Sector-Specific Methodologies for Subnational Energy Modeling

Ookie Ma,[1] Ricardo Oliveira,[2] Evan Rosenlieb,[2] and Megan Day[2]

1 U.S. Department of Energy
2 National Renewable Energy Laboratory

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Technical Report
NREL/TP-7A40-72748
April 2019

Sector-Specific Methodologies for Subnational Energy Modeling

Ookie Ma,[1] Ricardo Oliveira,[2] Evan Rosenlieb,[2] and Megan Day[2]

1 U.S. Department of Energy
2 National Renewable Energy Laboratory

Suggested Citation
Ma, Ookie, Ricardo Oliveira, Evan Rosenlieb, and Megan Day. 2019. *Sector-Specific Methodologies for Subnational Energy Modeling*. Golden, CO: National Renewable Energy Laboratory. NREL/TP-7A40-72748. https://www.nrel.gov/docs/fy19osti/72748.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Technical Report
NREL/TP-7A40-72748
April 2019

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Strategic Priorities and Impact Analysis Office. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

NREL prints on paper that contains recycled content.

Acknowledgments

The authors thank the following people for their contributions to the development of these methodologies and for providing substantive comments that informed the analysis and improved the structure and content of the report: Kevin Kurkul and Tom Cyrs of the World Resources Institute, Derik Broekhoff of the Stockholm Environment Institute, Galen Maclaurin, Michael Gleason, Brian Bush, and Gian Porro. We also gratefully acknowledge funding for this work from the Office of Strategic Programs of the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy under contract number 28759.

List of Acronyms

ACS5	5-year American Community Survey
CBECS	Commercial Building Energy Consumption Survey
CoStar	CoStar Realty Information, Inc.
DOE	U.S. Department of Energy
EPA	U.S. Environmental Protection Agency
GBS	General Building Stock
GHG	greenhouse gas
GHGRP	Greenhouse Gas Reporting Program
HIFLD	Homeland Security Infrastructure Foundation-Level Data
IET	Industrial Energy Tool
IPF	iterative proportional fitting
NAICS	North American Industrial Classification System
NREL	National Renewable Energy Laboratory
PBAplus	Principal Building Activity Plus
RII	Regional Indicators Initiative

Definitions

Census blocks are the smallest geographic areas used by the Census Bureau and are bounded by visible features, such as streets, roads, streams, and railroad tracks. Census blocks nest within all other census geographic entities including cities, townships, and counties.

Census tracts are small, statistical subdivisions of a county or county equivalent and generally have a population size between 1,200 and 8,000 people, with an optimum size of 4,000 people.[1]

Minor civil divisions are the primary governmental divisions of a county and include townships and towns. The minor civil divisions in 12 states, including New England states, New Jersey, New York, and Pennsylvania, also serve as general-purpose local governments that can perform the same governmental functions as incorporated places.

[1] https://www.census.gov/geo/reference/webatlas/tracts.html

Executive Summary

The National Renewable Energy Laboratory (NREL) and the United States Department of Energy (DOE) Office of Energy Efficiency and Renewable Energy Strategic Programs sought to address gaps in sub-national energy data. This report describes the methodologies developed to estimate the electricity and natural gas consumption and expenditures for the residential, commercial, and industrial sectors at sub-state geographies and made available on the State and Local Energy Data[2] site to enable more strategic energy decisions by a variety of stakeholders.

To estimate energy consumption at the sub-state level, researchers proportionally allocated energy sales and revenue values reported by the Energy Information Administration from state and utility geographies down to smaller geographies. Within each sector, population distributions are developed through the selection of cohorts assumed to have homogeneous energy use characteristics. In the residential sector, cohorts are based on locational, occupancy, physical, and demographic characteristics of the housing units. In the commercial sector, cohorts are based on climate zone, location, and use types of commercial square footage. The industrial sector is unique in that large greenhouse gas emitters are required to report at the facility level. These are treated individually, but the remaining cohorts are based on the North American Industrial Classification System (NAICS) codes and employment size class for business establishments. Energy values by cohort come from available statistical samples. These are then rescaled to match aggregate values from electric and natural gas utility reported sales and volumes as well as business reported fuel expenditures.

These methodologies model only electricity and natural gas consumption and expenditures and do not address other forms of building energy consumption and expenditures such as bottled gas (generally propane) or heating oil or on-site industrial consumption of coal, coke, or diesel fuel for process energy.

Residential sector modeling generally yields more accurate results due to greater data availability and uniformity of per customer energy use. The opposite is true of the industrial sector where significant disparities in energy consumption by industry and among facilities within industries cause modeling challenges. In small towns where a single industrial facility can comprise most of the sector's consumption, modeled industrial sector data may be unreliable for planning purposes. Overall, a comparison with reported values by sector demonstrates that these sector-specific modeling methodologies provide reasonable results which can inform subnational-level energy decision making where measured, utility-reported data is unavailable. In addition, the modeled data provides robust, standardized, and disaggregated national data.

[2] https://apps1.eere.energy.gov/sled/

Table of Contents

List of Figures

List of Tables

1 Introduction

Patterns of energy consumption and expenditures vary significantly across the United States and among the various subpopulations of households and businesses. However, most energy data are available only at the aggregate state or utility service territory level, masking how energy use is distributed at finer resolutions. For many important energy-related problems, understanding the distributional characteristics is critical for informing potential solutions. As such, estimation approaches become necessary for filling informational gaps.

Several research efforts have estimated disaggregated energy use. Many of these approaches cannot be easily generalized because they rely upon specific information not commonly available across geographic or sectoral domains. For instance, urban building energy modeling typically relies upon the availability of detailed building stock information such as geometry, envelope and mechanical systems, and use types.[3,4] Other efforts, such as Project Vulcan, which models county-level energy data, fill some data gaps but leave others. For instance, Parshall et al.[5] have translated Vulcan data to energy use data but do not include electricity use since in Vulcan, electricity is attributed to the point of generation rather than the point of consumption. Site-level data exist for most large electricity generation facilities, either through reporting to the Energy Information Administration (EIA)[6] or the U.S. Environmental Protection Agency (EPA),[7] but not for most residential and commercial buildings or all but the largest industrial facilities.

Most estimation techniques rely upon combinations of bottom-up and top-down approaches. Bottom-up approaches apply representative values to underlying population distributions. Taking a simple example, the energy use intensity of commercial buildings may be assumed to vary by use type. These values can then be applied to the prevalence of different use types in different locations using building area as the weighing factor.

Top-down approaches begin with aggregate values and attempt to break down those values into progressively smaller geographic units and population cohorts of interest, each higher level constraining the values of the associated lower levels. For instance, modeled estimates of population growth at the national level are often used to support state-level projections by forcing the sum of state-level values to equal the associated national-level values.[8]

Combined approaches are advantageous because they allow flexibility in re-summation of values to reflect different geographies and cohorts and the full leveraging of existing aggregate

[3] Reinhart, Christoph F., and Carlos Cerezo Davila, 2016. "Urban Building Energy Modeling – A Review of a Nascent Field." *Building and Environment.* https://www.sciencedirect.com/science/article/pii/S0360132315003248

[4] B. Howard, L. Parshall, J. Thompson, S. Hammer, J. Dickinson, and V. Modi, "Spatial Distribution of Urban Building Energy Consumption by End Use," *Energy and Buildings*. https://qsel.columbia.edu/assets/uploads/blog/2018/publications/Spatial-Distribution-of-Urban-Building-Energy-Consumption-by-End-Use.pdf

[5] Lily Parshall, Kevin Gurney, Stephen A. Hammer, Daniel Mendoza, Yuyu Zhou, Sarath Geethakumar. 2010. "Modeling Energy Consumption and CO_2 Emissions at the Urban Scale: Methodological Challenges and Insights from the United States," *Energy Policy*. https://pdfs.semanticscholar.org/3f35/975247c2c263441904ceca205dc5a5999f8a.pdf

[6] EIA Form 861.

[7] EPA Greenhouse Gas Reporting Program (GHGRP).

[8] University of Virginia Weldon Cooper Center, Demographics Research Group. (2019). Virginia Population Estimates. Retrieved from https://demographics.coopercenter.org/virginia-population-estimates

information. One common summation of interest is the city level. A city energy inventory is usually conducted as part of a greenhouse gas (GHG) emissions inventory, for instance, since energy use drives a significant fraction of emissions.

Developing inventories is a resource-intense activity requiring willingness not only on the part of the local government, but also private entities such as electric and natural gas utilities and other fuel suppliers, to disclose data. Even in the absence of public support and private cooperation, estimated inventories can support improved community energy planning. Furthermore, estimates for cities with inventories based on reported data can be used to validate estimation approaches. Validation should be undertaken with caution, however. While GHG emissions inventories seek to use primarily observational data, in practice inventories rely upon both measured and modeled values.

Uses for disaggregated energy information extend beyond city energy planning. For instance, the changing nature of electric loads has led to efforts to improve their characterization for electric system planning. The distribution of different types of energy consumers and the associated end-uses informs systems planners on how electric loads may be impacted by external factors such as weather and how electric loads may evolve as new technologies and end-uses penetrate the market.

This report describes the methodologies developed to estimate electricity and natural gas consumption and expenditures for the residential, commercial, and industrial sectors at sub-state geographies. Sector-by-sector descriptions of the bottom-up methodology are provided first. Next, the top-down calibration process is discussed. Finally, estimates are validated against existing values from city and county-level reported data.

2 Methodology

Estimates of residential, commercial, and industrial electricity and natural gas use are performed similarly, with differences primarily stemming from the different availability of datasets by sector. Within each sector, population distributions are developed through the selection of cohorts assumed to have homogeneous energy use characteristics. In the residential sector, cohorts are based on locational, occupancy, physical, and demographic characteristics of housing units. In the commercial sector, cohorts are based on climate zone, location, and use types of commercial square footage. The industrial sector is unique in that large GHG emitters report at the facility level. These are treated individually, but the remaining cohorts are based on the North American Industrial Classification System (NAICS) codes and employment size class for business establishments. Energy values by cohort come from available statistical samples. These are then rescaled to match aggregate values from electric and natural gas utility reported sales and revenues as well as business reporting of fuel expenditures. Final values are calculated at the city and county levels. In this work, cities are generally defined as incorporated places. However, in 12 states, including New England states, New Jersey, New York, and Pennsylvania, minor civil divisions are also included to capture coterminous towns and townships that also serve as general-purpose local governments that can perform the same governmental functions as incorporated places. In the state of Hawaii, which has only one incorporated place, counties are used in lieu of incorporated places. Detailed sector-by-sector methodologies are provided below.

2.1 Energy Estimates for the Residential Sector

Estimates of residential energy consumption are based on cross-tabulations of U.S. Census housing data from the 2016 5-year American Community Survey (ACS5).[9] Average energy expenditures by different housing unit types are weighted by their housing unit counts to develop census tract-level estimates of energy expenditures. Those energy expenditure estimates are then rescaled to give energy consumption values, as described later. They are finally aggregated to cities and counties based on the census block-level occupied housing unit counts. (Further discussion of the spatial apportionment process is provided in Section 2.2.4).

Spatial allocation of different housing unit types relies on the use of an iterative proportional fitting (IPF) algorithm.[10,11] IPF is used sequentially to build increasingly complex cross-tabulations. Census tract-level published tables from the ACS5[12] are used as the marginal totals, and cross-tabulations of the ACS5 Public Use Microdata Samples for the corresponding Public Use Microdata Areas are used as the seeds in the IPF algorithm. The resulting cross-tabulations (see Table 1) include housing unit tenure, building year of first construction, number of units in the building, primary heating fuel type, number of persons, and household income. Finally,

[9] U.S. Census Bureau American Community Survey 2012-2016 ACS 5-year Estimates https://www.census.gov/programs-surveys/acs/data.html

[10] IPF is applied in Lovelace, R. (2014). *Introducing Spatial Microsimulation with R: A Practical*. National Centre for Research Methods Working Paper. http://eprints.ncrm.ac.uk/3348/4/spat_microsimulation_R.pdf

[11] IPF is also introduced in Pitchard, D. R. & Miller, E. J. (2012). Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation 39*(3), 685–704. http://www.springerlink.com/index/10.1007/s11116-011-9367-4

[12] U.S. Census Bureau American Community Survey 2012-2016 ACS 5-year Estimates https://www.census.gov/programs-surveys/acs/data.html

household income and number of persons are collapsed to a single variable using the U.S Department of Housing and Urban Development definition of area median income.

