Big Data Analytics In the Building Industry

BY MICHAEL A. BERGER; PAUL A. MATHEW, MEMBER ASHRAE; TRAVIS WALTER

Catalyzed by recent market, technology, and policy trends, energy data collection in the building industry is becoming more widespread. This wealth of information allows more data-driven decision-making by designers, commissioning agents, facilities staff, and energy service providers during the course of building design, operation and retrofit.

There is increased interest among the energy-efficiency practitioner community in using real-world data for "data-driven" analysis. Some tools focus on using detailed data for a given building,[1] while others use empirical data on many buildings, most notably the Energy Star Portfolio Manager tool.[2] The U.S. Department of Energy's Building Performance Database (BPD) is the largest publicly available data source for energy-related characteristics of commercial and residential buildings in the United States, collected from federal, state, and local governments, utilities, and private companies.

With over 870,000 records from commercial and residential buildings across the country, the BPD provides anonymized building energy use and asset data with analytical capabilities to help energy service providers, real estate owners and managers, policy makers, and energy consultants make decisions about energy efficiency and retrofit projects.[3] To date, the BPD has more than 10,000 users, the majority of them designers and energy service providers. The BPD's web interface[4] is free-to-use and allows extensive dataset management and customization.

This article examines some of the promises and perils of having large amounts of building data at the user's fingertips and how to use such data and statistical analysis tools effectively to support decision-making by energy professionals.

Promise: Benchmarking and Sanity Checking

The BPD offers a set of tools designed to assist energy professionals by supporting building benchmarking and

Michael A. Berger is a scientific engineering associate, Paul A. Mathew is a staff scientist and department head of whole building systems, and Travis Walter is a scientific engineering associate at Lawrence Berkeley National Laboratory in Berkeley, Calif.

sanity checking building model outputs.[5] Users can explore the available data across geographic regions, and compare physical and operational characteristics to gain a better understanding of market conditions and trends in energy performance. The interface allows users to define datasets, which are sets of buildings that share similar characteristics, through the selection of filters such as climate zone, facility type, floor area, and various building system characteristics such as lighting type, HVAC type, etc.

ASHRAE's *Procedures for Commercial Building Energy Audits*[6] and the emerging ASHRAE Standard 211 P, *Standard for Commercial Building Audits,* call for energy benchmarking to be done as one of the very first steps in an audit. Consider the use case of an energy auditor evaluating an office building in Seattle. The auditor can go to the BPD website and begin the benchmarking process by selecting a dataset that includes all office buildings in the state of Washington, and then selectively refine the dataset based on building characteristics and available data.

The BPD's Explore Histogram tool can be used to visually explore a range of important building characteristics, from floor area to Energy Star Rating to energy use intensity (EUI). The user can explore these histograms using the dashed vertical quartiles markers, as well as the hover-over feature, which gives more specific data for each histogram bar when selected. In *Figure 1a,* the user can see that the median EUI for offices in Washington is 163 kBtu/ft^2·yr (1851 MJ/m^2·yr), with an interquartile range of 125 to 213 kBtu/ft^2·yr (1420 to 2419 MJ/m^2·yr).*

The user can then further customize their dataset to enable more targeted comparisons. For example, the user can specify a dataset that includes only office buildings with floor areas greater than 100,000 ft^2 (9290 m^2) in Seattle. *Figure 1b* shows the histogram of source EUI for this more targeted dataset, which contains 242 buildings, and has a median EUI of 171 kBtu/ft^2·yr (1942 MJ/m^2·yr) and an interquartile range of 163 to 213 kBtu/ft^2·yr (1851 to 2419 MJ/m^2·yr).

