BlocPower User Guide

MASSIVE DATA INSTITUTE

Version 1.1, June 2023
Environmental Impact Data Collaborative,
Massive Data Institute

---

1. Introduction

A majority of total U.S. emissions stem from four economic sectors: buildings, electricity, industry, and transportation. Reductions in all four areas are essential to reach the Paris agreement's Nationally Determined Contribution (NDC) of cutting emissions to 50% below 2005 levels by 2030. Princeton's Net Zero Initiative (Larson et al, 2021) sets out six pillars of economy-wide decarbonization, leading with improving end-use energy efficiency and electrification for buildings. Increasing electrification by itself is projected to reduce final energy use due to intrinsically more efficient technology like heat pumps.

High-resolution data on building equipment and energy consumption is essential to plan and measure the impact of decarbonization interventions. Existing data sources for this purpose have limited coverage. The Energy Information Administration's (EIA) Residential Energy Consumption Survey (RECS) is a key product, but it is a sample-based survey released infrequently and is not representative below the state level. Other options include proprietary providers of tax assessment data (like CoreLogic) or restricted data from energy and utility companies with personally identifiable information (PII).

The Environmental Impact Data Collaborative (EIDC) has partnered with BlocPower - a Brooklyn-based climate technology company that uses building-level energy use data to guide community decarbonization projects. As a key output of this partnership, EIDC has worked with

BlocPower to add a large building-level dataset to our data lake, containing energy equipment and consumption data for more than 121 million buildings. The data is partially sourced from tax assessment records, which provide data on building system types and attributes like built year and area. This data then serves as inputs to an Automatic Building Energy Modeling (AutoBEM) developed by Oak Ridge National Laboratory, to generate modeled estimates of building energy use.[1] Cloud-based building-level data pipelines are used by BlocPower to deploy their software-as-a-service (SaaS) solution - BlocMaps - to provide building decarbonization insights at city and local levels.[2]

The following sections provide a walkthrough of the data and its possible use cases. Section 2 provides an overview of the data availability and completeness, and then conducts basic validation checks with other peer-reviewed building and energy use datasets. Section 3 begins with a showcase of commonly used exploratory data analysis and visualizations of the key variables. Then, it explores methods to combine and analyze this dataset with existing administrative datasets (CEJST and LIHEAP) to derive insights, using methods like spatial clustering and regression modeling.

2. Data Overview

In this section we discuss the data availability and conduct several validation exercises.

2.1 Data availability

The data is provided through the following datasets on EIDC's Redivis data repository:

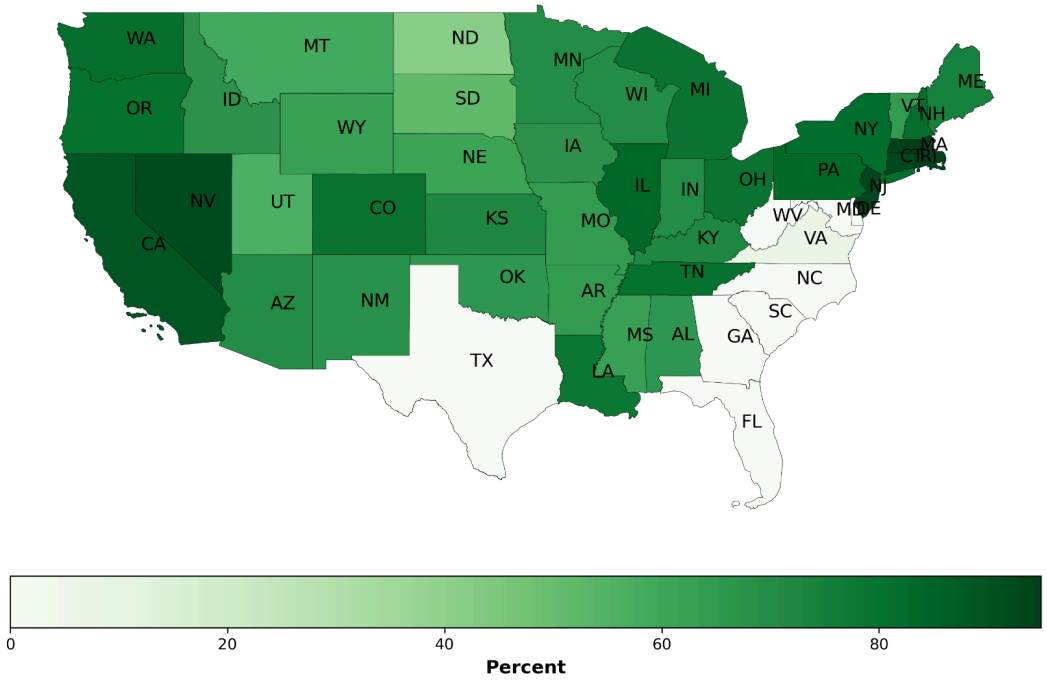
1. BlocPower_Core: Individual building level data for 51 states and territories (see appendix A1 for a full list of variables).