Table 1. Cross-Tabulation of ACS5 Data

Variable	Categories	ACS5 Published Table	ACS5 Microdata Sample
Tenure	Owner-Occupied and Renter-Occupied	Included below[a]	TEN
Building year of first construction	2010 and later, 2000-2009, 1980-1999, 1960-1979, 1940-1959, and 1939 and before	B25036, B25127[b]	YBL
Number of units in the building	1 Unit Detached, 1 Unit Attached, 2 Units, 3-4 Units, 5-9 Units, 10-19 Units, 20-49 Units, 50 and More Units, and Mobile and Other Units	B25032, B25124[b], B25127[b]	BLD
Primary heating fuel type	Utility Gas, Bottled Gas, Electricity, Fuel Oil, Wood, Coal, Solar, Other, and None	B25117	HFL
Number of persons	1-Person, 2-Person, 3-Person, 4-Person, 5-Person, 6-Person, 7 or More Persons	B25009, B25124[b]	NP
Household income	0-5K, 5-10K, 10-15K, 15-20K, 20-25K, 25-35K, 35K-50K, 50-75K, 75-100K, 100-150K, 150K and more	B25118	HINCP
Area median income	0-30%, 30-50%, 50-80%, 80-100%, greater than 100%	not available	not available

[a] Many of the published tables in the ACS5 are broken out by tenure.

[b] The ACS5 published tables include some cross-tabulations, including building year by number of units and number of persons by number of units. These cross-tabulations are incorporated in the IPF sequence to improve agreement between housing unit estimates and published Census values.

For each Public Use Microdata Area, corresponding energy expenditure values are calculated for each housing unit type in the above cross-tabulation. This requires modification to the microdata samples. To improve data accuracy, we removed survey responses for which housing energy costs are included in other housing costs, or energy costs across multiple fuels type are combined. Census tract-level housing unit counts are used to develop weighted averages.

The ACS5 provides energy expenditures for three categories of fuel types: electricity, gas, and other fuels. We assume that electricity is always provided by a utility. However, this assumption does not hold true for gas, which includes both utility-delivered natural gas and various forms of bottled gas. To make separate estimates of utility and bottled gas, we assume that housing units reporting bottled gas as their primary heating fuel type do not have access to utility gas. Those gas expenditures are subtracted from the associated census tract-level values to yield estimates of only utility gas expenditures.

As demonstrated in Section 4 Validation, resulting values are in good agreement with published inventory values. Nevertheless, this approach has several shortcomings aside from the more general issues associated with self-reported survey data. First, electricity and gas expenditures are taken for only a single month and that month is not reported publicly. Given the strong seasonal variation of energy use, extrapolation of annual values from monthly values hinges upon the quality of the sampling to cover all months of the year. This cannot be verified with the public data. Second, the ACS5 includes only occupied housing units. Unoccupied housing units may consume a significant fraction of energy. Third, elimination of responses reporting no expenditures associated with fuel use introduces unknown biases into the estimates. Fourth, the availability of piped utility natural gas is unknown and must be inferred from the prevalence of utility gas as a heating fuel source.[13] Finally, the characteristics of census tracts may differ significantly from their larger Public Use Microdata Area (PUMA), leading to poor IPF estimation. (For simple comparison, there are roughly 4,500 PUMAs and 73,000 census tracts in the U.S.)

2.2 Energy Estimates for the Commercial Sector

The commercial sector methodology starts with commercial customer electricity and natural gas sales and revenue data reported at the utility level by the EIA.[14,15] Utility service territories are mapped using ABB Velocity Suite data. The EIA-reported utility sales and revenues are proportionally allocated to census tracts based on building area by use type and assumed energy use intensities. A key missing component for energy estimation is a reliable dataset of all commercial buildings in the United States. To address this gap, we combined and normalized several datasets to create a comprehensive commercial building stock inventory. (See Appendix B for a summary flowchart of the commercial methodology.)

Developing a commercial building inventory included two distinct parts. First, based on the number of buildings and total square footage, an initial energy use value is assigned to every census tract. This value is calculated by using the Commercial Building Energy Consumption Survey (CBECS) reported energy use intensity (the amount of energy used by buildings of different use types and size) by the total square footage per building type reported by our commercial building inventory dataset. The total of the energy use values calculated in this step do not sum to reported utility sales; however, it provides an indication of the relative energy use across census tracts. EIA totals are then allocated proportionally according to the relative energy use across census tracts calculated in the previous step. The datasets used for commercial sector modeling are listed in Table 2.

[13] In some cases, the ACS5 incorrectly assigns utility gas heated housing units to areas lacking utility gas availability, which can lead to errors in this modeled data.

[14] EIA. 2016. "Annual Electric Power Industry Report (Survey Form No. EIA-861)." U.S. Energy Information Administration, Washington, DC, USA. Available online: http://www.eia.gov/electricity. Accessed December 1, 2018.

[15] EIA. 2016. "Annual Report of Natural and Supplemental Gas Supply and Disposition (Survey Form No. EIA-176)." U.S. Energy Information Administration, Washington, DC, USA. Available online: http://www.eia.gov/naturalgas. Accessed December 1, 2018.

Table 2. Datasets Used for Commercial Sector Modeling

EIA Form 861	Form 861 is a survey that gathers data on electricity sales, revenue, peak load, and energy efficiency.
EIA Form 176	Form 176 provides natural gas sales data at the utility level.
EIA CBECS	CBECS is a survey that collects sample information about the U.S. commercial building stock, including building activity, physical structure, and energy use.
Federal Emergency Management Administration's General Building Stock (GBS)	The Federal Emergency Management Administration's natural hazard model, or HAZUS, is used to assess the impacts of earthquakes, floods, hurricanes, and tsunamis. HAZUS depends on multiple datasets bundled within the model, including the GBS. The GBS provides square footage by occupancy and building type, building count by occupancy and building type, valuation by occupancy and building type, and general occupancy mapping[16] at the census block level. The GBS provides information on 33 building types, covering residential, commercial, industrial, and government buildings.
CoStar Realty Information, Inc. (CoStar)	CoStar provides commercial building data for purchase under specific license terms. While the GBS is an aggregated dataset at the census block level, CoStar provides information at the building level for over 4.5 million commercial buildings in the United States. This dataset was aggregated at the census block level to match the GBS.
ABB/Ventyx Energy Velocity Suite Utility Territories	The proprietary ABB Velocity Suite platform provides utility service territory datasets, both for electric and natural gas utilities, in a geographic information system format. These data are used to map utilities to other geographical units, such as census tracts and cities.[17]

2.2.1 Calculating Energy Intensity

The first step in disaggregating energy consumption data from utility service territories to census tracts is to estimate the commercial energy intensity of each census tract using energy consumption by building use type and size from CBECS. CBECS is organized by building type, census region, census division, and Building America climate zone. As demonstrated in Figure 1, census regions and divisions are among the largest levels of the census geography hierarchy. Buildings may have dramatically different energy use characteristics depending on their location within a multi-state census division. To address this problem, Building America climate zones were used.

[16] https://www.fema.gov/summary-databases-hazus-multi-hazard

[17] ABB/Ventyx Energy Velocity Suite. (2016). Local Distribution Company Territories. Retrieved June 1, 2018.

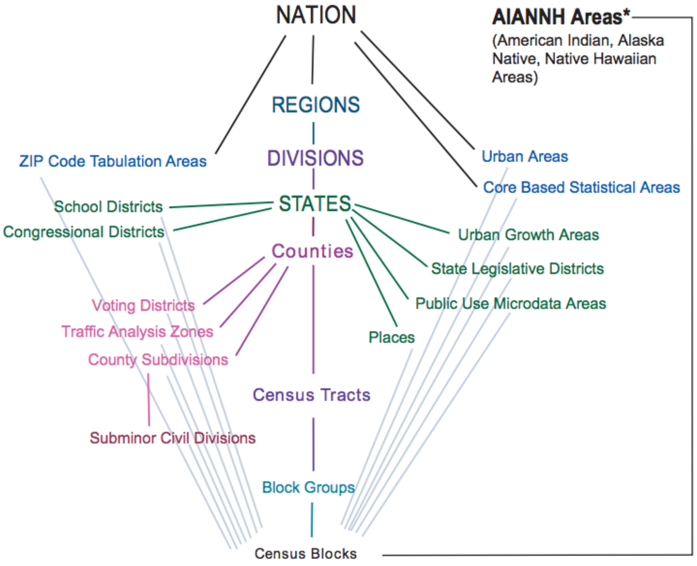

Figure 1. Census geographies hierarchy[18]

The U.S. Department of Energy (DOE) developed the Building America climate zones under the Building America initiative.[19] The climate zones follow regional climate patterns (see Figure 2). As such, a single state, census region, or division may contain several climate regimes that affect building energy use.

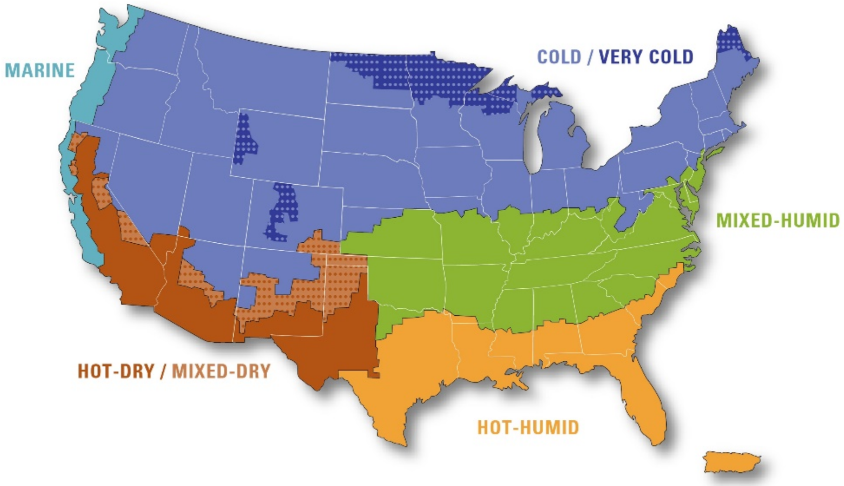

Figure 2. Building America climate zones[20]

Building America climate zones conform to county administrative boundaries, and since census tracts also nest within counties, it is possible to assign climate zones to all census tracts. After

[18] https://www.census.gov/geo/reference/hierarchy.html
[19] https://www.energy.gov/eere/buildings/building-america-bringing-building-innovations-market
[20] DOE https://www.energy.gov/eere/buildings/climate-zones

assignment, energy consumption is divided by square footage to obtain energy use and expenditures per square foot and by building use type. The data are then aggregated by climate zone, building type, and square footage class, which produces a dataset with energy use values per square foot per building use type at the census tract level.

These summed energy intensity values are joined to the building inventory based on building type and square footage class. The analysis was conducted using GBS and CoStar building inventories separately.

To join the CBECS data with the building inventory, we aligned CoStar building types with the Principal Building Activity Plus (PBAplus) building type descriptor system.[21] GBS, having fewer building types than PBAplus, was mapped to DOE Prototypes,[22] which is a system developed by the DOE to support the energy modeling software Energy Plus. In the final step, PBAplus building types were mapped to DOE Prototypes.

The post-processed CBECS building energy consumption data was allocated to the census tract-level building inventory using building type, PBAplus or DOE Prototype, square footage class, and climate zone as keys. This created an intermediate dataset with energy use per square foot and total square footage per building type at all census tracts, which were summed and summarized to obtain total commercial energy use for each census tract. The census tract commercial energy consumption and expenditure estimates were then calibrated to state totals from EIA Forms 861 and 176 using the approach described in Section 3 Calibration.

2.2.2 Data Integration

This analysis was performed using the different commercial building inventories datasets separately. While the original intent was to combine GBS and CoStar data into a single, more comprehensive inventory dataset, CoStar has less coverage of rural areas and public buildings than GBS. In addition, the GBS displays data projected to 2010 while CoStar covers data through May 2018 and each dataset has different building type descriptors, with CoStar providing more varied building type categories than GBS.

Thus, combining the datasets proved too complex for the scope of this project. It is important to note that this comprehensive, combined commercial building stock dataset could be explored in future research to improve the overall method. Instead, a simple process was implemented to select the buildings dataset with the higher commercial energy consumption for each census tract.

Figures 3 and 4 show the distribution of each dataset. Note how the GBS presents a higher distribution in both urban and rural settings while Costar presents a higher square footage total in both urban and rural areas, indicating that CoStar more accurately captures building area data, and the GBS data includes twice as many commercial buildings as Costar in rural areas. As CoStar data have a higher total building area and uses building-specific data inputs instead of

[21] PBAplus is a building activity classification system used by CBECS. See https://www.eia.gov/consumption/commercial/building-type-definitions.php

[22] https://www.energy.gov/eere/buildings/commercial-reference-buildings

modeled data, it is assumed that even though CoStar data have fewer urban buildings, it has a higher degree of accuracy for each building.