Additionally, the auditor may want to document how EUI varies by vintage. The BPD's Explore Table tool allows users to generate a table that groups the active dataset by one characteristic, and provides statistics on another variable for each resulting subgroup. *Table 1* shows a sample of

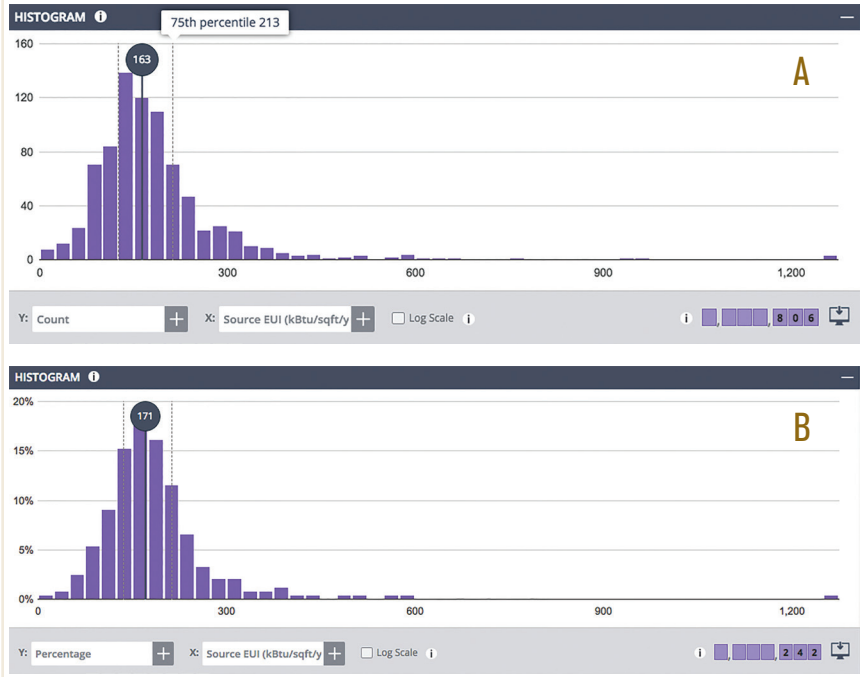

FIGURE 1 Histograms of source EUI for A) office buildings in Washington state; and B) large (>100k ft^2) office buildings in Seattle.

TABLE 1 Sample table of source EUI for large offices in Seattle, grouped by vintage.

YEAR BUILT	COUNT	MEAN (KBTU/FT^2·YR)	STANDARD DEVIATION KBTU/FT^2·YR)	MINIMUM (KBTU/FT^2·YR)	25TH PERCENTILE KBTU/FT^2·YR)
1960 to 1970	12	224.958	106.907	134.855	157.399
1970 to 1980	17	236.248	79.137	135.497	185.885
1980 to 1990	28	193.612	100.711	85.288	133.085
1990 to 2000	18	232.763	103.138	117.185	179.863
2000 to 2010	33	201.96	76.255	99.093	151.53

Note that additional data is available on the website, such as 50th Percentile, 75th Percentile and Maximum.

* Note that new data are continually being added to the BPD. The numbers given in this article reflect the data in the BPD at the time the article was written.

Dataset Comparison Methods

The BPD provides three methods to compare datasets: (1) visual comparison of histograms; (2) actuarial analysis; and (3) regression analysis. These methods are available on the Compare tab of the BPD web interface.

Visual Comparison of Histograms

The most basic form of comparative analysis is to present histograms of two datasets of interest overlaid together. This allows users to compare the spread and shape of the datasets in question, as well as quantify the difference in quartile values.

Actuarial Analysis Method

The actuarial approach used by the BPD is a method that represents the difference in a numerical characteristic between two datasets as a probability distribution. It does this by repeatedly and randomly sampling pairs of points from the two datasets and calculating the difference between the two.[9] These differences populate the resulting difference histogram. To improve computational efficiency, the sampling continues until the resulting distribution is identified as stable, which typically occurs after fewer than 20,000 pairwise comparisons.

This method was designed to be capable of comparing numerical values across any two datasets; therefore, its results are very sensitive to the underlying data. For example, if the input datasets have high statistical variability, the calculated distribution of differences will be more uncertain. This means a small dataset with a large number of outliers can skew the resulting distribution. Additionally, this method does not account for underlying differences in each pairwise comparison; therefore, if a user is looking for the difference in source EUI when switching from a furnace to a heat pump, the method does not automatically normalize for a building's climate zone. This means results are more reliable when the datasets being compared have fewer parameters that vary, making an understanding of the underlying datasets crucial.

