.07		
	CBCECS	648	94	1,307	2,564	3,366	5,311	2,371	0.00	0.00		
	CDD	BPD	5,177	7	570	710	976	2,925	811	-0.04	-0.02	
		CBCECS	648	73	358	583	1,000	3,280	800	0.01	0.02	
	Operating	BPD	214	4	40	50	60	168	54	0.41	6.36	
	Hours	CBCECS	648	0	42	54	70	168	58	0.32	15.95	

For each data field we report the number of buildings (N), the minimum (Min), and maximum (Max) values, the quartiles of each distribution (Q1–Q3), and the average value (Avg), as well as the correlation coefficient (ρ) and slope of least square regression line (m) relating each data field to energy use intensity.

York City and in San Francisco (see **Table 3**). Despite evidence of a weak correlation in select peer groups, fitted regression lines relating operating hours to EUI have approximately zero slope for nearly all peer groups. In other words, we observe no change in EUI with respect to different operating hours.

The lack of a relationship linking EUI to operating hours is surprising, and seems to suggest that buildings use as much energy while unoccupied as they do while occupied. Unoccupied buildings can often improve energy performance by scheduling reductions in loads designed to support occupant needs, such

TABLE 3 | Table showing range and distribution of reported values in BPD and CBECS for office, education and retail buildings.

Location	Field	Source	N	Min	Q1	Q2	Q3	Max	Avg	ρ	m
San Francisco	Site	BPD	1,131	0.04	0.90	3.19	12.25	165.56	10.88	0.29	10.42
	Energy	CBECS	633	0.00	0.44	2.12	10.93	1163.65	16.30	0.29	7.85
	Site	BPD	1,131	15	372	629	902	6,920	780	1.00	1.00
	EUI	CBECS	633	6	589	1,160	1,961	16,405	1,576	1.00	1.00
	Floor	BPD	1,131	0.15	2.32	5.82	18.00	182.48	14.87	-0.01	-0.24
	Area	CBECS	633	0.09	0.56	2.18	7.43	148.64	8.95	0.10	7.74
	Year	BPD	480	1850	1916	1942	1979	2011	1947	0.00	0.05
	Built	CBECS	633	1850	1960	1977	1991	2003	1973	0.07	5.10
	HDD	BPD	39	2,263	2,263	2,263	2,283	2,302	2,273	0.32	14.91
		CBECS	633	0	729	1,124	2,536	6,144	1,581	0.04	0.06
	CDD	BPD	39	9	9	9	11	12	10	0.32	200.29
		CBECS	633	11	309	497	893	3,105	613	-0.00	-0.02
	Operating	BPD	150	33	54	60	70	168	66	0.56	9.06
	Hours	CBECS	588	0	45	60	98	168	78	0.32	10.81
	New York	Site	BPD	717	0.00	1.81	4.02	15.39	374.34	19.31	0.08
Energy		CBECS	669	0.01	0.87	3.80	19.92	777.43	32.53	0.30	7.85
Site		BPD	717	29	536	888	1,496	9,203	1,223	1.00	1.00
EUI		CBECS	669	19	833	1,458	2,517	25,345	2,072	1.00	1.00
Floor		BPD	717	0.17	1.56	8.58	25.35	255.48	22.39	-0.17	-5.24
Area		CBECS	669	0.10	0.91	3.39	12.63	130.06	13.42	0.11	9.82
Year		BPD	396	1799	1920	1949	1970	2011	1945	-0.04	-0.89
Built		CBECS	669	1771	1940	1966	1983	2003	1958	0.02	1.59
HDD		BPD	433	2,198	2,456	2,513	2,513	2,765	2,455	-0.50	-4.75
		CBECS	669	1,906	2,856	3,197	3,589	5,393	3,259	-0.10	-0.42
CDD		BPD	433	472	708	710	710	945	725	0.36	8.38
		CBECS	669	123	303	437	576	862	445	0.10	1.35
Operating		BPD	151	31	55	64	75	122	67	0.45	5.27
Hours		CBECS	635	0	50	64	116	168	84	0.35	15.41

For each data field we report the number of buildings (N), the minimum (Min), and maximum (Max) values, the quartiles of each distribution (Q1–Q3), and the average value (Avg), as well as the correlation coefficient (ρ) and slope of least square regression line (m) relating each data field to energy use intensity.

as HVAC and lighting. Thus the lack of a relationship between EUI and operating hours may point to an opportunity to reduce consumption while buildings are unoccupied.