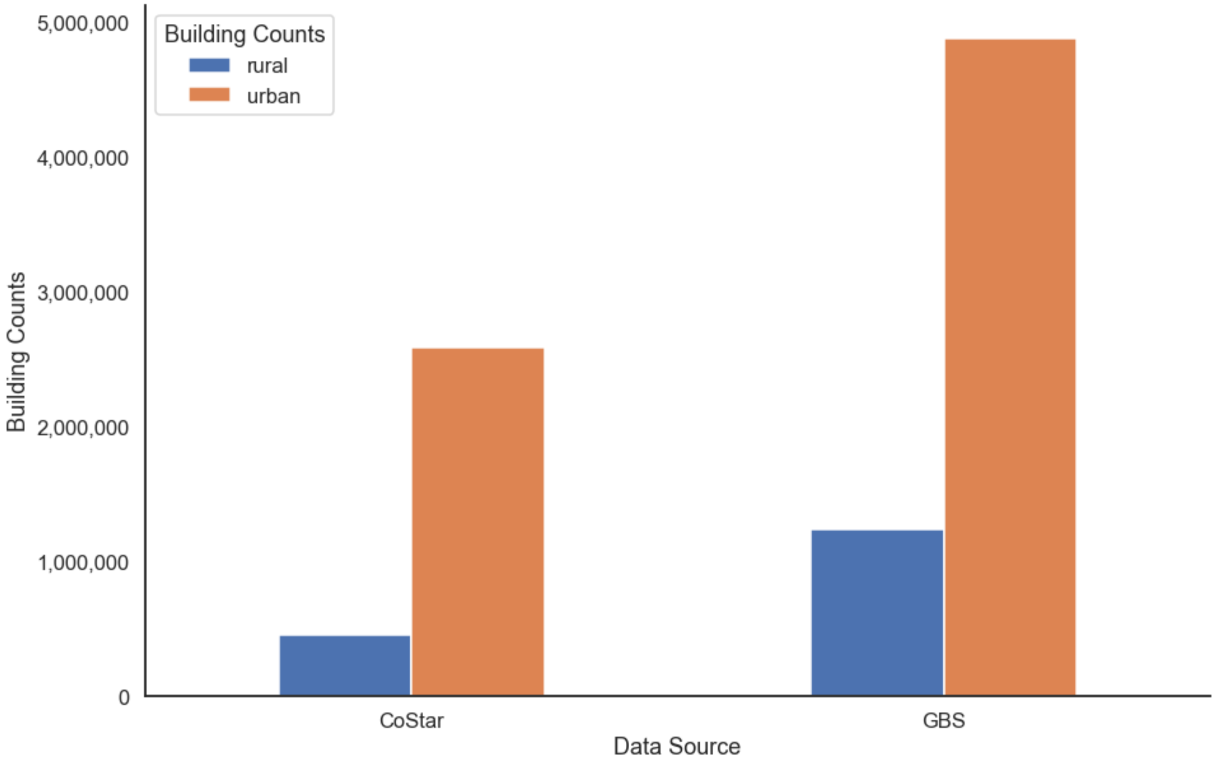

Figure 3. Building count distribution between CoStar and GBS

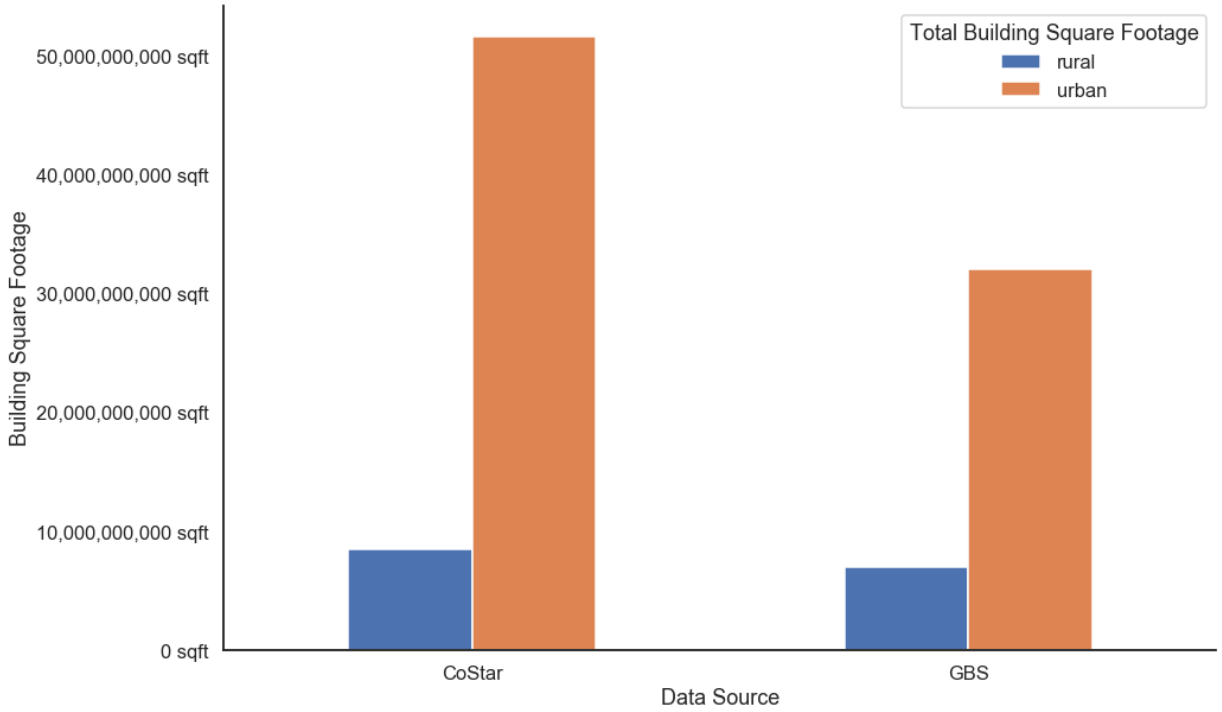

Figure 4. Total building square footage distribution between CoStar and GBS

Table 3 shows the aggregated totals for electricity consumption for the state of Colorado with the sum of total buildings and areas.

Table 3. Modeled Electricity Energy Consumption for Colorado

	Building Counts (CoStar)	Building Counts (GBS)	Building Square Footage (CoStar)	Building Square Footage (GBS)	Total MWh Consumption (CoStar)	Total MWh Consumption (GBS)
Rural	7,314	22,190	141,592,708	143,718,864	2,759,376	3,941,318
Urban	45,430	82,861	916,964,757	545,993,028	17,379,394	16,272,724

As shown in Table 3, CoStar's more detailed data on building area yield a higher total energy consumption for urban tracts despite having 54% fewer buildings than the GBS in urban tracts.

2.2.3 Aggregating to Cities

Once tract-level energy estimates are finalized, they are aggregated to cities using a mapping process between tracts and cities. This mapping process uses a weighting measure that allows for tracts that are not fully contained in a single city to be aggregated accordingly.

While the residential sector weighting is based on census block-level occupied housing unit counts, the commercial sector weighting uses commercial building inventory data, in this case GBS, because of the wider coverage, at the census block level. Census blocks are always completely contained within a census tract. This means that it is possible to build census tracts from census blocks. By knowing which census blocks make up a tract, it is then possible to effectively count how many buildings are inside and how many buildings are outside a city, and from this generate weights for that census tract.

Table 4 is an example of the final mapping. In this example, only 50% of the energy computed for tract 100002 would be assigned to city 01, and the remaining energy would either be assigned to an adjacent city or discarded. Then, after the weight is applied, the consumption values and expenditures are summed to obtain the total energy consumption and expenditure for a city.[23]

Table 4. Example of Mapping between Cities and Tracts with Weights

Tract ID	City ID	Weight
100001	01	1
100002	01	0.5

2.3 Energy Estimates for the Industrial Sector

The industrial sector methodology uses county-level estimates for 2014 industrial electricity and natural gas consumption from the Industrial Energy Tool and a point-level facility inventory to estimate the percent of state industrial energy that each city within that state consumes. That percentage is then multiplied by 2016 EIA estimates for state total industrial energy consumption to scale those city percentage estimates to 2016. The datasets used for industrial sector modeling are listed in Table 5.

[23] Jurisdictional boundaries are based on the U.S. Census Bureau 2013 Topologically Integrated Geographic Encoding and Referencing (TIGER)/Line Shapefiles: https://www.census.gov/geo/maps-data/data/tiger.html.

Table 5. Datasets Used for Industrial Sector Modeling

EIA Form 861	Form 861 is a survey that gathers data on electricity sales, revenue, peak load, and energy efficiency. Industrial sector estimates use only aggregated state-level Form 861 data.[24]
EIA Form 176	Form 176 provides natural gas sales data at the utility level. Industrial sector estimates use only aggregated state-level Form 176 data.[25]
Industrial Energy Tool (IET) county-level estimates	NREL developed the IET to be "an open-source, transparent, and flexible tool for exploring industrial sector energy and emissions scenarios." Using a variety of datasets, the tool provides industrial energy use estimates disaggregated to the county level and the four- to six-digit NAICS subsector (depending on subsector and region).[26]
Homeland Security Infrastructure Foundation-Level Data (HIFLD)/HSIP Gold 2012 Facility inventory	The HIFLD program was established by the HIFLD subcommittee to "address improvements in collection, processing, sharing, and protection of national geospatial information across multiple levels of government."[27] As part of the data provided by the program, HIFLD obtains data on the locations of businesses across the country from Dunn & Bradstreet. While the data are not part of the data that HIFLD makes available to the general public, as of 2012 it became available to government entities for planning and research purposes (under the name "HSIP Gold"). The subset of the HSIP Gold dataset used is a catalog of all manufacturing, mining, and agricultural facilities in the United States with coordinates of the location, NAICS code of the sector of the facility, and an estimate of the number of employees.
EPA Greenhouse Gas Reporting Program (GHGRP)	The GHGRP tracks the GHG emissions and energy use of large point sources of GHGs. This dataset allows for the resolution of exact amounts of natural gas consumption data for these large facilities. The dataset does not include electricity consumption data as it only tracks fuel sources that emit carbon at the point of consumption.[28] As GHGRP facilities are also generally included in the HSIP gold dataset, we filtered out facilities from the HSIP gold dataset that appear to already be counted by the GHGRP dataset to avoid duplicate data. This was accomplished through spatially filtering those HSIP facilities in very close proximity to GHGRP facilities with the same NAICS code and manually removing those with facility names that appeared to match those of the GHGRP facility.

2.3.1 Calculating Energy Intensity

As the IET leverages many datasets and uses sophisticated methods to provide highly disaggregated industrial energy estimates, it serves as a natural starting point for further resolving estimates to municipal boundaries. While the IET datasets are estimates, the open source nature of the tool provides transparency and enables our final estimates to be traced back to publicly reported data.

[24] EIA. 2016. "Annual Electric Power Industry Report (Survey Form No. EIA-861)." U.S. Energy Information Administration, Washington, DC, USA. Available online: http://www.eia.gov/electricity. Accessed December 1, 2018.

[25] EIA. 2016. "Annual Report of Natural and Supplemental Gas Supply and Disposition (Survey Form No. EIA-176)." U.S. Energy Information Administration, Washington, DC, USA. Available online: http://www.eia.gov/naturalgas. Accessed December 1, 2018.

[26] McMillan, Colin and Vinayak Narwade. 2018. The Industry Energy Tool (IET): Documentation. Golden, CO: National Renewable Energy Laboratory. NREL/TP-6A20-71990. https://www.nrel.gov/docs/fy19osti/71990.pdf.

[27] "Welcome | HIFLD," accessed December 6, 2018, https://gii.dhs.gov/hifld/.

[28] OAR U.S. EPA, "Greenhouse Gas Reporting Program (GHGRP)," Policies and Guidance, U.S. EPA, June 10, 2014, https://www.epa.gov/ghgreporting.

To apportion the IET data from counties to cities, it is necessary to use another dataset to understand the distribution of industrial activity within counties. Industrial energy consumption is far less predictable and uniform than the residential and commercial sectors. Industrial energy use within a city or county is often distributed among a small number of facilities, and consumption for these facilities can easily range over several orders of magnitude. With the possibility of a single or several facilities accounting for most of the industrial energy consumption in many counties, it becomes particularly important to pinpoint the exact location of industrial facilities. The correct estimation of a small city's industrial energy consumption may hinge on whether a large meat packing plant is correctly identified as being inside city limits or across the street in the unincorporated county, for example.