2. BlocPower_Summary**:** Aggregated zip code and county-level statistics.

[1] "BlocPower collects data from over 100 million buildings from external sources, such as the Department of Energy's National Laboratories. These laboratories store their data using intermediate data format files, which require BlocPower to use EnergyPlus to process and render simulations of individual buildings."
[2] See Appendix A2.

Figure 1 shows the data availability by states, with darker colored states having more complete data. The lighter colored states have nearly all values missing for address, and the key building attributes like heating and cooling system types. This arises because the building attributes data could not be matched to an address.

Fig 1. State-wise distribution of buildings with complete data (%)

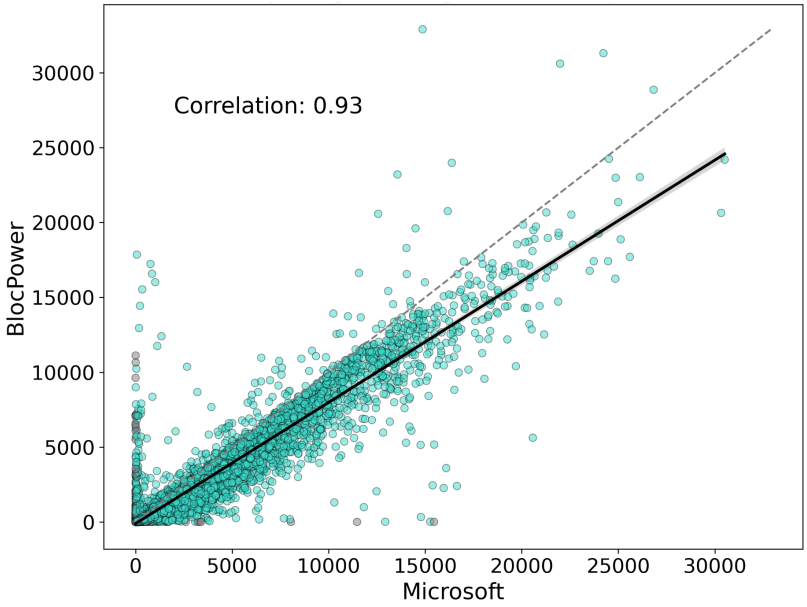


Note: Figure plots percent of non-missing values in the 'address' column for each state.

2.2 Data Validation: Comparing BlocPower and Microsoft Building Data

A simple validation of the dataset is provided by comparing aggregated statistics of buildings at the zipcode with other data sources. One such independent source of data on buildings is provided by Microsoft's Building Footprints dataset. This contains 129,591,852 building footprint polygon geometries for the US mainland in GeoJSON format, derived from computer vision algorithms on satellite imagery (updated till 2020). Figure 2 plots compares the number of buildings in the Microsoft and BlocPower datasets, aggregated to the zipcode level.

Fig 2. Scatterplot, buildings in each zip code (Microsoft vs BlocPower)



Note: Zip Code level building counts for six states - New York, Pennsylvania, Minnesota, Tennessee, Washington, Massachusetts, Colorado and Arizona. Zipcodes with 10 or fewer buildings colored in gray.