Regression Analysis Method

The regression method provides a powerful comparison tool, using a selective multiple regression model to predict the distribution of differences between two datasets. This method accounts for differences in physical and operational characteristics, such as climate zone and facility type, and building equipment, such as cooling system and airflow controller, to predict energy savings due to building retrofits.[9] However, the stringent requirements of the model mean some datasets cannot be compared. For instance, when comparing differences in EUI due to a building system (e.g., lighting), if one of the two system types in question (e.g., LEDs vs. fluorescent T8s) composes less than 5% of the dataset, too little data exists for the regression model to accurately attribute the effect of that system type on EUI, and the regression analysis will return an error.

the exported table for large offices in Seattle, grouped by year built and analyzed by source EUI.

A similar investigation can be made for other factors that one would expect to impact energy use, including air control type and operating hours. The table tool can help flush out high-level differences in energy use across a dataset, highlight building systems that warrant further analysis, and be exported for inclusion in an audit's benchmarking documentation.

These data visualization tools, combined with the high level of dataset customization, give users unprecedented capabilities for benchmarking buildings against their peers and checking that building simulation estimates of performance fall within believable ranges, all based on real-world empirical data.

Peril: Potentially Nonrepresentative Data

The BPD team collects, cleanses,† and combines data from disparate sources from all over the country, essentially "crowdsourcing" data, without regard to representivity. Therefore, the BPD is not statistically representative of the national building stock. Additionally, the database is constantly growing, with new data sources added regularly, and old data sources updated when possible. Past work has investigated the representativeness of the BPD by comparing the database to the nationally representative Commercial Building Energy Consumption Survey (CBECS) and Residential Energy Consumption Survey (RECS), both of which are included in the BPD,‡ and found some regions and building types to be overrepresented in the BPD.[7] This, and the unprecedented level of granularity the BPD enables, means some

† Each dataset collected by the BPD team is analyzed and cleansed, which includes removing spurious numerical values and duplicate buildings across datasets. For more information on the BPD's data preparation, quality control, site-source EUI conversion factors and more, see Custodio, et al.[8]

‡ Additional public datasets in the BPD include benchmarking ordinance data from cities including New York, Boston, San Francisco, and others.

peer comparison groups will be data rich, while others may have no buildings at all. Some peer groups may have large quantities of data about building systems, while others may only have energy data. There is also the possibility of selection bias in the underlying data, as those organizations that collect and contribute building energy data to the BPD may be more likely to have pursued energy efficiency and benchmarking.

Users should keep this in mind when exploring the database, as benchmarking a building against its peers in the BPD does not necessarily represent the exact standing of that building against the national population of buildings. As such, the BPD is not appropriate for analyses that are critically dependent on a statistical sample, e.g., national or regional estimates of total energy use.

FIGURE 2 Comparisons of Datasets 1a and 1b, by electric EUI. A) Overlaid histograms of Datasets 1a (purple) and 1b (yellow). B) Difference histogram calculated with the actuarial analysis method. C) Difference histogram calculated with the regression analysis method.

Promise: Leveraging Big Data to Analyze Technology Impacts

In addition to peer group benchmarking, large datasets such as the BPD enable data-based analysis of the energy impacts of changing building characteristics or systems. The BPD offers three methods of comparative analysis (see "Dataset Comparison Methods" for more details). The following sections will present examples of such analysis on both residential and commercial datasets.

Example 1: Single-Family Homes in Ohio

Our first example dataset contains single-family homes in Ohio. This dataset has 2,245 buildings in the BPD, and will be called Dataset 1.

Comparing No Cooling to Central Air Conditioning. Let us explore the use case of an energy service provider who seeks to understand the impact of central air conditioning on home electricity use to better market their services to customers. The user can create two datasets and directly compare the electricity use of customers from Dataset 1 with air conditioning to those without it. Dataset 1a contains only buildings from Dataset 1 that have "No Cooling" as their cooling type; this dataset has 309 homes in it. Similarly, Dataset 1b only has homes with "Central Air Conditioning" (CAC) as their cooling type; this dataset has 1,287 homes in it.