In addition, per-facility consumption is highly dependent on the subsector. This is demonstrated by comparing the EIA national electricity use estimates by manufacturing sector for 2014 with the Census Bureau Statistics of US Businesses national counts of facilities by sector for 2014 (the most recent year for which both are available). For example, it is estimated that 6,891 apparel manufacturing facilities (NAICS code 315) in the nation used 803 million kilowatt-hours for an average use of approximately 0.1165 million kilowatt-hours per facility. The 4,558 primary metal manufacturing facilities (NAICS code 331) used 138,437 million kilowatt-hours of electricity, for an average annual use of approximately 30.37 million kilowatt-hours per facility.[29] This shows an average difference of over two orders of magnitude. Thus, it is important to estimate both the spatial and subsector distribution of all industrial activity within the county.

In a survey of available data, the HSIP Gold dataset met three criteria: comprehensively cataloging all industrial facilities in the United States, having point location attribution, and categorizing facilities to a subsector. While the 2012 vintage is less than preferable for 2016 estimates, it is assumed that the turnover in facilities between 2012 and 2016 would, in general, not be large enough to overcome the dataset's advantages. One other weakness of the dataset is that it does not inventory construction activities in any way, whereas construction is considered a subsector by EIA and is estimated at the county level by the IET. The nature of retail electricity and natural gas use for construction is spatially and temporally dynamic. Unlike a factory, the point of consumption for a construction company moves throughout the year and between years as different projects are started and completed. Therefore, the error in assessing the value for any particular year could be particularly high. Additionally, the IET estimates that the construction sector only consumes about 5.2% of total industrial energy and the EIA Annual Energy Outlook 2016 estimates construction energy consumption at 3.6% of total industrial energy. This suggests that the construction sector is not likely to be a prime driver of industrial energy use for most cities.

EIA provides state-level industrial energy consumption data, including construction activity. As such, it is necessary to express the IET estimates at the county and sector level (hereafter called "*county-sector estimates*") as proportions of state industrial consumption, including construction. The county-sector estimates, including those for the construction sector, were summed to the

[29] U.S. Census Bureau, "2014 SUSB Annual Data Tables by Establishment Industry," accessed December 3, 2018, https://www.census.gov/data/tables/2014/econ/susb/2014-susb-annual.html; "U.S. Energy Information Administration (EIA) - Consumption & Efficiency Data."

state level to provide the total electricity and natural gas consumption estimates for the entire state. Then, the proportion of each state's total consumption that was attributed to each county-sector was calculated by dividing the county-sector estimate by the state sum. For the purposes of this document, this proportion will be known as the "*county-sector proportion.*"

Since the GHGRP data were used as an input to the IET tool, it was necessary to disaggregate the proportion of the IET fuel consumption estimates attributable to GHGRP facilities. This was achieved by summing up the natural gas consumption of all GHGRP facilities for each county-sector combination to find the total natural gas consumption attributable to GHGRP for each county-sector. This number was then divided by the IET county-sector estimates to find the proportion of natural gas consumed by GHGRP facilities for each county-sector. To find the proportion of the IET county-sector estimates that still need to be resolved to the city level using the HSIP dataset, it is only necessary to subtract this proportion from 1. For purposes of this document, this proportion will be referred to as the "*county-sector HSIP proportion.*"

Since the GHGRP facility consumption still had to be attributed to each city, the GHGRP facilities were spatially joined with a city boundary dataset to assign a city to each facility. Then, the natural gas consumption of all GHGRP facilities for each city-county-sector combination was summed. This sum was then divided by the total consumption for each county-sector to express the proportion of each county-sector's energy use attributable to GHGRP facilities in each city in each county. The proportion will be referred to as the "*city GHGRP proportion.*"

To determine the amount of activity within each county-sector that was attributable to each HSIP facility, the HSIP facilities were spatially joined with a county boundaries dataset to assign a county to each facility. Then, the total employees for each county-sector was summed. It is assumed that, within each subsector, the number of employees listed for the facility is a reasonable proxy for the size and therefore relative energy requirements of the facility. The proportion of the county-sector energy use that was attributable to each facility was then found by dividing the employee count for each facility by the county-sector sum.

To find the proportion of each county-sector energy estimate attributable to each city, the HSIP facilities were intersected with a city boundaries dataset to assign a city to each facility. It was then possible to sum the county-sector proportions for each facility up to the city level. The result of this is the total proportion of the county-sector energy consumption that is estimated to be used by HSIP facilities by each city in each county. This will be referred to as the *"city HSIP proportion.*"

This gives all the information necessary to calculate each city's proportion of the total state energy consumption attributable to each city-county-sector combination (see Appendix B for summary flowcharts of the industrial methodology). For electricity, the calculation is:

> County-Sector-City Proportion = County-Sector Proportion * city HSIP Proportion

For natural gas, the calculation is:

> County-Sector-City Proportion = County-Sector Proportion * ((County-Sector HSIP Proportion * city HSIP Proportion) + city GHGRP proportion)

The resultant data can be summed by city to express the proportion of total state energy use attributable to each city. It is important to note that these proportions in most cases do not sum to one due to industrial activity that took place outside of cities. The proportion is then multiplied by the 2016 State energy consumption and expenditures (as given by EIA form 861 for electricity or EIA form 176 for natural gas) to give industrial sector electricity and natural gas consumption by city for 2016. Note that care must be taken to properly account for unit conversions. EIA form 861 reports electricity sales in gigawatt-hours and revenue in millions of dollars. EIA form 176 reports natural gas sales in million cubic feet (MMcf) and prices in dollars per thousand cubic feet (Mcf), such that revenues must be calculated as:

$$\text{Revenue} = \text{Sales in MMcf} * \text{Dollars per Mcf} * 1{,}000$$

The city consumption numbers for both electricity and natural gas must be multiplied by 1,000 to be in the units reported in our final dataset of megawatt-hours and thousand cubic feet. The electricity revenue numbers also must be multiplied by 1,000 to be in the final units reported of thousands of dollars.

Potential shortcomings in this method of estimating industrial energy include the use of employee size as a proxy for energy use in industrial facilities. Employee size is used in both the IET tool county estimates and the assignment of county energy consumption to cities via the HSIP facility data. While the accuracy of this proxy is improved in both cases by using economic subsectors in addition to employee size, there is uncertainty as to the accuracy of this approach. In addition, the HSIP facility datasets do not include employee size estimates for the mining subsector. Another shortcoming is that the industrial facility catalogue provided by HSIP is not exhaustive and has a vintage several years prior to the estimates produced for 2016. Errors may also occur when a single facility dominates industrial energy consumption in small cities.

3 Calibration

Modeling energy consumption and expenditures at the census tract level represents a bottom up approach. Estimates can be improved through fully leveraging existing aggregate information as well as the data available at different geographies and cohorts used to build the bottom up approach. To combine the bottom up and top-down approach and create a more robust model, we rescale census-tract estimates using energy intensity weighting and sum to state totals reported in EIA Forms 861 and 176.

This rescaling or calibration process of allocating total state energy consumption to each census tract uses two key assumptions. For electricity, it was assumed that all tracts are served by a utility company. This assumption is supported by the high electrification rate found in the United States. For natural gas, it was assumed that not all tracts are served by utility gas, as other fuel types such as bottled gas, coal, or liquid fuel may serve as substitutes for utility gas.

To inform the subsequent steps, electric and natural gas utility territories were mapped to census tracts. This mapping was conducted using the Ventyx utility territories spatial data, which provides boundary information for electric and natural gas companies.[30] A simple intersection was conducted between the territories and the census tracts. Utility IDs were also normalized to facilitate merging the spatial representation of the territories and the data provided by EIA.

Mapping utility service territories to census tracts sought to achieve two goals. First, it was important to know which tracts should be excluded from the natural gas allocation process. Second, to rescale bottom up estimates using energy intensity weighting, each tract was assigned the energy sales rate (dollars per megawatt-hour for electricity and dollars per million cubic feet for natural gas) of the utility that serves it, which was obtained from EIA's data files. For tracts that had no electric utility initially assigned to it, as in states with unbundled utility service, the state average per-unit commodity and delivery service values price was used. For tracts where several utilities were present, the average of the prices was used.

Census-level consumption and expenditure estimates were geographically rescaled to match utility sales (electricity) or volumes (natural gas) and revenues. Different weighting approaches were used for the residential and commercial sectors.

3.1 Residential sector energy intensity weighting

Rescaling cannot be accomplished exactly for the residential sector because the number of utility customers does not correspond to the number of housing units and utility territories often overlap. Therefore, the calibration process first assigns utility customers by utility to census tract using the IPF algorithm. The geospatial overlay of Ventyx utility boundaries with census tracts yields the binary matrix $\boldsymbol{U}$ of 1's and 0's, with columns corresponding to utilities and rows to census tracts. Column marginal sums are utility customers and row marginal sums are total housing units, buildings, and establishments. The resulting matrix $\boldsymbol{C}$ is the approximate assignment. Note that the IPF does not converge as the column and row marginal sums total to

[30] ABB/Ventyx Energy Velocity Suite. (2016). Local Distribution Company Territories. Retrieved June 1, 2018.

different values. However, each iteration preserves the corresponding marginal sum, in this case, chosen to be utility customers to match the EIA data.

Per customer consumption and expenditures are assumed to be similar among customers within a single census tract but served by different utilities. Based on this assumption, per customer consumption $\boldsymbol{l}$ and per customer expenditures $\boldsymbol{l}'$ should also be similar whether taken from the utility-reported values, with $\boldsymbol{S}_{\boldsymbol{rep}}$ as a diagonal matrix of utility-reported sales and volumes per customer,

$$\boldsymbol{l} = \left[\boldsymbol{U} \cdot \left(\boldsymbol{C} \cdot \boldsymbol{S}_{\boldsymbol{rep}}\right)^{\boldsymbol{T}}\right] \cdot \mathbf{1}$$

or estimated from the customer perspective, with $\boldsymbol{R}_{\boldsymbol{est}}$ as a diagonal matrix of estimated consumption per customer,

$$\boldsymbol{l}' = [\boldsymbol{U} \cdot (\boldsymbol{C}^{\boldsymbol{T}} \cdot \boldsymbol{R}_{\boldsymbol{est}})] \cdot \mathbf{1}.$$

Matrix multiplication by the column vector of ones $\mathbf{1}$ yields a row-wise sum. The estimated values are then rescaled by the ratio of these quantities to give the following calibrated result

$$\boldsymbol{R}_{\boldsymbol{cal}} = \boldsymbol{R}_{\boldsymbol{est}} \cdot (\boldsymbol{l} \oslash \boldsymbol{l}')$$

where the operator $\oslash$ represents element-wise division. Derivation of these equations is provided in Appendix A. The diagonal elements of $\boldsymbol{R}_{\boldsymbol{cal}}$ are the calibrated census tract-level estimated consumption per customer.

The calibration approach described in Appendix A leads to improved agreement with available measured inventory data. However, there are several instances where this approach can lead to poor agreement:

- The equivalence of $\boldsymbol{l}$ and $\boldsymbol{l}'$ holds under the assumption that census tracts are relatively homogeneous in per-customer energy consumption. However, the calibration process is biased toward larger utilities in instances where two or more utilities share a single census tract and those utilities have very different per customer energy consumption values.
- There are data gaps and inaccuracies in the mapping of utilities to census tracts, leading to incorrect assignment of customers to utilities and customers to census tracts.
- The IPF approach may poorly reflect the true geographic distribution of utility customers.
- Utility service territories may be large, encompassing many cities and counties. For instance, in a severe case, most of the state of Rhode Island is served by a single distribution utility. The calibration process merely reflects the state averages and provides poor results for areas served by the state’s two much smaller municipal utilities.

3.2 Commercial sector energy intensity weighting

Differences in the number of buildings provided by the CoStar and GBS commercial building inventories and the number of customers reported on EIA’s Forms 861 and 176 are the primary cause of misalignment between state totals and census tract-level estimates. Nonetheless, the

CBECS-derived energy consumption data can be converted into weights that inform the relative energy uses among tracts, which are the metrics used to estimate the energy intensity of each tract.

Once the spatial disaggregation was complete, the energy consumption and expenditure calculations were finalized. Energy consumption values were obtained by distributing state totals to all relevant census tracts using the energy intensity weights obtained from CBECS. Expenditures are equal to consumption multiplied by the sales rate ($/energy unit).