It shows a high correlation of 0.92, and the best-fit regression line is slightly below the 45-degree line, denoted by the dashed line. Note that expecting perfect equality would not be appropriate as the datasets were generated at different points in time. Some zip codes in each data set have fewer than 10 buildings; these are likely due to missing data. Note that the Microsoft computer vision algorithms has a precision of 98.5 % and recall of 92.4% - implying that building counts will contain misclassifications.[3]

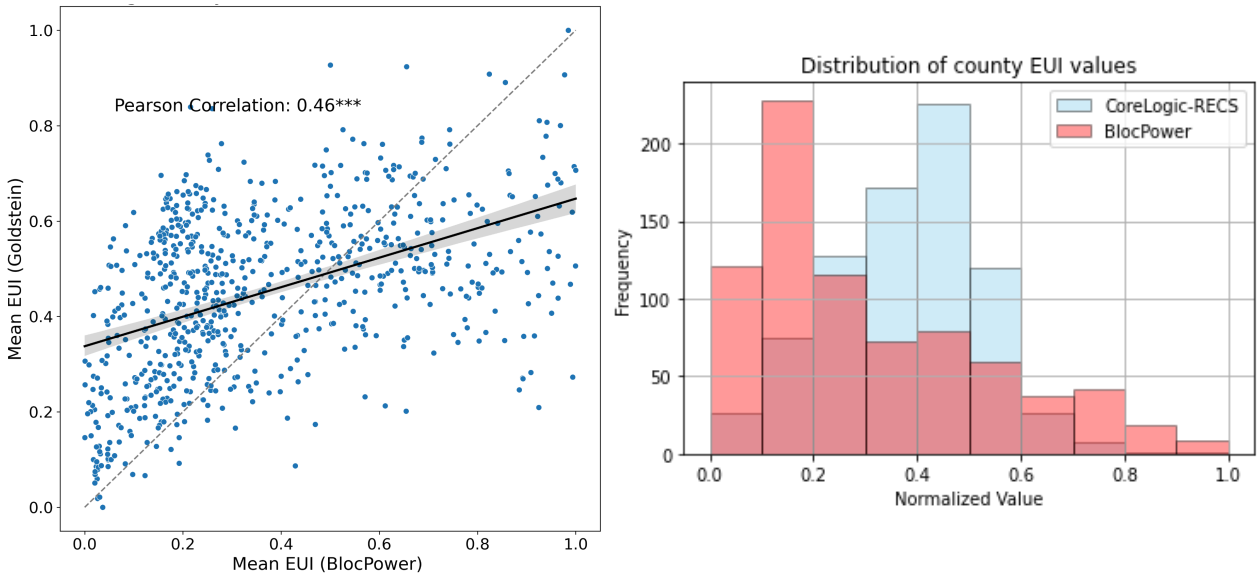
2.2 Data validation: Comparing Energy Use Intensity of zipcodes with Goldstein (2022)¶

Goldstein et al (2022) used proprietary data from CoreLogic for around 90 million buildings to estimate energy use intensity (EUI). Using the primary heating fuel type and cooling type (AC) of the building from CoreLogic, they generate EUI estimates with the regression models trained on the Residential Energy Consumption Survey (RECS) 2015. Their findings are covered in a later section. They also make available zip code level energy use intensity in slightly different

[3] Precision of 98.5% means of 100 entities classified as buildings, 1.5 are false positives..

units (kWH/square meters). We calculate the weighted average of EUI for each county, weighted by zipcode population, and compare it with county level EUI estimates generated from the BlocPower data. The distribution of normalized EUI values by county is also plotted for both datasets, by two overlapping histograms.

Fig 3. Comparing county-level average EUI across BlocPower and Goldstein datasets



Note: The subplot on the left shows normalized mean county EUI in the BlocPower dataset on the X-axis, and normalized mean county EUI data used in Goldstein et al (2022) for 741 counties. The dotted line is the 45-degree line, solid line is the OLS line. Subplot on the right plots overlapping histograms. Note that the BlocPower dataset has not been cleaned to match the sample selected by Goldstein (restricting to Single Family only, and minimum and maximum area thresholds).

The scatterplot clearly shows a high correlation coefficient of 0.46, though the points are quite dispersed from the 45-degree line of equality. This shows that relative position of counties in terms of EUI are positively correlated. However, the histogram figure suggests that the distributions of EUI across counties might be different. A Kolmogorov-Smirnov (KS) test confirms that the distributions are different, with a p-value<0.001.