Results from the literature reporting on correlations between EUI and year built are mixed. Kontokosta (2012b) and Kahn et al. (2014) observe a positive correlation between EUI and year built, while Kolter and Ferreira (2011) observes no correlation. Our own results show no correlation in any peer group. That we observe no correlation does not necessarily prove a relationship does not exist, but rather that year built has less pronounced impacts on EUI than other differences between buildings.

Walter et al. (2014) observe a relationship between temperature and energy use in individual buildings at high temperatures (or high CDD). Our own results show no similar correlation between EUI and either HDD or CDD. The lack of a correlation seems to suggest that buildings in moderate climates use as much energy as buildings in extreme climates. This result is surprising, as weather-sensitive loads (i.e., heating and cooling) often constitute a large portion of whole-building energy use.

If confirmed, this result could reveal an opportunity to improve energy performance among buildings in moderate climates. However, the lack of a correlation may also suggest that differences in weather have less pronounced impact on EUI than other differences between buildings; controlling for those differences could reveal correlations not evidenced in **Tables 2, 3**.

To further explore the relationship between energy and climate, we compare whole-building, electricity, and fuel intensity between buildings in an extreme climate (New York City) and a moderate climate (San Francisco), shown in **Figure 3**. Based on climate normals, New York typically experiences twice as many HDDs and 20 times as many CDDs annually than San Francisco (U.S. National Oceanic and Atmospheric Administration, 2013). However, for the years on record in BPD, buildings in New York and San Francisco experience on average about the same HDDs (4,400 and 4,100, respectively), and buildings in New York only experienced about 10 times as many CDDs than buildings in San Francisco (on average 1,300 and 18, respectively). The observed similarities in HDD

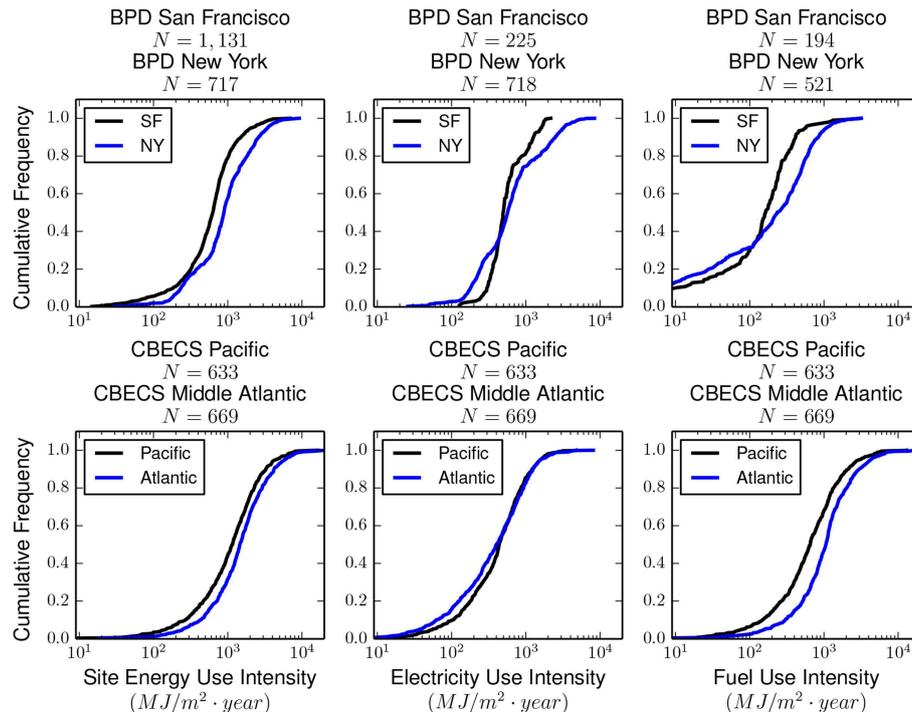


FIGURE 3 | Cumulative distribution of annual whole-building, electricity, and fuel energy consumption normalized by gross floor area (MJ/m^2) for commercial buildings in San Francisco and New York in the Building Performance Database (BPD), and commercial buildings in the Pacific and Middle Atlantic census divisions in CBECS.

are due to a particularly mild winter in New York in 2011-12 (U.S. National Oceanic and Atmospheric Administration, 2015), which coincides with the energy measurement interval for most BPD buildings in New York. Thus the similarities in energy use between buildings in New York and San Francisco, as shown in **Figure 3**, are likely attributable to atypical weather in New York.