4 Validation of Estimates to Measured Values

4.1 Validation of Residential Modeled Data

There are several publicly available sources of data for validation purposes. Xcel Energy, an investor-owned utility that operates in several western and midwestern states, reports community-level electricity and natural gas sales and revenue data. Also, many cities in Minnesota report consumption data through the Regional Indicators Initiative (RII). Further, Massachusetts reports community-level electricity data through Mass Save.[31] Finally, many local governments have undertaken GHG emissions inventories. As part of an intermediary step, those inventories often report electricity and natural gas consumption by sector.

One significant difficulty with validation is that the number of reported utility customers may differ significantly from the number of occupied housing units. There are several reasons why this occurs. Most commonly, there are discrepancies between the community boundaries used by the reporting entity and the community boundaries used by the U.S. Census. There are also difficulties with estimating the number of utility customers. While we can assume that most occupied housing units are served by an electric utility, there may be many unoccupied housing units that also are served by an electric utility, and multi-family units may share a single meter. Furthermore, many occupied housing units are not served by a natural gas utility, and multi-family units may share a single meter or common heating unit. Our assignment of natural gas customers to census tracts is approximate. Comparisons of state- and local-calibrated estimates to community data from Xcel Energy (electricity and natural gas) and Mass Save (electricity only) are shown in Figures 5 and 6, where each community-wide consumption data point reported is charted according to its difference with modeled data. State-level calibration rescales energy consumption estimates to match state-wide totals (see Figure 5). However, local-calibrated estimates (see Figure 6) rescale energy consumption based on utility-level totals (as described in Section 3.1 and Appendix A).

[31] "Mass Save® | Energy Assessments | Equipment Rebates | Incentives," accessed July 3, 2018, http://www.masssavedata.com/Public/GeographicSavings?view=U

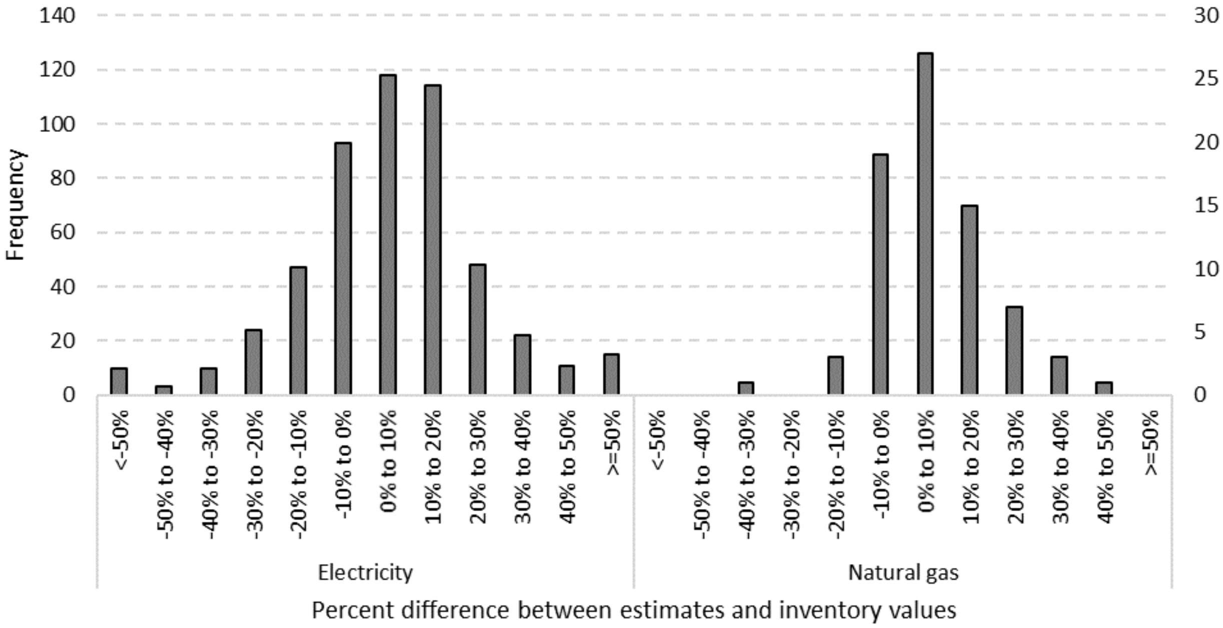

Figure 5. Comparison of state-calibrated residential estimates to community data from Xcel Energy and Mass Save.

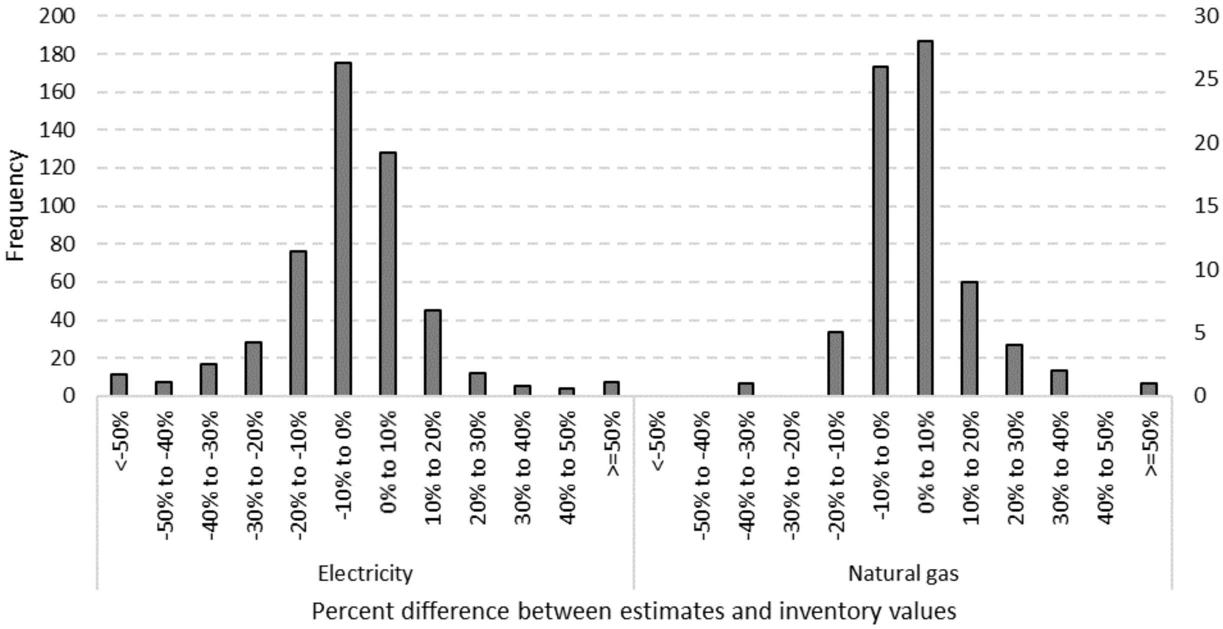

Figure 6. Comparison of local-calibrated residential estimates to community data from Xcel Energy and Mass Save.

This comparison illustrates that the majority of residential energy consumption estimates based on the methodology described herein are within +/- 10% of the 515 reported electricity and 76 reported natural gas city and county inventory values provided by Xcel Energy and Mass Save.

There are several outliers with large discrepancies. Many of those communities are very small, with less than 1000 occupied housing units, where potential errors in the Census survey data or differences in geographic boundary definitions between utilities and Census could generate large differences on a percentage basis. To add, the Mass Save data does not include customer counts, preventing basic checks against this issue.

4.2 Validation of Commercial and Industrial Modeled Data

High-quality validation datasets for the industrial and commercial sectors are scarce, a key reason this work was undertaken. Xcel Energy's Community Energy Reports[32] provide consumption and expenditure data for 200 cities located within the utility's service territory, which covers parts of six states. However, Xcel Energy only reports disaggregated commercial and industrial consumption in its Community Energy Reports when certain rules about the number of facilities are met to protect anonymity. In addition, the classification of a company as commercial or industrial is done according to a voluntary survey. When the survey is not submitted, the company is simply classified as commercial. This means the commercial energy consumption is likely overestimated while industrial energy consumption is underestimated. With no information on survey response rates, there is no way to estimate the extent to which this is the case.

Definitions of customer types can also conflict with definitions of building activity types. Datasets such as data from EIA forms 861 and 176 and Xcel Community Energy Reports report the total number of customers, which reflects the total number of connections and not the total number of buildings serviced. The total number of customers is usually lower than the total number of buildings, as evident in Figure 7.

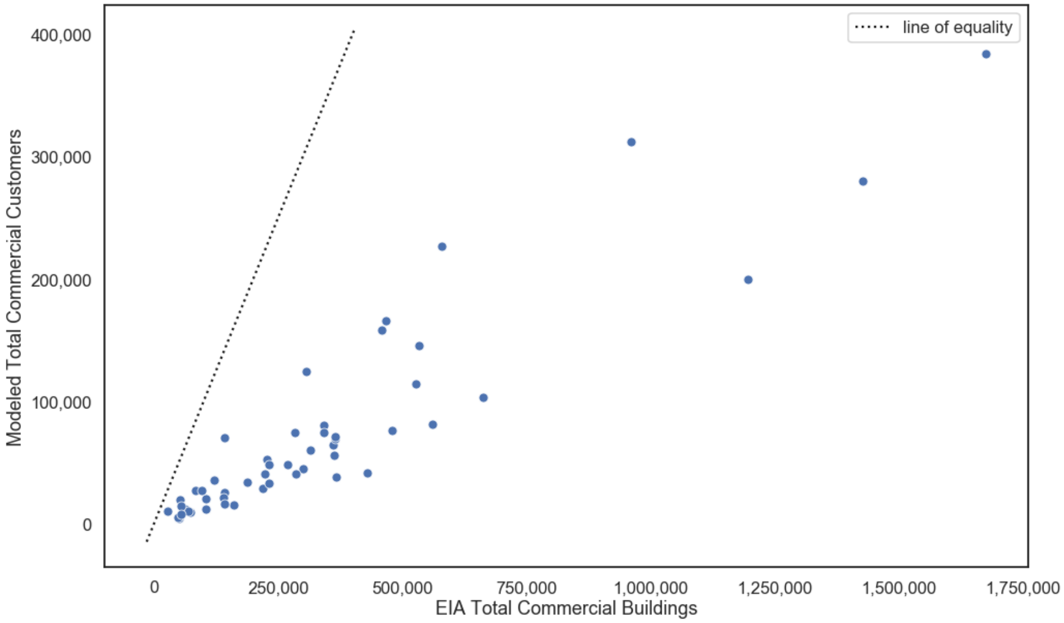

Figure 7. Modeled total buildings versus EIA reported commercial customers per state

[32] https://www.xcelenergy.com/working_with_us/municipalities/community_energy_reports

It is difficult to assess how much this difference between buildings and customers impacts the final energy estimates. Since the tract energy intensity was calculated based only on the total square footage of the building, one can argue that this difference may not have a strong impact. However, it is important to note that the current approach does not take into consideration the difference between occupied and unoccupied commercial buildings. By only using the building square footage and use type, the method assumes all buildings are occupied and therefore are using energy at the average rate for buildings of the same class.

One high-quality dataset is municipal electricity use for industrial and commercial sectors gathered by the Mass Save program led by the State of Massachusetts.[33] The Mass Save program collects data on energy use for every municipality in the state, except for some that are censored for certain anonymity reasons. Still, the dataset gives data on combined commercial and industrial retail electricity use for 299 cities. While there are still certain discrepancies between what is reported and what our estimates include (for instance, the Mass Save data likely include electricity use by the construction sector whereas our estimates do not include this), the dataset presents one of the best and largest sources for validation available for commercial and industrial electricity consumption.

Our aggregated estimates for commercial and industrial electricity use vs. those reported by Mass Save are shown in Figure 8. Due to the skewed distribution of electricity use, both estimates are log transformed.