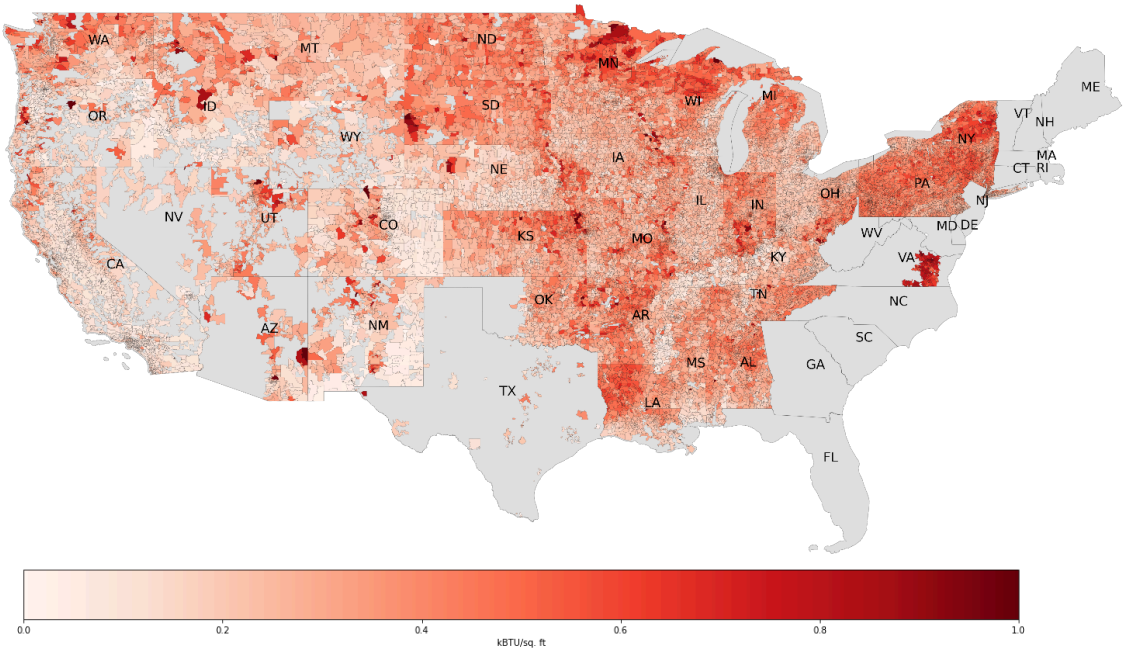
3. Exploratory Data Analysis

To provide a sense of how the BlocPower data can be used, we present several different preliminary analyses of the data, showing how the data can be used to assess energy use attributes for precise spatial areas.

3.1 Median building EUI across space

The following maps depicts the median energy use intensity across all buildings in each zip code. As we see, data is missing for certain zip codes and states, which appear in grey. There seems to be considerable variation in EUI, and we see that the highest EUI values are in the Northeast.

Fig 4. Choropleth map - zipcode level energy use intensity



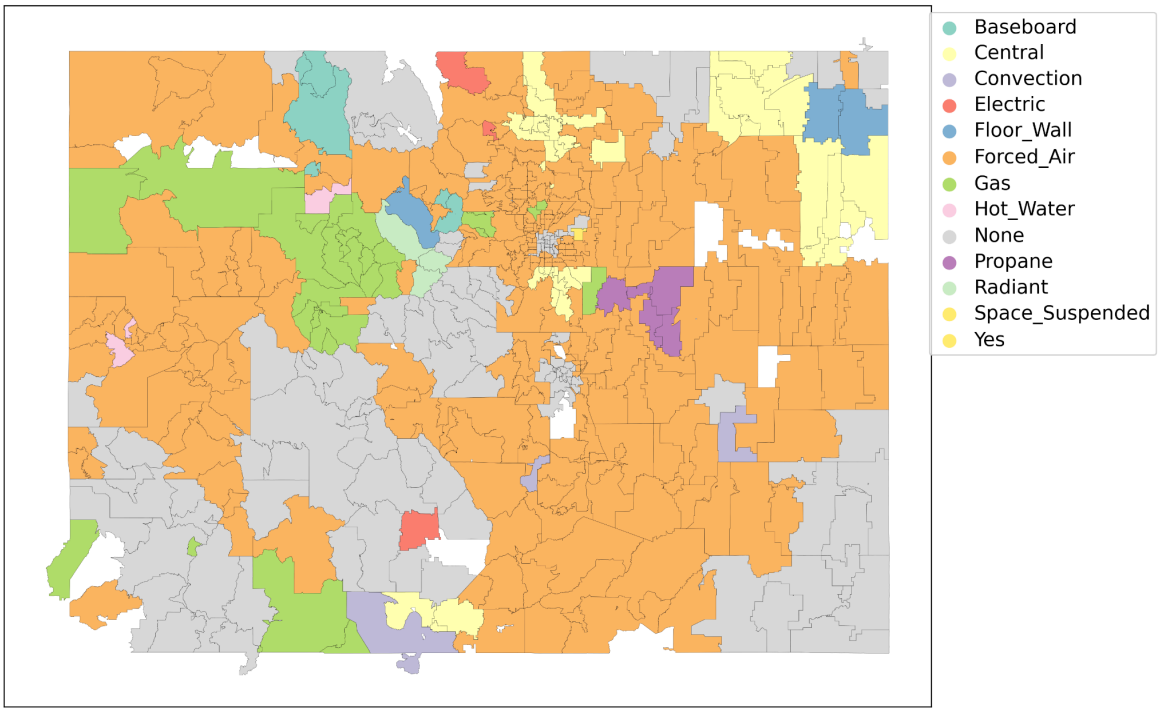
Note: Values normalized within states. Energy Use Intensity is expressed in units of kBTU/sq.ft

