Leveraging Natural Language Processing for Efficiencies without Increasing Nonresponse Bias

For

Federal Committee on Statistical Methodology

October 27, 2022 | Washington, DC

The analysis and conclusions contained in this presentation are those of the authors and do not represent the official position of the U.S. Energy Information Administration or the U.S. Department of Energy.

By

Sarah Grady, Francisco Cifuentes, S. Grace Deng, and Katie Lewis

About the Residential Energy Consumption Survey (RECS)

- A nationally-representative and state-representative household survey about energy characteristics and energy usage patterns in homes

- Conducted in two phases
  - Collects data on household energy characteristics, usage patterns, and demographics in the **Household Survey**
  - Uses respondents' answers to questions about energy suppliers and account numbers to obtain detailed bills during the **Energy Supplier Survey**

- Data used by EIA's Office of Energy Analysis, modelers, and researchers interested in patterns in household energy usage

- First conducted in 1978; last conducted in 2020

Data science is great for efficiency

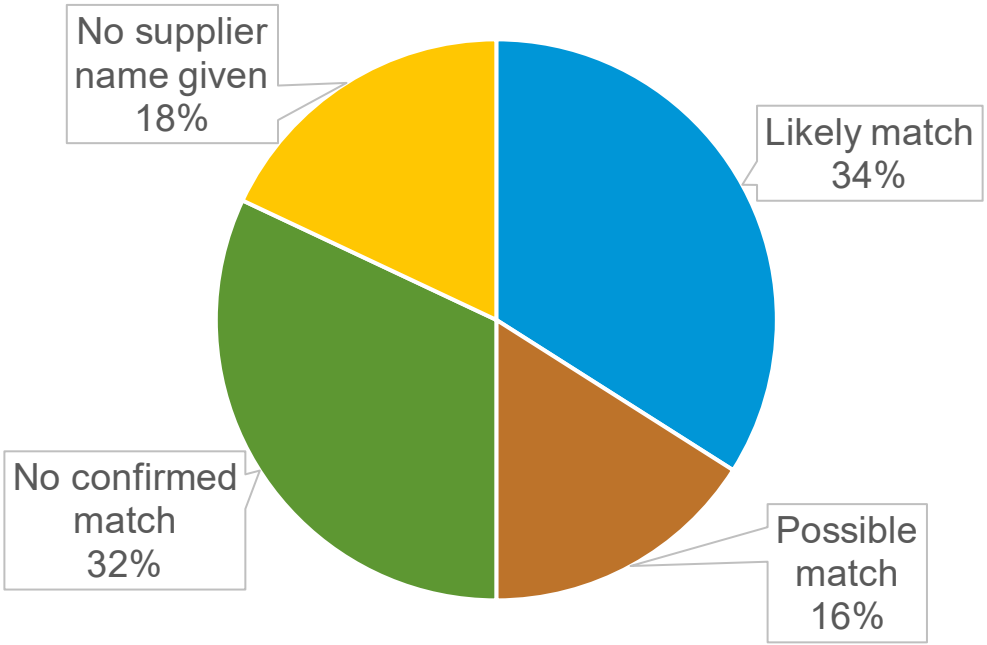
- Comes with great responsibility
- RECS matching algorithm for Energy Supplier Survey (ESS)
  - Built to automate the process of matching respondents' reports of their energy suppliers to the actual energy suppliers (e.g. Pepcom = PEPCO)
  - Needed for the ESS collection of bills
  - New for 2020, where RECS sample is about 3 times larger than in previous collection years
- When under constraints, the algorithm tempts us to only work the "easy to match" cases

About the matching algorithm (Martin et al 2022)

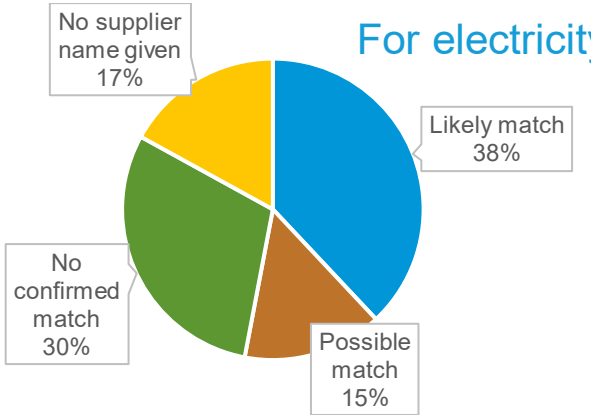
- Python script searches for variations on supplier names
- Compares a household-provided supplier to a reference list
- Calculates a score based on the Levenshtein distance between input text and reference list candidates
  - A value between 0 and 1 where 0 = identical
  - Between 0 and 0.2 = likely match

