Dina Stretiner

Project Report

Exploring the Association between Energy Burden and County Demographics

Introduction

Energy burden is defined as the percent of gross income that a household spends on energy costs: electricity, heating and cooling. While energy burden can vary greatly depending on the type of heating fuel used and efficiency of the HVAC equipment, it can be beneficial for policymakers to step back and understand the macro trends behind high energy burdens. This is where county demographics come in.

Energy burden is important because it demonstrates the level of energy affordability. Households that forgo energy purchases to save money are energy insecure and are vulnerable to adverse health outcomes, such as hypothermia or heat stroke. Policymakers need to deploy targeted assistance to vulnerable populations to ensure equitable access to energy.

Research Question

This project will attempt to find associations between certain demographic variables, such as county population, percent minority, unemployment rate, vehicle counts and the average energy burden. Relevant population is all of United States – this is represented by data at the county level. The aim of this project is build an explanatory model between the relevant predictors and the continuous response – average energy burden by county. However, if linear model assumptions are not met, we will switch to a logistic regression model with a binary outcome: high/low energy burden.

We will perform necessary transformation to ensure that variables are roughly on the same scale, identify possible interactions with graphical methods and select the best model by removing statistically insignificant variables and comparting model performance measures (R^2, AIC).

Results – Linear Model

The data set was obtained from the National Renewable Energy Laboratory Website and is titled "Equitable Energy Investment Prioritization Data Set" (2021). Its goal is to provide metrics around renewable energy deployment potential and energy justice. The former consists of estimates with a high degree of uncertainty and is not directly applicable to our project. Below are variables relevant to our project with the year(s) listed:

1. **State**, *categorical*
2. **County population**, *continuous*, 2013 - 2017
3. **Reference Vehicle Counts**, *continuous*, 2018: battery electric vehicle, hybrid electric gasoline vehicle, plug-in hybrid electric vehicle, internal combustion engine vehicle
4. **Energy Burden**, *continuous*, 2018: energyburden_1_prop – proportion of census tracts within each county with less than 4% energy burden, energyburden_2_prop – between 4% and 7% energy burden, energyburden_3_prop – between 7% and 10% energy burden, energyburden_4_prop – over 10%
5. **Unemployment Rate**, *continuous,* 2020
6. **Percent of work force employed in mining, quarrying or oil and gas**, *continuous*, 2015 – 2019
7. **Rural-urban continuum code**, *categorical*, 2013: codes 1 through 9 – represent urban and rural counties with a certain population, the higher the code the more rural the county is.
8. **Farming-dependent county code**, *categorical*, 2015: yes/no
9. **Population loss county code**, *categorical*, 2015: yes/no
10. **Persistent Poverty County Code,** *categorical*, 2015: yes/no
11. **Minority Indicator**, *continuous,* 2013 – 2017, (minority_4_prop - .20) + (minority_5_prop -.20)*1.7
12. **Low Income Indicator,** *continuous,* 2013 – 2017, (lowincome_4_prop - .20) + (lowincome_5_prop - .20)*1.5
13. **Less than High School Indicator,** *continuous,* 2013 – 2017, (lessthanhs_4_prop - .20) + (lessthanhs_5_prop -.20)*2.1
14. **Cancer Indicator** (environment hazard), *continuous,* 2013 – 2017, (*_5_prop - .20)

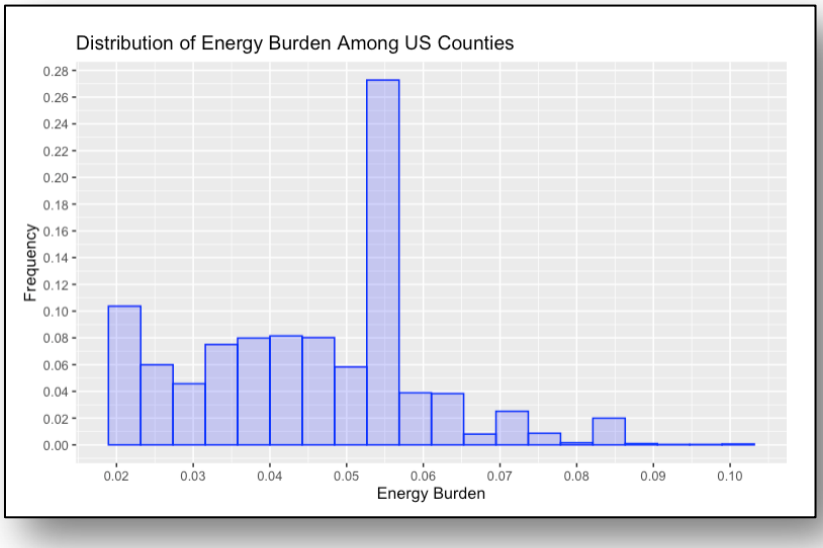
    ***_4_prop**: Proportion of census block groups within each county that fell within the 4th national quintile for the specified demographic variable or environmental hazard indicator (60-80th percentile)