3.2 Top heating/cooling system types by zip code

We calculate the most common heating system type, by choosing the heating system with the highest share of buildings in a zip code. The map for Colorado is given below. This data allows

identification of zip codes with high prevalence of particular kinds of heating systems, say, baseboard heating.

Fig 5. Choropleth map - Most prevalent heating system type per zip code, Colorado

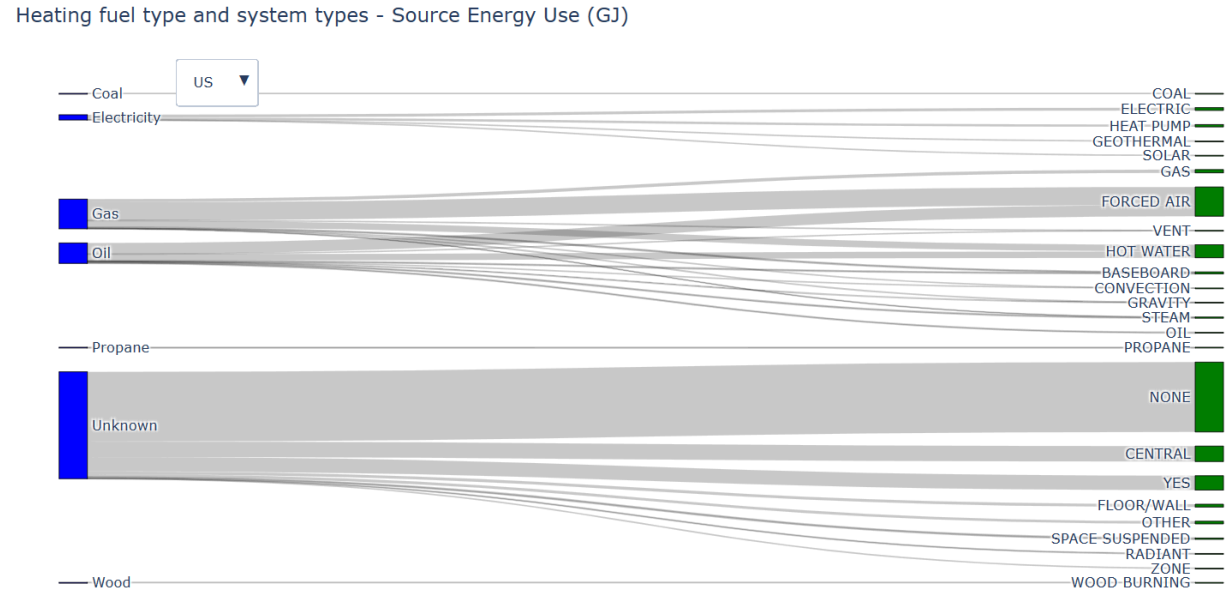


Note: Most prevalent heating system type selected as the most frequent value of column 'heating_system_type' in the zipcode, including unknown values.