Figure 2a shows that the median electric EUI for homes without cooling is 10 kBtu/ft^2·yr (114 MJ/m^2·yr) and the median for homes with CAC is 17 kBtu/ft^2·yr (193 MJ/m^2·yr), which represents an increase of 70%. This is consistent with the results

from the actuarial analysis, which shows CAC to increase EUI by 63%; however, the regression analysis results suggest a smaller increase in EUI of only 43%. The regression analysis attempts to account for other building characteristics, including heating type, number of occupants, window types, etc., and would suggest that underlying factors are causing the actuarial and median differences to overestimate the impact of CAC.

Comparing Single-Pane to Double-Pane Windows.

Similarly, consider an energy service professional who is interested in empirical evidence of the impact of double-pane windows on source EUI. To do this, the user could create two datasets. Dataset 1c contains only buildings from Dataset 1 that have "Single-Pane" window glass layers; this dataset has 79 homes in it. Dataset 1d only has homes with "Double-Pane" window glass layers; this dataset has 433 homes in it.

Figure 3a shows that the median EUI for homes with single-pane windows is 90 kBtu/ft^2·yr (1022 MJ/m^2·yr), and the median for homes with double-pane windows is 89 kBtu/ft^2·yr (1011 MJ/m^2·yr), which represents a decrease of 1%. The result from the actuarial analysis, shown in *Figure 3b,* estimates double-pane windows to increase EUI by 3%. Lastly, as shown in *Figure 3c,* the regression analysis calculates a 2% decrease in EUI when switching from single-pane to double-pane windows. These results demonstrate the uncertainty, or "noise," that can arise from empirical data analysis. The energy service professional should interpret these results as inconclusive of whether or not double-pane windows significantly reduce energy use in Ohio homes at the portfolio level.