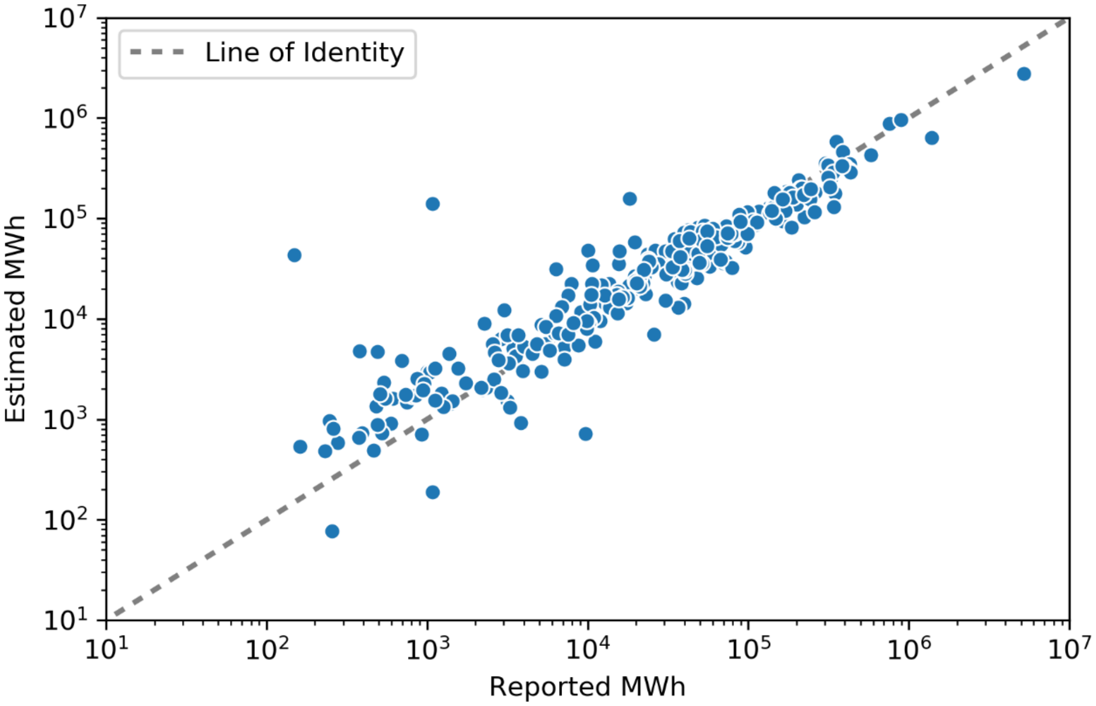

Figure 8. Plot of estimated commercial and industrial retail electricity consumption for municipalities and townships in Massachusetts vs. reported values from the Mass Save program

[33] "Mass Save® | Energy Assessments | Equipment Rebates | Incentives," accessed December 6, 2018, https://www.masssave.com/.

A simple linear regression of the log-transformed estimated values predicting the log-transformed reported values gives an intercept of -0.69 and a slope of 1.05, meaning that there is very little bias in the estimates, and an R-squared value of 0.87, implying that most of the variance in the estimated values is due to the actual underlying difference in electricity consumption. There are certain outlier values where the estimates are several times larger or a small fraction of the reported values; however, these outliers seem to be more common for cities with small reported values as opposed to large reported values. This makes sense as facilities datasets used for the estimates have more accurate coverage within larger cities than small towns where a single facility can dominate energy consumption.

From the perspective of a local planner, it matters less how closely the dataset tends to reflect energy use and instead how much the estimate for a single city can be relied on to track the city's actual energy use. From this perspective, it may be more constructive to understand the distribution of the percent errors between the estimates and reported values. Such a histogram, with bins like the bins used for the residential validation, is shown in Figure 9.

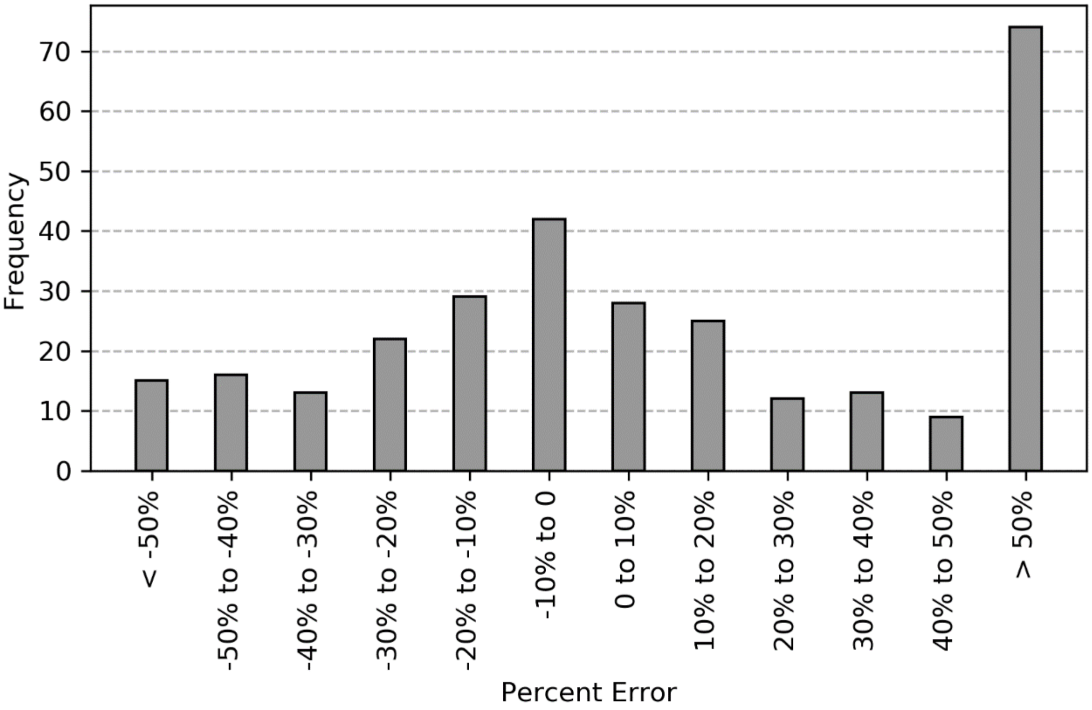

Figure 9. The percent error of estimated commercial and industrial electricity consumption vs. reported values in the Mass Save dataset

While the estimates can predict the reported values with low bias, the average percent error is very high compared to the residential data. While 52.8% of the cities have an estimate value within 30% of the reported values, 30.8% of the values have an error greater than 50%. This is not due to a bimodal distribution but rather to a preponderance of small towns with reported consumption values of under 10,000 MWh, for which the model reliably tends to overestimate consumption. Therefore, while the estimates may represent an advancement in the ability to estimate these values at such a scale, they also underscore the need for better data on commercial and industrial energy use and facilities to be collected both at the local and national level. While our estimated values seem to represent an unbiased and relatively low noise representation of the true consumption at an aggregate level, they may be unacceptably inaccurate to inform local planning decisions, particularly for smaller towns.

Because this validation exercise only represents electricity consumption from one state, we thought it necessary to seek at least one other validation dataset. The other validation dataset used is from the Minnesota RII program. Like Mass Save, the program reports combined commercial and industrial consumption, in this case for both electricity and natural gas. Data for only 46 cities were available for 2016, much smaller than the Mass Save dataset of 299 cities, but still enough for approximate validation.

A simple linear regression of the estimated electricity consumption values and reported values gives a relatively biased fit, with an intersect value of 2.2, a slope value of 0.82, and an R-Squared valued of 0.72. However, a single outlier data point was identified that had an outsized effect on these coefficients. With the outlier removed, the results improved significantly to an intercept value of 0.42, a slope value of 0.97 and an R-Squared value of 0.84.

A simple linear regression of the estimated natural gas consumption values and reported values also gives a relatively biased fit, with an intersect value of 3.8, a slope value of 0.7, and an R-squared valued of 0.7. However, the same point that was an outlier for the electricity values also was a very influential outlier for the natural gas values. With the outlier removed, the results improved significantly to an intercept value of 1.05, a slope value of 0.9 and an R-squared value of 0.85. Figures 10 and 11 show scatterplots of the log-transformed values for electricity and natural gas estimates vs. the reported values without the outlier removed.

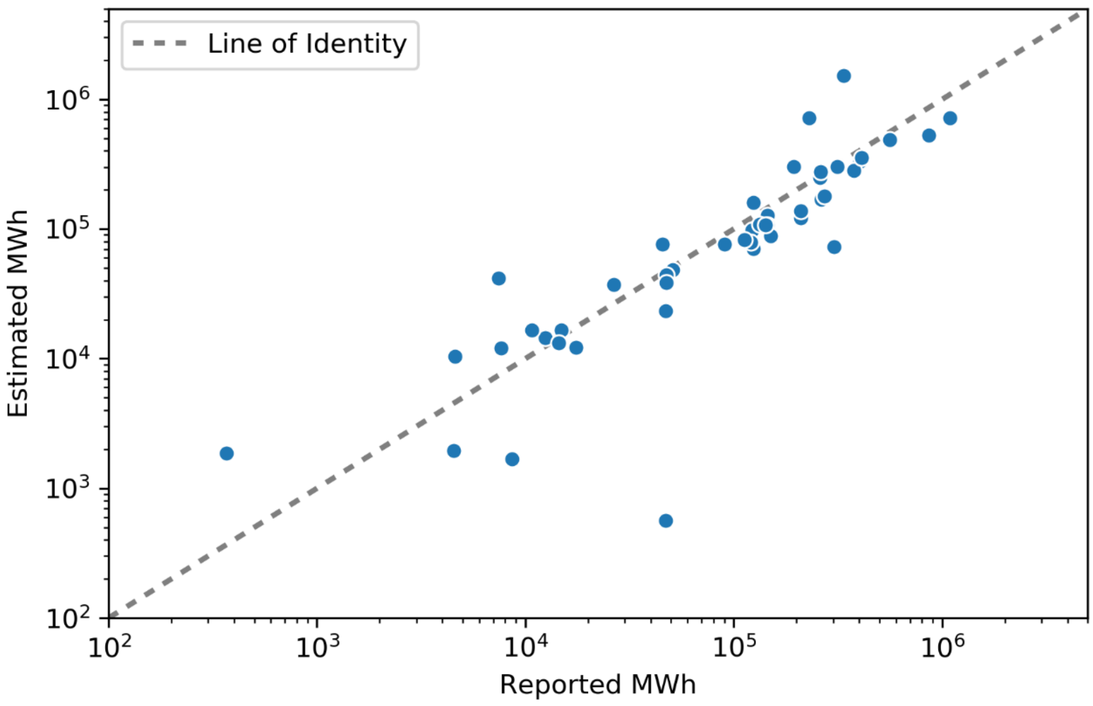

Figure 10. Plot of estimated commercial and industrial retail electricity consumption for select Minnesota municipalities vs. reported values from the Minnesota RII program

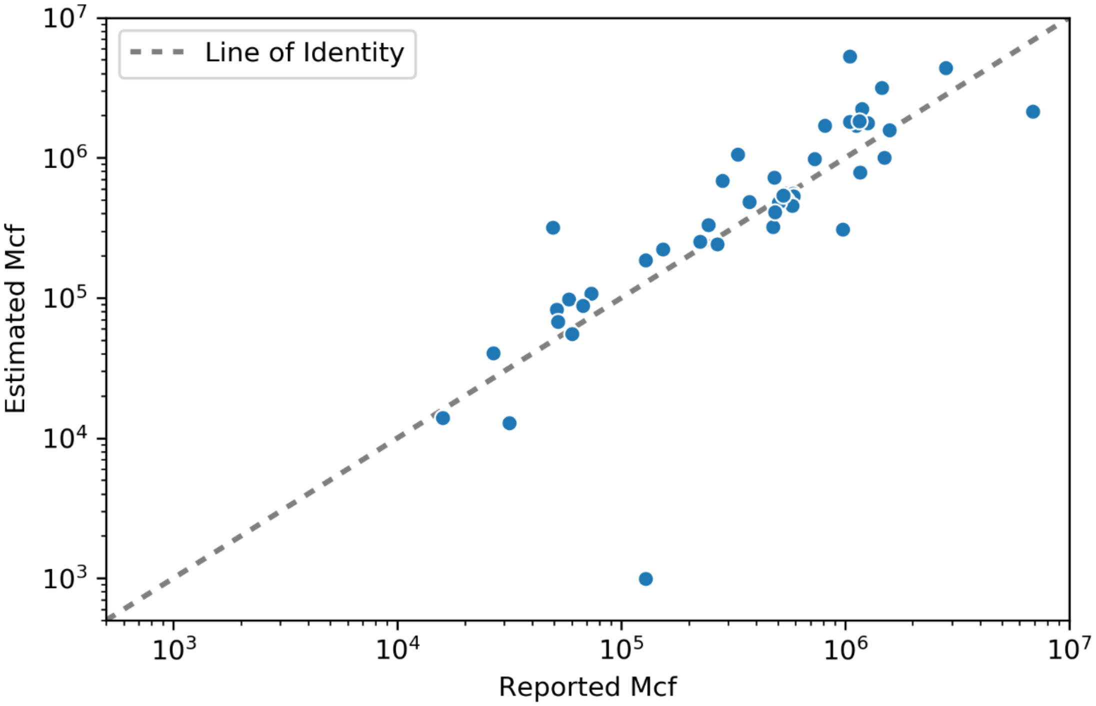

Figure 11. Plot of estimated commercial and industrial retail natural gas consumption for select Minnesota municipalities vs. reported values from the Minnesota RII program

Taken together, these plots mostly reinforce the story that the Mass Save dataset shows. While significant outliers occur, once they are removed, the validation shows the estimates are unbiased predictors of reported values. The natural gas validation shows there may be more of a bias with natural gas data than with electricity, with a slight bias towards underestimation. Since the Minnesota RII dataset represents a sample with some self-selection bias (unlike the Mass Save dataset, which is an almost comprehensive dataset for the entire state), some bias may exist due to the types of towns that report to RII in addition to the small sample size. If the RII tends to have data from smaller towns, which the Mass Save data suggest we do a poorer job of estimating, this could give the appearance of bias in our estimates that does not exist in the entire dataset.