Martin, M., Good, C., Amsbary, M., Cifuentes, F. (2022) Using Natural Language Processing to Help Develop a Frame of Energy Suppliers. FedCASIC virtual conference

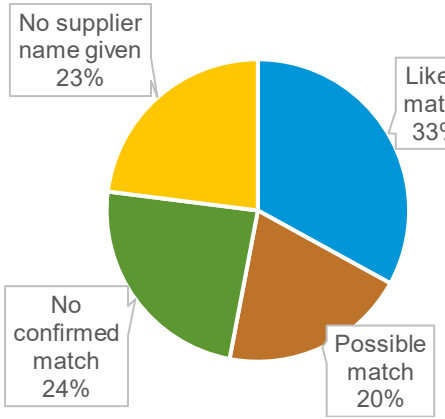
Matching algorithm flag results for electricity, natural gas, fuel oil, and liquid petroleum



For electricity only



For natural gas only

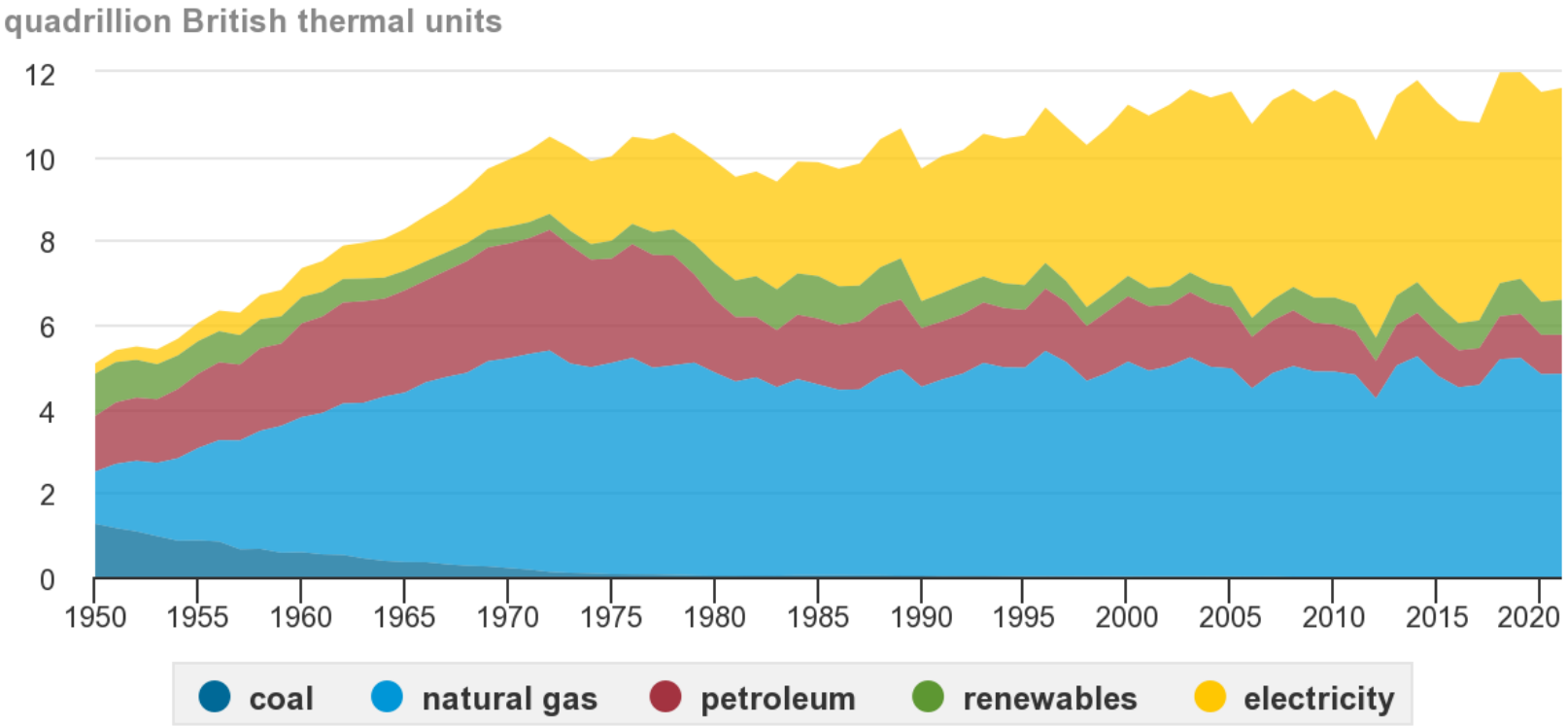


Martin, M., Good, C., Amsbary, M., Cifuentes, F. (2022) Using Natural Language Processing to Help Develop a Frame of Energy Suppliers. FedCASIC virtual conference

Research questions

- Does a relationship exist between match rate and respondent characteristics?
- Between match rate and consumption?
- What would consumption estimates look like that just use the matched cases?
- Can we develop a methodology to prioritize cases and make sure that we are not biasing our estimates by only spending time on the easy cases?
- DISCLAIMER– The RECS program is not seeking ways to decrease the size of its Energy Supplier Survey; it has never only worked the “easy cases”

U.S. residential sector energy consumption by energy source, 1950 to 2021



Data source: U.S. Energy Information Administration, *Monthly Energy Review*, Table 2.2, April 2022, preliminary data for 2021