Because most of the buildings included in this analysis (for both BPD and CBECS) report only 1 year of energy use data, the current datasets do not necessarily capture typical regional and seasonal trends in energy use. Both datasets could benefit by incorporating multiple years of energy use data to reduce the impact of atypical weather on regional trends in energy use. Further, collecting monthly or daily interval energy use data may provide further insight into seasonal energy use patterns in heating/cooling loads.

6. CONCLUSION

Current empirical building data reveals limited insight into the factors that drive energy use in buildings. Our results show that, with the exception of floor area, the building characteristics collected today are largely uncorrelated with energy consumption. However, we exercise caution in drawing conclusions based on these results due to the limited size and depth of the current datasets. Although we control for building type, building size, and location, controlling for other characteristics or for different combinations of characteristics may reveal trends that are not evidenced in the current analysis.

A large, detailed, and high-quality dataset with information about building energy use and characteristics is needed to support a more detailed analysis. Unfortunately, accuracy and detail are key limitations of the self-reported data that constitutes much of the BPD, while size is costly to achieve when administering a detailed survey such as CBECS. As a result of these limitations, no such dataset currently exists.

Key limitations of the current datasets include: low resolution of energy use data, sparse detail in BPD records, and the small number of buildings in CBECS. Addressing these limitations in future data collection efforts may increase the range of applications these datasets can support. The data may be used to identify specific opportunities to improve energy performance; for example, the lack of correlation between energy use and operating hours could suggest that buildings can improve performance by reducing consumption during unoccupied hours. These results could also be used to target outreach efforts to specific subsets of the building stock likely to see the highest energy savings, such as buildings with relatively low operating hours.

Although more data can support more detailed analysis, the current datasets are capable of supporting broad analysis of buildings, for example to: (1) examine trends in energy use, (2) identify low-performing subsets of the building stock, or (3) evaluate the performance of buildings relative to their peers. Additional research is needed to determine whether the datasets are capable of supporting applications examining energy use drivers to: (1) identify inefficiencies in the building stock, (2)

predict energy savings due to building retrofits, or (3) confirm or refute our current understanding of building physics.

The correlations and distributions of values we observe in BPD are largely corroborated in CBECS, and vice versa. Thus despite concerns we and others raise with data quality in self-reported datasets, we did not find that data quality changed broad conclusions drawn from the data. We caution that newer, more detailed, or higher quality data may be warranted to support certain applications of the data.

Finally, we find the magnitude and slope of trends to be nearly identical across all peer groups and datasets. These similarities suggest that certain trends in energy use are shared among diverse subsets of the building stock. That the same trends are present in both a representative and a clustered dataset further supports the conclusion that energy use correlates are shared between buildings of different types and buildings in diverse climates.

DATA AVAILABILITY

The datasets analyzed for this study can be found in the Buildings Performance Database <http://bpd.lbl.gov> and

Commercial Buildings Energy Consumption Survey <https://www.eia.gov/consumption/commercial/data/2012/>.

AUTHOR CONTRIBUTIONS

MS defined the scope and direction of the work, and helped to prepare the manuscript for publication. LD did the analysis and drafted the manuscript.

FUNDING

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Paul Mathew, Richard Brown, Travis Walter, Claudine Custodio, Andrea Mercado, and Michael Berger of Lawrence Berkeley National Laboratory for their support on this project.