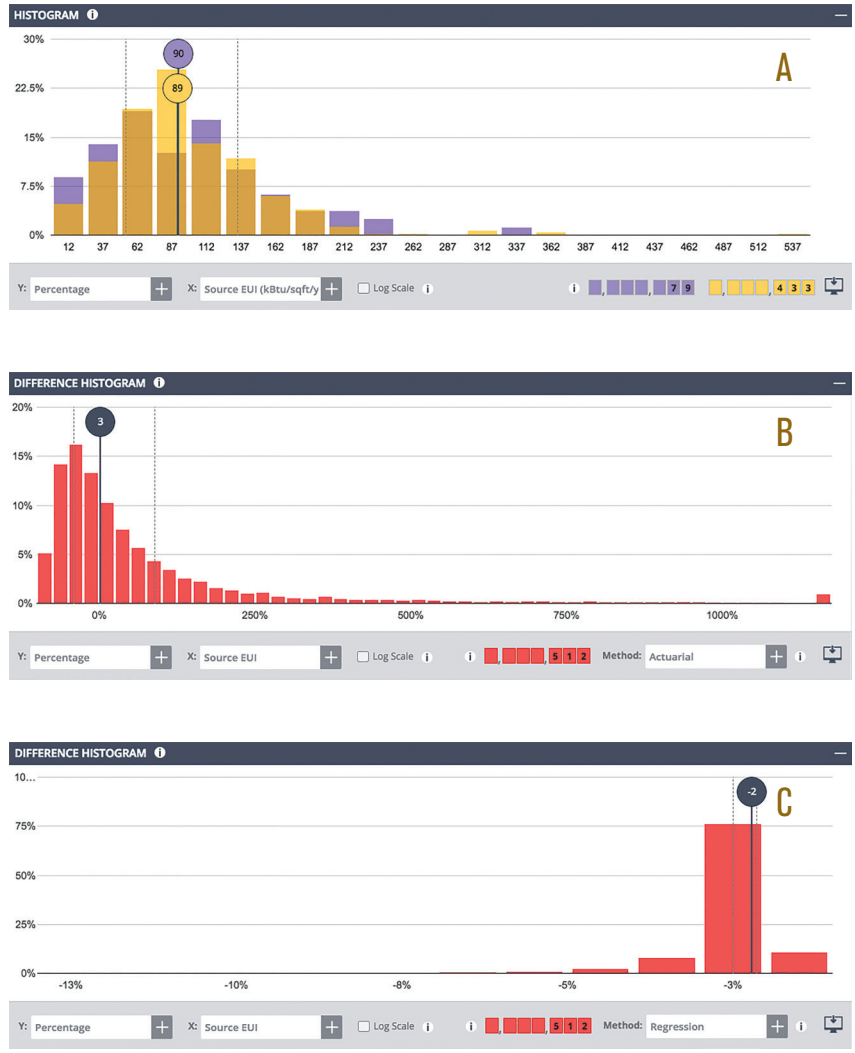

FIGURE 3 Comparisons of datasets 1c and 1d, by source EUI. A) Overlaid histograms of Datasets 1c (purple) and 1d (yellow). B) Difference histogram calculated with the actuarial analysis method. C) Difference histogram calculated with the regression analysis method.

Example 2: California Offices

Similar analyses can be done for commercial buildings. Let us use a dataset for office buildings in California, which we will call Dataset 2. Dataset 2 has 3,448 buildings in it.

Comparing CAV to VAV Systems. Consider a facilities portfolio manager that wants to better understand the impact of a retrofit from constant air volume (CAV) HVAC control systems to more efficient variable air volume (VAV) controls. The user can define Datasets 2a and 2b to include only buildings with CAV and VAV airflow control, respectively. Dataset 2a has 71 buildings, and Dataset 2b has 104 buildings.

Advertisement formerly in this space.

Figure 4 shows consistent and intuitive results, with both the visual comparison and actuarial method showing a 22% decrease in source EUI when switching from CAV to VAV, and the regression method showing a 27% decrease in source EUI.