Note: Electricity excludes losses in electricity generation and delivery. Petroleum includes heating oil, liquefied petroleum gas (propane), and kerosene. Renewables includes wood, geothermal energy, and solar energy.

Household and respondent characteristics related to differences between subgroups in "likely" match rate for electricity

Comparisons >= 10 percentage points	Comparisons >=5 percentage points	Comparisons >=3 percentage points
Housing type	Owner vs renter	Households that received a disconnection, shut off, or nondelivery notice within previous year
Urbanicity	Household income	Year home built
Region, census division, and state		Square footage
Respondent's race/ethnicity		Female vs male respondents
Number of people in household		
Respondent's level of education		

Household and respondent characteristics related to differences between subgroups in "likely" match rate for natural gas

Comparisons >= 10 percentage points	Comparisons >=5 percentage points	Comparisons >=3 percentage points
Census division and state	Housing type	Number of people in household
Respondent's race/ethnicity	Urbanicity	
	Year home built	
	Lived in home at least a few years vs just moved into home in 2020	
	Region	
	Respondent's level of education	
	Household income	

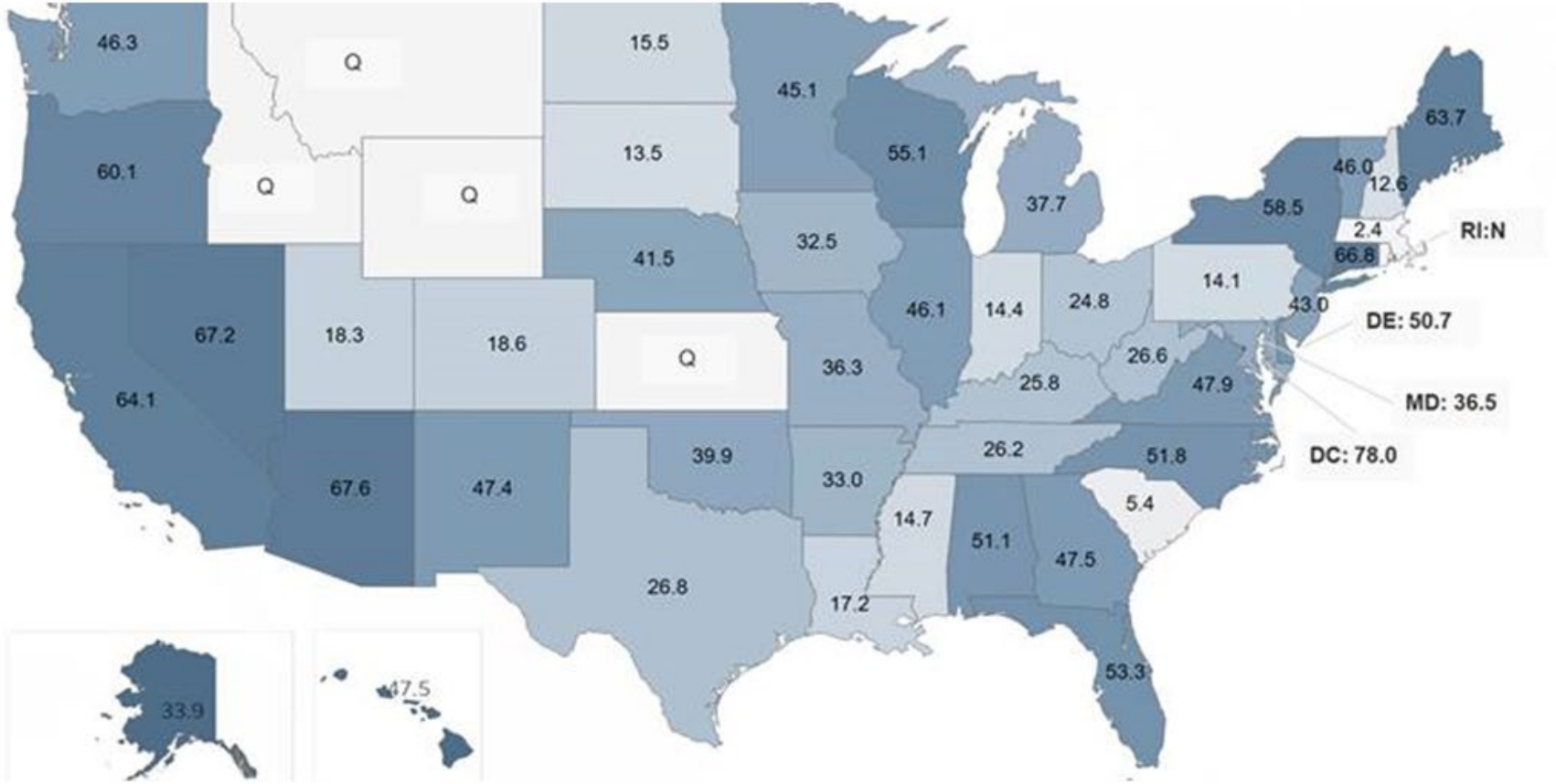
Selected differences in "likely" electricity match rate by demographics