REFERENCES

- ASHRAE (2004). *ANSI/ASHRAE/IESNA STANDARD 90.1-2004*. Technical report, American Society of Heating Refrigerating and Air-Conditioning Engineers, Inc.
- Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychol. Methods* 2, 131–160. doi: 10.1037/1082-989X.2.2.131
- Brown, R. E., Walter, T., Dunn, L. N., Custodio, C. Y., Mathew, P. A., Cheifetz, D. M., et al. (2014). “Getting real with energy data: using the buildings performance database to support data-driven analysis and decision-making,” in *2014 ACEEE Summer Study on Energy Efficiency* (Pacific Grove, CA).
- Bryand, R. E., Katz, R. H., and Lazowska, E. D. (2008). *Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science, and Society*. Technical report, Computational Research Association.
- Custodio, C. Y., Walter, T., Dunn, L. N., Mercado, A., Brown, R. E., and Mathew, P. A. (2014). *Data Preparation Process for the Buildings Performance Database*. Technical Report LBNL-6724E, Lawrence Berkeley National Laboratory.
- Deru, M., Field, K., Studer, D., Benne, K., Griffith, B., Torcellini, P., et al. (2011). *U.S. Department of Energy Commercial Reference Building Models of the National Building Stock*. Technical Report NREL/TP-5500-46861, National Renewable Energy Laboratory.
- Hsu, D. (2014a). How much information disclosure of building energy performance is necessary? *Energy Policy* 64, 263–272. doi: 10.1016/j.enpol.2013.08.094
- Hsu, D. (2014b). Improving energy benchmarking with self-reported data. *Build. Res. Informat.* 42, 641–656. doi: 10.1080/09613218.2014.887612
- Hsu, D. (2015). Identifying key variables and interactions in statistical models of building energy consumption using regularization. *Energy*. 83, 144–155. doi: 10.1016/j.energy.2015.02.008
- Issenberg, S. (2012). *The Victory Lab: The Secret Science of Winning Campaigns*. New York, NY: Crown Publishing Group.
- Kahn, M. E., Kok, N., and Quigley, J. M. (2014). Carbon emissions from the commercial building sector: the role of climate, quality, and incentives. *J. Publ. Econ.* 113, 1–12. doi: 10.1016/j.jpubeco.2014.03.003
- Kavgic, M., Mavrogiani, A., Mumovic, D., Summerfield, A., Stevanovic, Z., and Djurovic-Petrovic, M. (2010). A review of bottom-up building stock models for energy consumption in the residential sector. *Build. Environ.* 45, 1683–1697. doi: 10.1016/j.buildenv.2010.01.021
- Kolter, J., and Ferreira, J. (2011). “A large-scale study of predicting and contextualizing building energy use,” in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence* (San Francisco, CA).
- Kontokosta, C. E. (2012a). *Local Law 84 Energy Benchmarking Data*. Technical report, New York City Mayor’s Office of Long-Term Planning and Sustainability.
- Kontokosta, C. E. (2012b). “Predicting building energy efficiency using New York City benchmarking data,” in *2012 ACEEE Summer Study on Energy Efficiency in Buildings* (Pacific Grove, CA).
- Mathew, P. A., Dunn, L. N., Sohn, M. D., Mercado, A., Custodio, C., and Walter, T. (2015). Big-data for building energy performance: lessons from assembling a very large national database of building energy use. *Appl. Energy* 140, 85–93. doi: 10.1016/j.apenergy.2014.11.042
- McKinsey & Co. (2009). Pathways to a low-carbon economy: version 2 of the global greenhouse gas abatement cost curve. Available online at: https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/Sustainability/cost%20curve%20PDFs/Pathways_lowcarbon_economy_Version2.ashx
- Mills, E., Friedman, H., Powell, T., Bourassa, N., Claridge, D., Haasl, T., et al. (2004). *The Cost-Effectiveness of Commercial-Buildings Commissioning: A Meta-Analysis of Energy and Non-energy Impacts in Existing Buildings and New Construction in the United States*. Technical report, Lawrence Berkeley National Laboratory. LBNL-56637.
- Pacala, S., and Socolow, R. (2004). Stabilization wedges: solving the climate problem for the next 50 years with current technologies. *Science* 305, 968–972. doi: 10.1126/science.1100103
- U.S. Department of Energy (2014). *EnergyPlus Documentation*. (accessed June 4, 2015).
- U.S. Department of Energy (2015a). *Building Performance Database*. (accessed March 9, 2014).
- U.S. Department of Energy (2015b). *Building Performance Database*. (accessed June 1, 2015).
- 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 Buildings Energy Consumption Survey (CBECS)*. (accessed November 14, 2014).
- U.S. National Oceanic and Atmospheric Administration (2013). *NOAA’s 1981-2010 U.S. Climate Normals*. (accessed March 16, 2015).
- U.S. National Oceanic and Atmospheric Administration (2014). *Integrated Surface Database*. (accessed (January 6, 2015).

- U.S. National Oceanic and Atmospheric Administration (2015). *Heating/Cooling Degree Day Monthly Summary: Climate Prediction Center*. (accessed May 12, 2015).
- Walter, T., Price, P. N., and Sohn, M. D. (2014). Uncertainty estimation improves energy measurement and verification procedures. *Appl. Energy* 130, 230–236. doi: 10.1016/j.apenergy.2014.05.030
- Williams, J. H., DeBenedictis, A., Ghanadan, R., Mahone, A., Moore, J., Morrow, W. R., et al. (2011). The technology path to deep greenhouse gas emissions cuts by 2050: the pivotal role of electricity. *Science* 335, 53–59. doi: 10.1126/science.1208365

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Sohn and Dunn. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.