    ***_5_prop**: Proportion of census block groups within each county that fell within the 5th national quintile for the specified demographic variable or environmental hazard indicator (80-100th percentile)

These variables represent historical data that was compiled by NREL from "Low Income Energy Affordability Data (LEAD)", "Atlas of Rural and Small-Town America" and US Environmental Protection Agency. It should be noted that the data comes from different

years, which raises questions about consistency. We found that to be a minor concern because most of the variables would not change that much in a span of four to five years.

The first step of the analysis was creating the response variable of energy burden. It involved multiplying proportion of census tracks to the midpoint burden in the relevant bucket. For example, if 54% of census tracts in the county had burden 0-4%, 30% had 4-7%, 10% had 7-10% and 6% - over 10%, we would calculate the overall weighted average energy burden as follows: 0.54 * 2% + 0.3 * 5.5% + 0.1 * 8.5% + 0.06 * 10%. This yielded the following distribution:



75% of the counties ended up having estimated average burden less than 5.5%. This number was also the most frequently occurring value (as seen on the graph), which was not surprising, given the nature of our weighted average calculation.

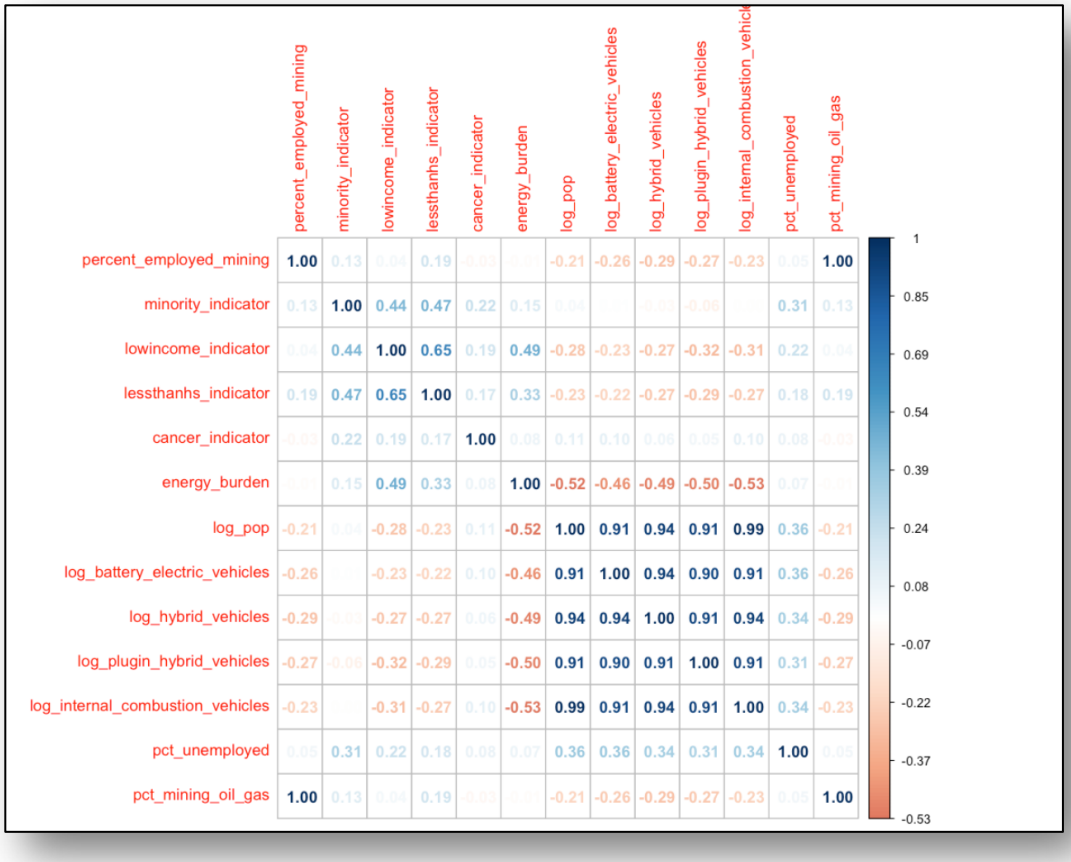
Numerical summaries of other variables revealed a need for transforming vehicle counts and population variables because of difference in scales and outliers.