Higher match rate	Percentage	Lower match rate	Percentage
Multifamily homes	43%	Mobile homes	29%
Urban areas	42%	Rural	29%
		Urban cluster	28%
Respondent with Master's degree	42%	Respondent with a high school credential	32%
Respondent with Master's degree	42%	Respondent with less than a high school degree	30%
Bachelor's degree	41%		
Non-Hispanic, Asian respondents	50%	Non-Hispanic, Black respondents	38%
		Non-Hispanic, White respondents	36%
		Non-Hispanic respondents who are American Indian, Alaska Native, Native Hawaiian, other Pacific Islander or more than one race	36%

Overall likely match rate = 38%

Range in likely match rates = 27% to 50%

Likely match rates to an electricity supplier by state

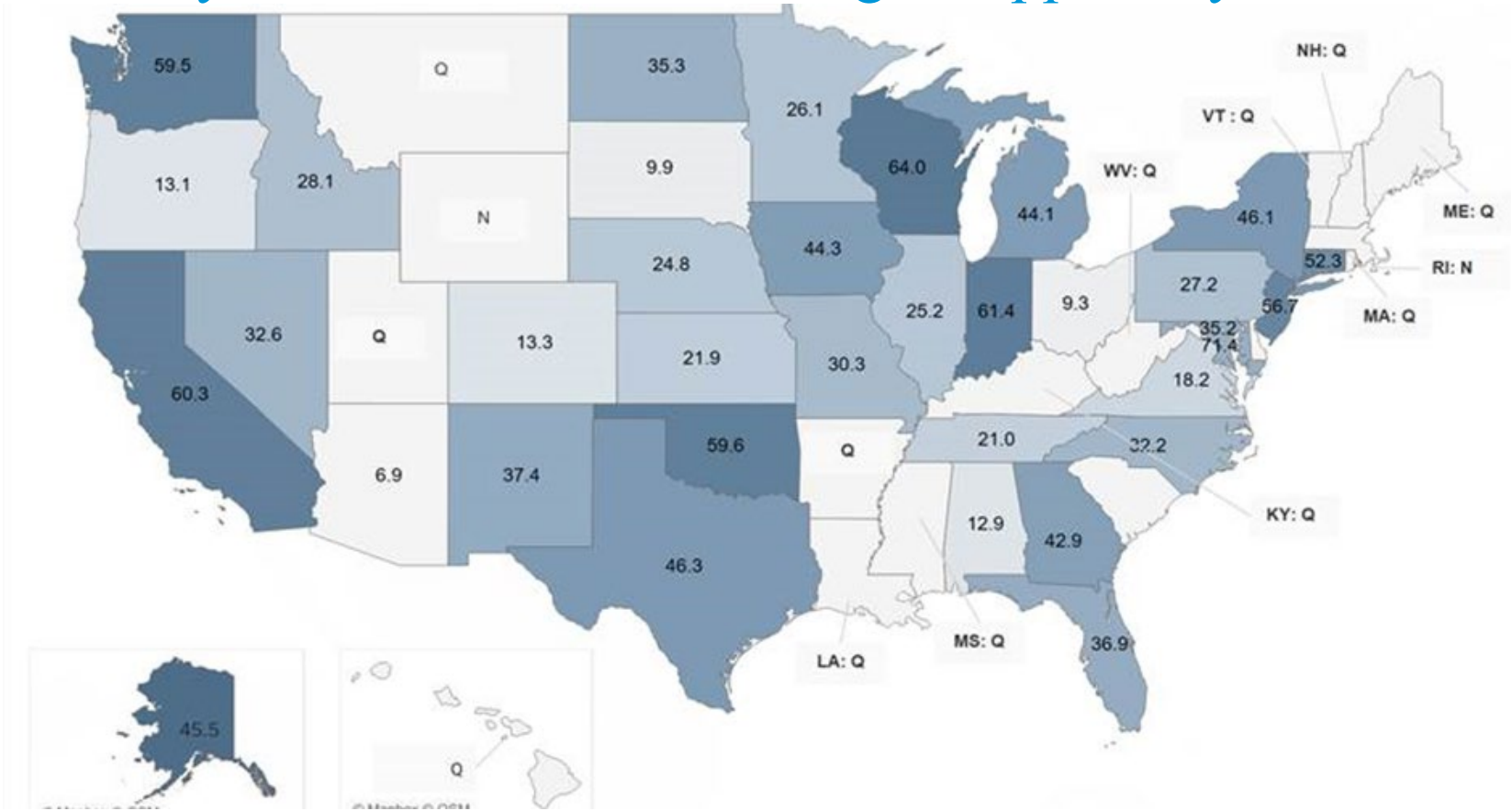


Selected differences in "likely" natural gas match rate by demographics

Higher match rate	Percentage	Lower match rate	Percentage
Non-Hispanic, Asian respondents	**43%**	Non-Hispanic, Black respondents	31%
		Non-Hispanic, White respondents	32%
		Non-Hispanic respondents who are American Indian, Alaska Native, Native Hawaiian, other Pacific Islander or more than one race	28%

Overall likely match rate = 33%
Range in likely match rates = 26% to 43%

Likely match rates to a natural gas supplier by state



A higher percentage of these households had no matches to an electricity supplier

Higher rate of NO match	Percentage	Lower rate of NO match	Percentage
Mobile homes	36%	Multifamily homes	**25%**
Urban clusters	40%	Urban areas	**26%**
Rural areas	38%		