5 Conclusion

The method of proportionally allocating energy use and expenditure values reported by EIA from utility geographies down to smaller geographic levels is shown to be a valid approach through comparing the resulting estimates with reported values at the city-level. This method provides scalability, enabling users to obtain energy estimates for a single county or city or for the whole nation.

The quality of the estimates is directly related to the quality of the input data. This is particularly true for the quality of the commercial building inventory data used to generate the energy intensities where a more comprehensive dataset with a wider range of building types and more extensive spatial coverage yields better results. Of the three sectors, greater existing data availability and uniformity of building-level energy use in the residential sector enable more accurate modeling. The opposite is true of the industrial sector where significant disparities exist in energy consumption by industry and among facilities within industries, making estimation challenging. In aggregate, validation demonstrates that estimated consumption provides an unbiased, relatively low noise representation of true consumption. However, increased inaccuracy in individual, small cities and towns, particularly for industrial estimates, makes modeled industrial data less reliable for planning purposes in these jurisdictions.

Overall, validation demonstrates that these sector-specific methodologies provide reasonable results that can inform city-level energy decision making where measured, utility-reported data are unavailable. The methodologies result in a robust, standardized national dataset of city-level energy consumption and expenditures available on the State and Local Energy Data site.[34]

One area for potential improvement in future versions is the validation step. The scarcity of standardized city-level data makes it challenging to validate the results at a national scale. Validation with city-level reported data should be conducted to the extent possible, with the goal of identifying patterns of overprediction and underprediction. Such patterns can inform which input dataset and metrics to revise and improve.

[34] https://apps1.eere.energy.gov/sled/

Appendix A. Derivation of Calibration Equations

The calibration approach seeks to identify groupings of census tracts within states that share a set of utilities. Then, the average per customer consumption over those census tracts should equal the corresponding average per-customer sales by the associated utilities. This assumes that states are separable in the sense that there are smaller than state-level groupings.

Start with a matrix $\boldsymbol{C}$, which maps utility customers (columns) to census tracts (rows), and a matrix $\boldsymbol{U}$, which is a binary representation of $\boldsymbol{C}$:

$$U_{ij} = \begin{cases} 1 \text{ if } C_{ij} \neq 0 \\ 0 \text{ if } C_{ij} = 0 \end{cases} \qquad \textbf{Eq. A.1}$$

Then, the following quantity provides the total number of utility customers shared by each census tract

$$\boldsymbol{c} = [\boldsymbol{U} \cdot \boldsymbol{C}^{\boldsymbol{T}}] \cdot \mathbf{1} \qquad \textbf{Eq. A.2}$$

where matrix multiplication by the column vector of ones $\mathbf{1}$ yields a row-wise sum. For example, suppose there are four census tracts labeled 1, 2, 3, and 4 and three utilities labeled A, B, and C, as illustrated in Figure A1.

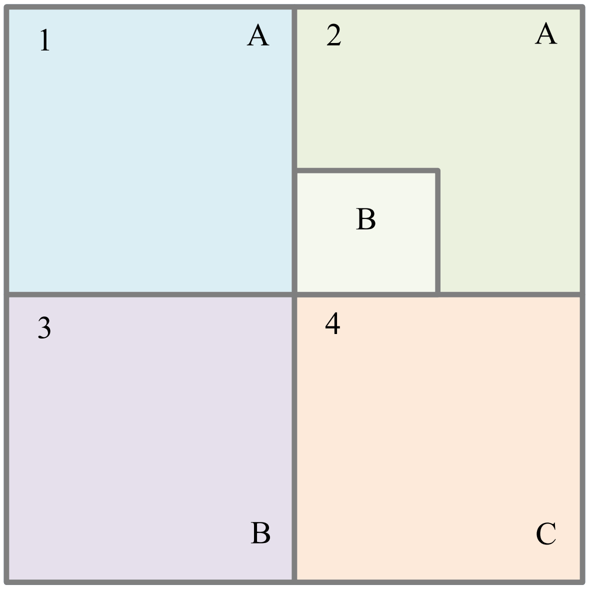

Figure A1. Example census tracts 1, 2, 3, and 4 and utility service areas A, B, and C

Further suppose this corresponds with the following mapping:

$$\boldsymbol{C} = \begin{bmatrix} 400 & 0 & 0 \\ 200 & 50 & 0 \\ 0 & 300 & 0 \\ 0 & 0 & 600 \end{bmatrix}$$

which assigns 400 customers of utility A to census tract 1, 50 customers of utility B to census tract 2, and so forth. Then, Eq. A.2 gives:

$$\boldsymbol{c} = \begin{bmatrix} 600 \\ 950 \\ 350 \\ 600 \end{bmatrix}$$

which shows that that census tract 1 shares 200 utilities customers in census tract 2 to give a total of 600 customers, census tract 2 shares 400 customers in census tract 1 and 300 customers in census tract 2 to give a total of 950 customers, and so forth.

The average per customer sales for each grouping may be calculated accordingly:

$$\boldsymbol{s} = \left\{\left[\boldsymbol{U} \cdot \left(\boldsymbol{C} \cdot \boldsymbol{S_{rep}}\right)^{\boldsymbol{T}}\right] \cdot \mathbf{1}\right\} \oslash \{[\boldsymbol{U} \cdot \boldsymbol{C^T}] \cdot \mathbf{1}\} \qquad \textbf{Eq. A.3}$$

where the operator $\oslash$ represent element-wise division. Here, we have simply inserted $\boldsymbol{S_{rep}}$ as a diagonal matrix of utility reported sales and volumes per customer to produce a customer-weighted average.

This same quantity may be calculated based on consumption per customer at the county level. In this case, the numerator is revised to read:

$$\boldsymbol{r} = \{[\boldsymbol{U} \cdot (\boldsymbol{C^T} \cdot \boldsymbol{R_{est}})] \cdot \mathbf{1}\} \oslash \{[\boldsymbol{U} \cdot \boldsymbol{C^T}] \cdot \mathbf{1}\} \qquad \textbf{Eq. A.4}$$

where $\boldsymbol{R_{est}}$ is a diagonal matrix of estimated consumption per customer.

Under the assumption that utility customers of different utilities but within a single census tract have the same average per-customer consumption, the equality $\boldsymbol{r} = \boldsymbol{s}$ should hold. Then, we can try to enforce the equality accordingly:

$$\boldsymbol{R_{cal}} = \boldsymbol{R_{est}} \cdot (\boldsymbol{s} \oslash \boldsymbol{r}) \qquad \textbf{Eq. A.5}$$

where the diagonal elements of $\boldsymbol{R_{cal}}$ are the calibrated census tract-level estimated consumption per customer. This equation is equivalent to equation Y if we simply set the numerator of $\boldsymbol{s}$ as $\boldsymbol{l}$ and the number of $\boldsymbol{r}$ as $\boldsymbol{l'}$.

To complete the simple example, if we assume:

$$\boldsymbol{S_{rep}} = \begin{bmatrix} 14 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 16 \end{bmatrix}$$

and

$$\boldsymbol{R_{est}} = \begin{bmatrix} 2300 & 0 & 0 & 0 \\ 0 & 2100 & 0 & 0 \\ 0 & 0 & 1700 & 0 \\ 0 & 0 & 0 & 2700 \end{bmatrix}$$

then

$$\boldsymbol{s} = \begin{bmatrix} 14 \\ 12.5 \\ 10 \\ 16 \end{bmatrix}$$

and

$$\boldsymbol{r} = \begin{bmatrix} 2233 \\ 2058 \\ 1757 \\ 2700 \end{bmatrix}$$

The calibrated per-customer consumption values are then:

$$\boldsymbol{R_{cal}} = \begin{bmatrix} 2300 & 0 & 0 & 0 \\ 0 & 2100 & 0 & 0 \\ 0 & 0 & 1700 & 0 \\ 0 & 0 & 0 & 2700 \end{bmatrix} \cdot \left(\begin{bmatrix} 14 \\ 12.5 \\ 10 \\ 16 \end{bmatrix} \oslash \begin{bmatrix} 2233 \\ 2058 \\ 1757 \\ 2700 \end{bmatrix} \right)$$
$$= \begin{bmatrix} 14.4 & 0 & 0 & 0 \\ 0 & 12.8 & 0 & 0 \\ 0 & 0 & 9.7 & 0 \\ 0 & 0 & 0 & 16 \end{bmatrix}$$

We can check this numerical result by re-calculating the utility values:

$$S_{est} = \{\boldsymbol{C} \cdot [\boldsymbol{1} \cdot \boldsymbol{R_{cal}}]\} \oslash \{\boldsymbol{1} \cdot \boldsymbol{C}\} = [13.9 \quad 10.1 \quad 16] \qquad \textbf{Eq. A.6}$$

showing good agreement with the diagonal values in $\boldsymbol{S_{rep}}$.

In this example, we have provided numerical values of similar magnitudes to the values used in the residential calibration. A value of 14 corresponds to 14 MWh/year, which is within the normal range of annual household electricity consumption. A value of 2,300 corresponds to \$2,300/year in electricity expenditures, also within the normal range. Residential estimates come from the 2016 ACS5, which samples expenditures and not consumption. However, this analysis assumes that expenditures are roughly proportional to consumption.

Note that the choice of $\boldsymbol{R_{cal}}$ is not unique, and in general, there are many possible values that provide agreement with the utility-reported values. However, this approach has the benefit of making a choice that retains some sense of the original guess $\boldsymbol{R_{est}}$, as most choices of $\boldsymbol{R_{cal}}$ are not physically plausible. One could further refine this approach by taking the calibrated value as the seed value in a numerical minimization, such that the utility values are matched exactly. Minimizing the difference between the values in Eq. A.6 and the diagonal elements of Eq.A.5 leads to the following result. Differences are less than 2%.

$$\boldsymbol{R_{fit}} = \begin{bmatrix} 14.6 & 0 & 0 & 0 \\ 0 & 12.8 & 0 & 0 \\ 0 & 0 & 9.5 & 0 \\ 0 & 0 & 0 & 16 \end{bmatrix}$$

Improving agreement with utility values is desirable; however, in practice, performing this minimization step is challenging. Any errors or unusually divergent values in $\boldsymbol{C}$ and $\boldsymbol{S_{rep}}$ can lead to $\boldsymbol{R_{fit}}$ values that deviate significantly from the original $\boldsymbol{R_{est}}$ guesses. Thus, in this report, we omit this second step and limit the calibration to Eq. A.5.