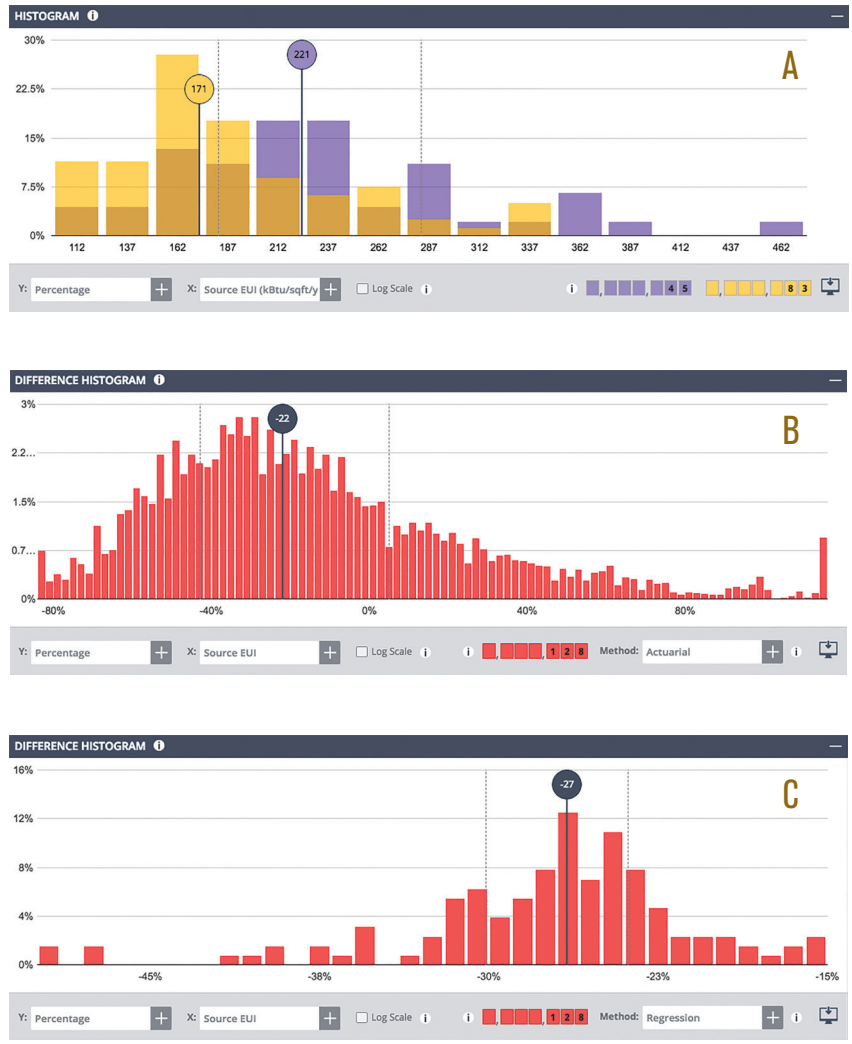

FIGURE 4 Comparisons of dataset 2a and 2b, by source EUI. A) Overlaid histograms of Datasets 2a (purple) and 2b (yellow). B) Difference histogram calculated with the actuarial analysis method. C) Difference histogram calculated with the regression analysis method.

Peril: Misinterpretation of Comparison Results

Tools such as the BPD put powerful analytical methods and large amounts of building data in the hands of the user; however, the complexity and depth of these tools can lead the user astray. The key to enabling users to craft representative datasets that produce robust and reliable results is an understanding of building physics principles and basic data science. This is particularly true when comparing datasets.

Take the example of a user who wishes to understand the impact of switching their retail buildings, all of which are in Climate Zone 3B, from old T12 fluorescent lighting to newer, more efficient T8s. The user can quickly make two peer groups, the first comprised of all retail buildings in Climate Zone 3B with T12 lighting, and the second identical except with T8 lighting. Comparing each dataset's histogram medians for source EUI shows that buildings with T8s use 58% more energy than buildings with T12s. The actuarial method of comparison returns a similar value of 60% more energy. Finally, the regression method returns a value of 16% more energy when switching to T8s, which is still counterintuitive, but much closer to what one expects.

To understand, and avoid, issues such as these, the user should always explore the similarities and differences between their two datasets to gain an understanding of what amount and quality of data comprises each dataset. The BPD does not have complete data on all buildings' systems and characteristics, which makes it crucial to understand exactly what data is being used to estimate differences.

For retail buildings, data fields worth investigating include operating hours, floor area, and number of people, as major differences in these datasets may explain the nonintuitive results. *Figure 5* shows how the underlying data for retail buildings with T8 lighting has, on average, 4.7 times the square footage, 54% more people, and 44% longer operating hours than retail buildings with T12 lighting. These major differences between datasets are, to some degree, being accounted for by the regression model, but not by visual comparison of

histograms or the actuarial comparison method.

Conclusions

While the BPD is the largest publicly available dataset of its kind and leverages powerful statistical analysis methods, a sound understanding of the underlying data and building science is necessary to avoid misinterpreting results. A general caution for empirical data analysis is that underlying differences in datasets, as well as missing data, can mislead the unaware user. These pitfalls can be overcome with careful investigation of the available data and a foundational understanding of the principles of building energy use.

The Building Performance Database gives users a set of tools for performing real-world, data-based exploration and analyses on highly granular, and highly customizable, datasets. These tools can be used by building energy practitioners to benchmark buildings, check the validity of simulations and model outputs against real world data, and compare datasets using statistical methods to better understand technology impacts on building energy use.