Non-Hispanic, White respondents	32%	Non-Hispanic, Asian respondents	**20%**
Non-Hispanic respondents who are American Indian, Alaska Native, Native Hawaiian, other Pacific Islander or more than one race	31%		

Overall rate of no matches = 30%

Range in rates of having no match = 20% to 40%

A higher percentage of these households had no respondent data from the Household survey about supplier

Higher rate of no data for electricity	Percentage	Lower rate of no data for electricity	Percentage
Households with 6 or more members	26%	Households with 1 member	**14%**
Higher rate of no data for natural gas	**Percentage**	**Lower rate of no data for natural gas**	**Percentage**
Multifamily homes	36%	Single-family home	**21%**
Renters	33%	Owners	**21%**
Households that just moved	33%	Households where respondent had lived at address at least a few years	**22%**
Non-Hispanic respondents who are American Indian, Alaska Native, Native Hawaiian, other Pacific Islander or more than one race	32%	Non-Hispanic, White respondents	**21%**
Households with 7 or more members	31%	Households with 1 member	**20%**

Overall rate of providing no electricity supplier data on the Household survey = 17%

Range in rates of providing no electricity supplier data on the Household survey = 14% to 26%

Overall rate of providing no natural gas supplier data on the Household survey = 23%

Range in rates of providing no natural gas supplier data on the Household survey = 19% to 36%

Is there a relationship between match rate and annual consumption?