Appendix B. Methodology Flowcharts

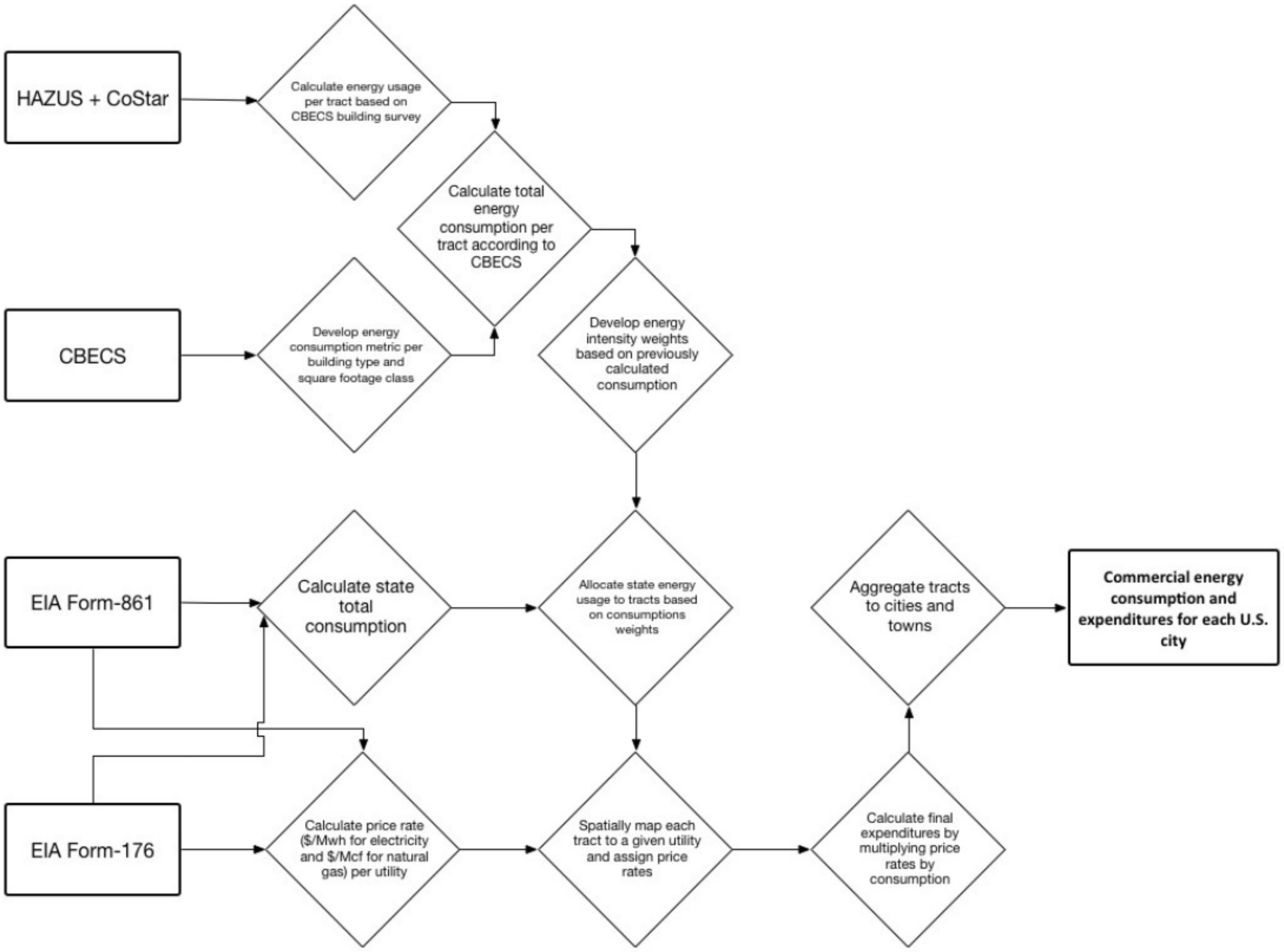
HAZUS + CoStar
Calculate energy usage per tract based on CBECS building survey
Calculate total energy consumption per tract according to CBECS
CBECS
Develop energy consumption metric per building type and square footage class
Develop energy intensity weights based on previously calculated consumption
EIA Form-861
Calculate state total consumption
Allocate state energy usage to tracts based on consumptions weights
Aggregate tracts to cities and towns
Commercial energy consumption and expenditures for each U.S. city
EIA Form-176
Calculate price rate ($/Mwh for electricity and $/Mcf for natural gas) per utility
Spatially map each tract to a given utility and assign price rates
Calculate final expenditures by multiplying price rates by consumption

Figure B1. Commercial energy consumption and expenditures methodology summary

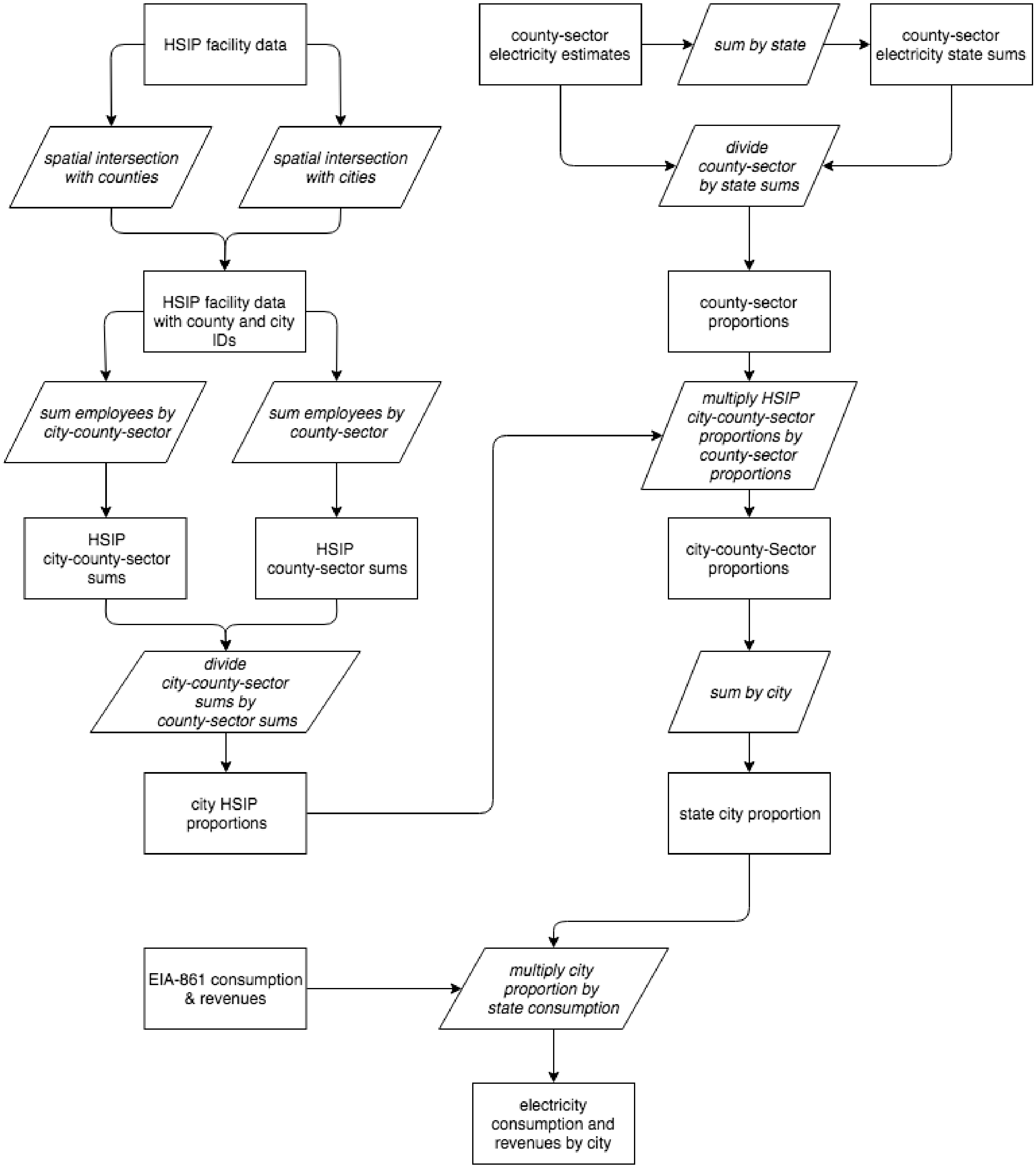
HSIP facility data
spatial intersection with counties
spatial intersection with cities
HSIP facility data with county and city IDs
sum employees by city-county-sector
sum employees by county-sector
HSIP city-county-sector sums
HSIP county-sector sums
divide city-county-sector sums by county-sector sums
city HSIP proportions
county-sector electricity estimates
sum by state
county-sector electricity state sums
divide county-sector by state sums
county-sector proportions
multiply HSIP city-county-sector proportions by county-sector proportions
city-county-Sector proportions
sum by city
state city proportion
EIA-861 consumption & revenues
multiply city proportion by state consumption
electricity consumption and revenues by city

Figure B2. Industrial electricity consumption and expenditures methodology summary

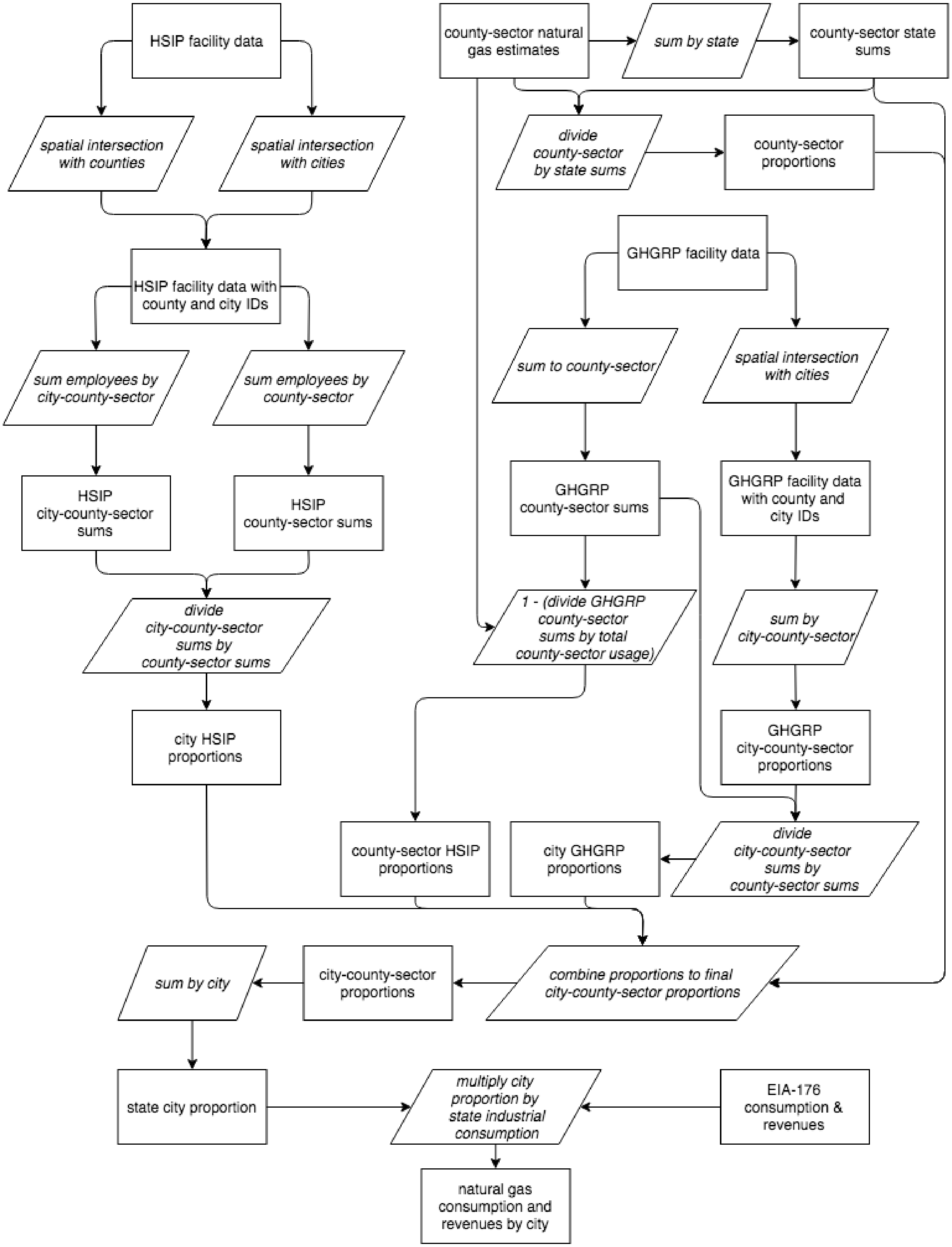
HSIP facility data
spatial intersection with counties
spatial intersection with cities
HSIP facility data with county and city IDs
sum employees by city-county-sector
sum employees by county-sector
HSIP city-county-sector sums
HSIP county-sector sums
divide city-county-sector sums by county-sector sums
city HSIP proportions
county-sector natural gas estimates
sum by state
county-sector state sums
divide county-sector by state sums
county-sector proportions
GHGRP facility data
sum to county-sector
spatial intersection with cities
GHGRP county-sector sums
GHGRP facility data with county and city IDs
1 - (divide GHGRP county-sector sums by total county-sector usage)
sum by city-county-sector
GHGRP city-county-sector proportions
county-sector HSIP proportions
city GHGRP proportions
divide city-county-sector sums by county-sector sums
sum by city
city-county-sector proportions
combine proportions to final city-county-sector proportions
state city proportion
multiply city proportion by state industrial consumption
EIA-176 consumption & revenues
natural gas consumption and revenues by city

Figure B3. Industrial natural gas consumption and expenditures methodology summary