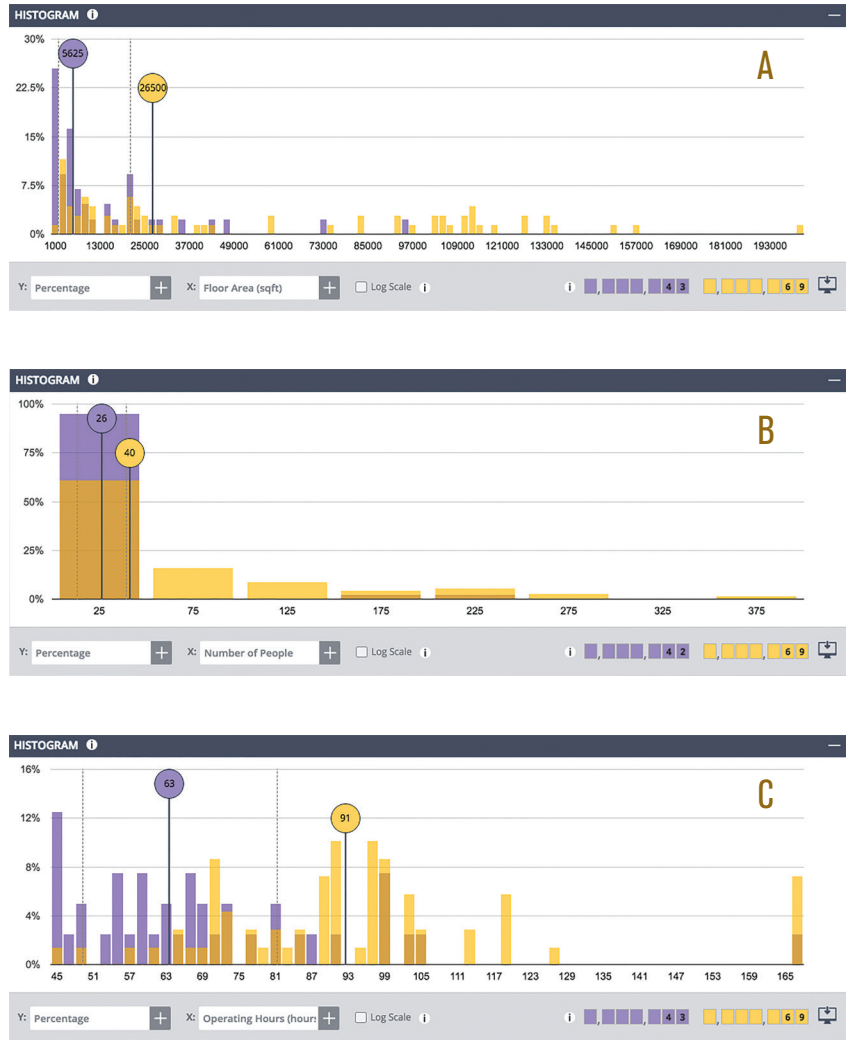

FIGURE 5 Overlaid histograms of T12 (purple) and T8 (yellow) lighting for retail buildings in California for A) floor area, B) number of people, and C) operating hours per week.

Acknowledgments

The authors gratefully acknowledge the many data contributors to the BPD, and energy professionals who shared how they use the tool. The BPD is sponsored by the U.S. Department of Energy's Building Technologies Office.

References

1. Granderson, J., M.A. Piette, G. Ghatikar. 2011. "Building energy information systems: user case studies." *Energy Efficiency* 4(1):17–30.

2. Energy Star. "Portfolio Manager." http://tinyurl.com/jh7kgvf.

3. U.S. DOE. 2015. "It's All About the Data: How the Building Performance Database is Informing Decisions on Energy Efficiency." U.S. Department of Energy. http://tinyurl.com/jh7kgvf.

4. LBNL. "Building Performance Database." Lawrence Berkeley National Laboratory. https://bpd.lbl.gov/.

5. Brown, R.E., et al. 2014. "Getting real with energy data: using the buildings performance database to support data-driven analyses and decision-making." ACEEE Summer Study on Energy Efficiency in Buildings. http://tinyurl.com/j86oy54.

6. ASHRAE. 2011. *Procedures for Commercial Building Energy Audits, Second Edition*. Atlanta: ASHRAE.

7. Mathew, P.A., et al. 2015. "Big-data for building energy performance: Lessons from assembling a very large national database of building energy use." *Applied Energy* 140:85–93.

8. Custodio, C.Y., et al. 2015. "Data Preparation Process for the Buildings Performance Database." LBNL-6724E. Lawrence Berkeley National Laboratory. http://tinyurl.com/jugn74w.

9. LBNL. "BPD API Documentation." Lawrence Berkeley National Laboratory. http://tinyurl.com/hyr6rwl. ■