3.3 Energy Use by heating technology

The uniquely granular nature of the data can be leveraged to create detailed energy source and destination diagrams for heating fuel/system types. The interactive Sankey plot below allows you to examine the energy breakdown (total site energy use in GJ) for different states.

Fig 6. Sankey plot - total site energy use (GJ) for heating fuel and system types, United States



Note: This also shows the amount of energy attributed to buildings which do not use any fuel/system for heating, marked by unknown and none. Note that the total energy use in those buildings can't be attributed solely to heating.

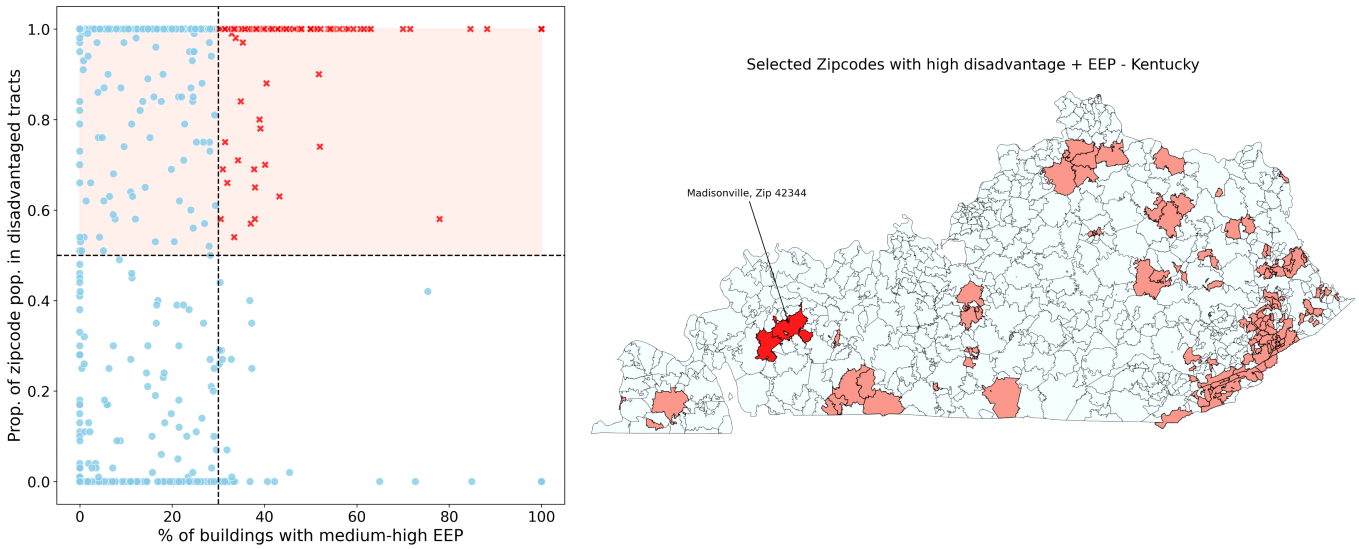
3.4. Exploring energy efficiency and environmental justice

This section sketches a simple framework to prioritize zipcodes for energy efficiency investments. This is inspired in part by Helleno et al's (2022) approach of combining energy efficiency with social equity priorities, as well as Tong et al (2021's) quadrant analysis diagram. To do this, we overlay the Climate and Economic Justice Screening Tool (CEJST), after aggregating the disadvantage classification to the zipcode level.[4] We generate the proportion of population in each zip code which lives in a tract classified as disadvantaged in the CEJST. The other variable in the framework will be the percent of buildings in each zipcode classified as having either high or medium energy efficiency potential (EEP) by BlocPower.

[4] We employ a raster-based method to map CEJST data from census tracts to ZIP Code Tabulation Areas (ZCTAs). Since these areas can overlap, we use a population density raster layer to proportionately map the population within these overlapping areas. This process produces a "crosswalking weight" that can be then used to reweight the percentage of disadvantaged populations from census tracts to matched ZCTA. This ensures a more accurate representation of the demographics across different geographic scales.