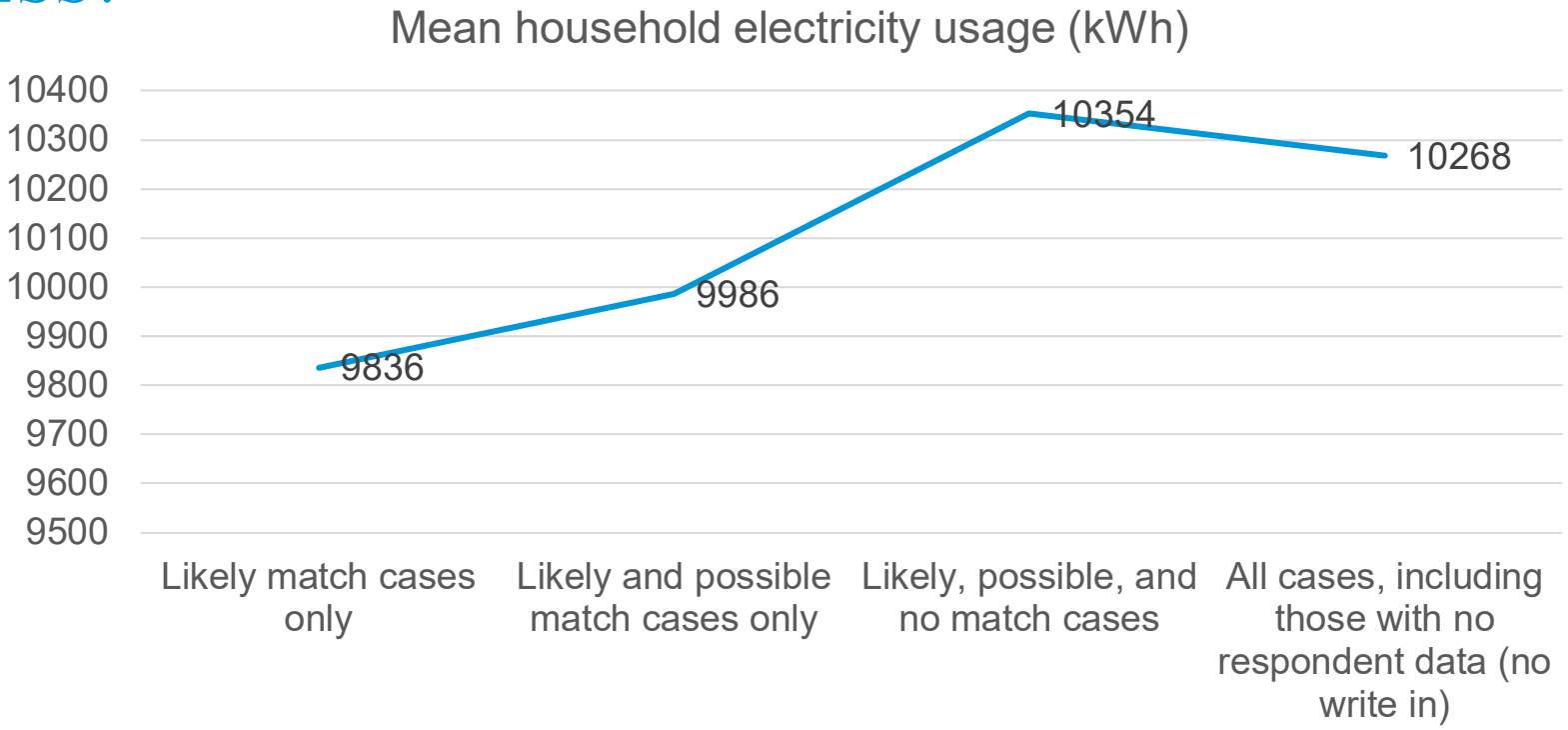
Categories of estimated electricity usage
High estimated electricity usage (>14,000 kWh)
Medium estimated electricity usage (>=10,000 & <14,000 kWh)
Low estimated electricity usage (<10,000 kWh)

25% of these cases are “no match” compared to 35% of high electricity consumers and 32% of medium electricity consumers

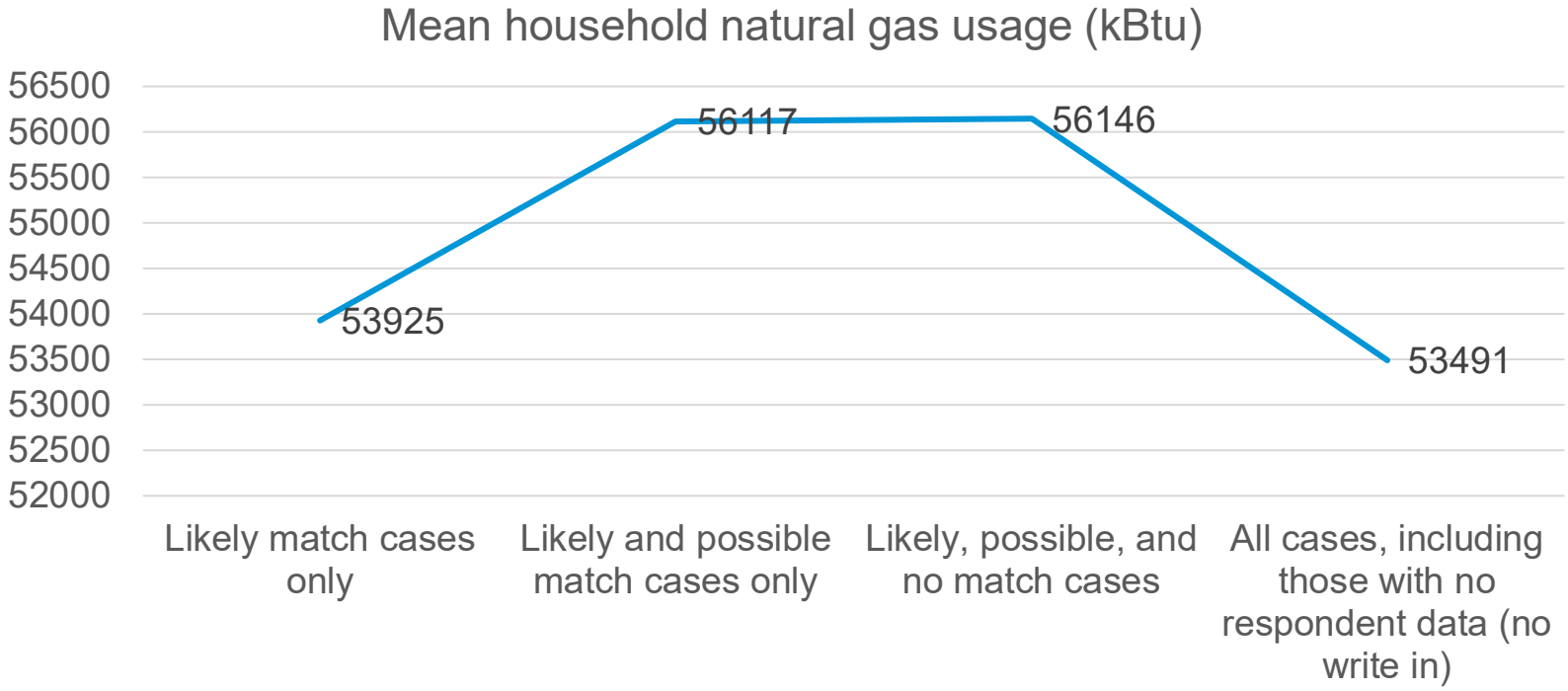
Categories of estimated natural gas usage
High estimated natural gas usage (>80,000 kBtu)
Medium estimated natural gas usage (>=25,000 & <80,000 kBtu)
Low estimated natural gas usage (<25000 kBtu)

34% of these cases had no data compared to 19% of high electricity consumers and 20% of medium electricity consumers

What do preliminary consumption estimates look like when we only have likely electricity matches included in the ESS?



What do preliminary consumption estimates look like when we only have likely natural gas matches included in the ESS?