We select the state of Kentucky to illustrate this analysis. Figure 7 plots the share of disadvantaged population on the Y-Axis and percent of buildings with medium-high EEP on the X-Axis. Partly, this graph is inspired, but different, from the quadrant analysis performed in Tong et al (2021). Depending on the threshold of each variable chosen, a different set of zip codes fall in the 'optimal zone', the top right quadrant. Suppose we select 3 adjacent zip codes with both high disadvantage and EEP, located in the top right quadrant: 42431, 42408, 42344 - in Madisonville, Kentucky.

Fig 7. Quadrant analysis - zipcodes with high disadvantage and energy efficiency potential, Kentucky

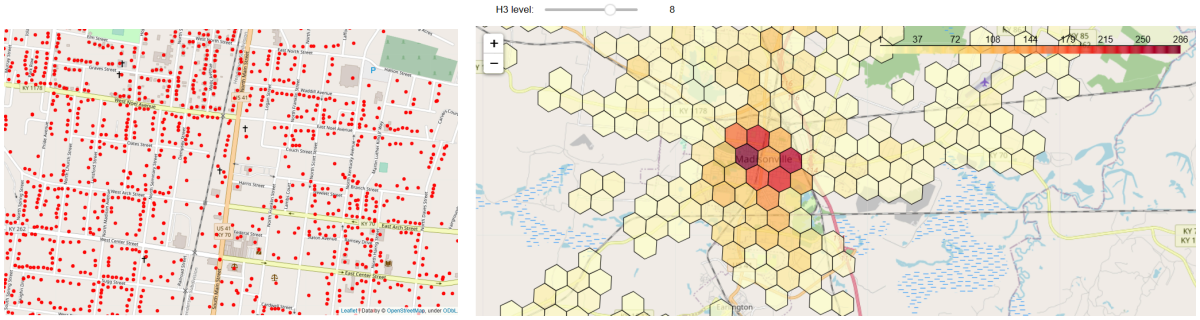


Note: Figure on the left shows quadrant analysis, with percent of buildings with medium-high energy efficiency potential on X-Axis, and prop. of zipcode population living in tracts classified as 'disadvantaged' in CEJST. Figure on the right shows the selected zipcodes in the upper right quadrant in the first figure, with the manually selected zipcodes colored bright red and annotated.

Once a set of zipcodes has been selected , we can visualize them on a map. It is interesting to note that several chosen zipcodes are part of local spatial clusters. Retrieve the original individual building level data for these three zipcodes from EIDC and geocode the building addresses to coordinates. Using the open source H3 hierarchical hexagonal spatial indexing method to plot spatial clusters of buildings with high EUI (defined as EUI above the zipcode's

median). Note that the interactive plot allows us to change the zoom (H3 level) parameter. (see H3 indexing)

Fig 8. Spatial clustering of high-EUI buildings in Madisonville, KY with H3 indexing



Note: Figure on the left maps the geocoded buildings as points in the chosen zipcodes in Madisonville, KY. Figure on right uses H3 indexing to bucket points with higher than average EUI into hexagons. Darker red means more buildings clustered in that location.

3.5. Regression analysis

We merge census demographic data at the zipcode level to the above dataset, along with aggregated statistics of the US Department of Health and Human Services (HHS) Low Income Home Energy Assistance Program (LIHEAP).[5] This results in four new variables:

- **Population per building**: population of zip code divided by total number of buildings
- **Percent white**: percentage of white from ACS 5-year sample from 2020
- **Percent of LIHEAP recipients:** LIHEAP recipients as percent of zip code population
- **Average LIHEAP:** Average LIHEAP amount paid

[5] LIHEAP, or the Low Income Home Energy Assistance Program, provides financial assistance for low income families to cover the costs of heating, cooling, and weatherizing their homes. The LIHEAP variables used in this dataset describe both the percentage of the population that received this energy assistance and the average amount of funding received given to each household for each zip code across the county. The total LIHEAP funding variable is the summation of several variables pertaining to funding directed specifically for heating, cooling, and crisis costs while the percentage of recipients is the proportion of LIHEAP recipients to a zip code's total population. These variables were aggregated by zip code via household-level data on LIHEAP recipients in order to ensure the anonymity of recipients and their personal information.

Using the above as independent variables, the following table shows the results of an OLS regression with median energy use intensity (EUI) of zip code as the dependent variable.