Approaches to case prioritization and reducing bias in the literature

- Use base weights and response propensity models to create measures of influence
  - E.g., Riddles, M.K., and Krenzke T. for the Program for International Assessment of Adult Competencies (2016); West, Chang, and Zmich (2021) for the National Survey of Family Growth
- Prediction of ability to meet RSE goals for key estimates combined with response propensity
  - E.g., “Relative importance measure” used in the 2018 Commercial Buildings Energy Consumption Survey to determine mid-collection subsampled cases (Westat final report for the 2018 CBECS (2021));
- Monitor Representativity Indicator (R-Indicator) during data collection and adjust
  - Also uses estimated response probabilities

Use match rate underrepresentation to prioritize cases

- Prioritize mobile homes; rural addresses; urban cluster addresses; respondents with a high school credential or less education; non-Hispanic, White respondents; Non-Hispanic respondents who are American Indian, Alaska Native, Native Hawaiian, other Pacific Islander or more than one race

- Also, implement stopping rules for cases that are overrepresented by match rate
  - non-Hispanic, Asian respondents

- Prioritize cases with no respondent data like households of 6 or more members, multifamily homes, renters, households that just moved

Another approach – key variables

- Region
- Census division
- Urban/rural classification
- Climate region
- Housing types by ownership or rental
- Year of construction
- Total square footage
- Number of household members
- Income
- Payment method for energy bills
- Main heating fuel

Categories	Percentage distribution in "likely" matches group	Percentage distribution in population	Percentage distribution of total energy consumption (trillion Btu) (2015)	Percentage of Household cases for which we want reported ESS data
New England	7.0	4.8	6.0	**6.0**
Middle Atlantic	11.5	13.0	15.8	**15.8**
East North Central	10.6	15.0	19.3	**19.3**
West North Central	7.3	6.9	8.0	**8.0**
South Atlantic	20.8	20.1	17.4	**20.1**
East South Central	5.7	6.0	5.5	**6.0**
West South Central	7.4	11.8	10.8	**11.8**
Mountain North	1.8	3.7	3.9	**3.9**
Mountain South	8.4	3.7	3.0	**3.7**
Pacific	19.6	15.0	10.4	**15.0**

Another approach - relative standard error (RSE) monitoring as ESS frame is created: "Publishability"

RSE comparisons of engineering model consumption estimates between partial and complete cases for Mountain North census division and detailed housing type

		Partial data (n=13,417)			Complete data (n=18,496)		
		N	Mean	**RSE**	N	Mean	**RSE**
Mountain North	Mobile homes	57	9,348	**17.2%**	82	8,613	**9.4%**
	Single-family detached	642	10,098	**3.5%**	860	9,967	**2.6%**
	Single-family attached	77	8,117	**14.2%**	93	7,902	**6.7%**
	Apartment with 2-4 units	27	6,909	**25.2%**	38	7,326	**8.7%**
	Apartment with 5+ units	81	6,312	**13.5%**	107	6,334	**4.5%**

Conclusions

- Be careful with data science algorithms designed for efficiency. Don't be tempted into cutting corners. Not all cases are of equal value to your estimates
  - We found relationships between match rates and demographic characteristics, geographic location, and energy consumption
- Consider case prioritization strategies when your resources are limited
- Consider planning and schedule implications of investigating potential for bias within data science projects at the design stage

Limitations and areas for future research

- This is hypothetical and was done for demonstration
- Our analysis focused on what the algorithm predicted; we dropped cases from analysis that ultimately could not be matched to a supplier at all or where the supplier did not provide a response
  - Future algorithm development could be refined by incorporating correlates of nonresponse or of cases that cannot be matched to a supplier
- Future research could simulate impacts on workload of various bias reduction/prioritization strategies under constraint scenarios

Thank you!

sarah.grady@eia.gov

francisco.cifuentes@eia.gov

shaofen.deng@eia.gov

katie.lewis@eia.gov

Supplementary slide – state-level analysis

- We suspect that match rates at the state level are highly related to whether the state's residents only use one energy supplier or one natural gas supplier
  - E.g., PEPCPO, Washington National Gas, and Washington, DC
- Small sample sizes obfuscate statistical differences at the state level
- No geographic cluster patterns

Supplementary slide - Is there a relationship between match rate and annual consumption?

Electricity consumers from greatest to least mean kWh	Natural gas consumers from greatest to least mean kBtu
No match (11,266 kWh)	“Possible” match (63,219 kBtu)
“Possible” match (10,648 kWh)	“Likely” match (59,981 kBtu)
No data (10,362 kWh)	No match (59,093 kBtu)
“Likely” match (10,203 kWh)	No data (49,590 kBtu)

Supplementary slide - How often was the matching algorithm wrong?

On average, data were received but from a different supplier than originally assigned for:	
3% of “likely” electricity matches	4% of “likely” natural gas matches
5% of “possible” electricity matches	6% of “possible” natural gas matches

Supplementary slide -What would caseloads look like if we worked the easy cases first?