Table 1: OLS Regression results - full sample

	Dependent variable: Median EUI		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
Pop. per building	-0.003***	0.002***	0.001***
	(-0.001)	(0.00003)	(-0.00001)
Percent White	1.600***	0.528***	0.879***
	(-0.127)	(-0.041)	(-0.081)
Average LIHEAP	0.054***	0.126***	
	(-0.004)	(-0.008)	
Percent LIHEAP recipients			4.827***
			(-0.913)
Constant	58.492***	152.541***	193.571***
	(12.614)	(16.277)	(15.975)
State Fixed effects	No	Yes	Yes
Observations	19,399	19,399	19,399
R^2	0.093	0.708	0.694
Adjusted R^2	0.093	0.708	0.694
Residual Std. Error	126.429 (df = 19395)	71.760 (df = 19362)	73.475 (df = 19362)
F Statistic	665.669*** (df = 3; 19395)	1,306.652*** (df = 36; 19362)	1,221.542*** (df = 36; 19362)
Note:			*p<0.1;* ***p<0.05;*** p<0.01
			Robust HC standard errors

Average LIHEAP and percentage LIHEAP recipients are correlated with each other, so they are not included in the same model. Model 1 does not contain state fixed effects. Model 2 and 3 contain state-level fixed effects. This accounts for unobserved state-level characteristics that could affect median EUI, such as building codes, energy tariffs and climatic conditions. Zip Codes with higher shares of white population, higher shares of LIHEAP recipients, and higher average LIHEAP payments have higher median EUI, on average.[6] This is similar to Goldstein et al's (2022) findings, where a area's racial composition had a statistically significant role in determining energy use and emissions, with majority White areas having higher energy use.

[6] The relationship and coefficients change significantly, depending on the state. Observations within states (zipcodes) are likely not independent. To model this dependence, use a spatial lag error term and/or cluster standard errors by state.

Tong et al (2021) find an inverse relationship for St. Paul, MN and Tallahassee, FL; where rising fraction of racial minorities in a block was associated with higher EUI.

References

Larson, E., Greig, C., Jenkins, J., Mayfield, E., Pascale, A., Zhang, C., ... & Swan, A. (2020). Net-Zero America: Potential Pathways, Infrastructure, and Impacts, interim report, Princeton University, Princeton, NJ.

Tong, K., Ramaswami, A., Xu, C., Feiock, R., Schmitz, P., & Ohlsen, M. (2021). Measuring social equity in urban energy use and interventions using fine-scale data. Proceedings of the National Academy of Sciences, 118(24), e2023554118. https://doi.org/10.1073/pnas.2023554118

Goldstein, B., Reames, T. G., & Newell, J. P. (2022). Racial inequity in household energy efficiency and carbon emissions in the United States: An emissions paradox. Energy Research & Social Science, 84, 102365. https://doi.org/10.1016/j.erss.2021.102365

Heleno, M., Sigrin, B., Popovich, N., Heeter, J., Figueroa, A. J., Reiner, M., & Reames, T. (2022). Optimizing equity in energy policy interventions: A quantitative decision-support framework for energy justice. Applied Energy, 325, 119771. https://doi.org/10.1016/j.apenergy.2022.119771

Appendix

This section contains the full list of variables available in the current dataset. It also contains links and screenshots of BlocPower's software-as-a-service (SaaS) platform for building retrofits, based on very similar variables.

A1. List of variables in BlocPower Core dataset

This section lists the 16 variables available in the current dataset and their data types.

Name	Type	Label
building_id	integer	Building ID
state	string	State
county	string	County
city	string	City
zip	string	Zip code
address	string	Address
area_sq_ft	integer	Building area in square feet
year_built	float	Year built
building_type	string	Building type
cooling_system_type	string	Type of cooling system
heating_system_type	string	Type of heating system
heating_fuel_type	string	Type of heating fuel
energy_use_intensity	float	EUI per square foot in 1000 British Thermal Units
energy_efficiency_potential	string	Level of Building's Potential Energy Efficiency (catgorical)
total_site_energy_GJ	float	Total site energy in Gigajoules
total_source_energy_GJ	float	Total source energy in Gigajoules

A2. BlocMaps decarbonization platform screenshot

This is a screenshot of BlocPower's proprietary BlocMaps platform. It is a decision support system to guide building upgrades, which generates projections of cost and emissions reduction for each building. This is based on data very similar in structure to the current one. For more information, click here.