Unemployment rate showed a 75th percentile of 8% with a maximum of 22.5% - a clear outlier.

```
  county_pop       bev_2018_reference_vehicle_counts hev_gasoline_2018_reference_vehicle_counts phev_2018_reference_vehicle_counts
Min.   :     102   Min.   :     0.0                 Min.   :     0                               Min.   :     0.00
1st Qu.:   11204   1st Qu.:    24.0                 1st Qu.:   153                               1st Qu.:    12.00
Median :   25878   Median :    75.0                 Median :   387                               Median :    37.00
Mean   :  103199   Mean   :   661.5                 Mean   :  2068                               Mean   :   311.86
3rd Qu.:   67371   3rd Qu.:   209.8                 3rd Qu.:  1104                               3rd Qu.:    99.75
Max.   :10098052   Max.   :147766.0                 Max.   :376448                               Max.   :107281.00
```

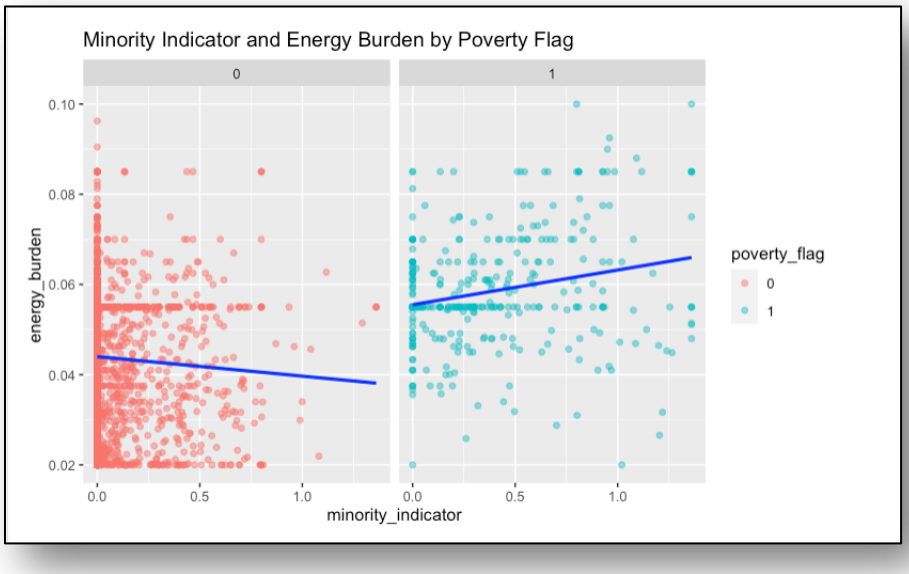
```
icev_gasoline_2018_r minority_indicator lowincome_indicator lessthanhs_indicator cancer_indicator unemprate2020
Min.   :     99      Min.   :0.00000    Min.   :0.0000      Min.   :0.0000       Min.   :0.00000   Min.   : 1.700
1st Qu.:   9455      1st Qu.:0.00000    1st Qu.:0.0000      1st Qu.:0.0000       1st Qu.:0.00000   1st Qu.: 5.200
Median :  21416      Median :0.00000    Median :0.1381      Median :0.1311       Median :0.00000   Median : 6.500
Mean   :  74209      Mean   :0.10723    Mean   :0.2115      Mean   :0.2747       Mean   :0.07696   Mean   : 6.706
3rd Qu.:  53762      3rd Qu.:0.09753    3rd Qu.:0.3389      3rd Qu.:0.4429       3rd Qu.:0.00000   3rd Qu.: 8.000
Max.   :6463113      Max.   :1.36000    Max.   :1.2000      Max.   :1.6800       Max.   :0.80000   Max.   :22.500
```

A correlation matrix helped assess the strength of bivariate relationships.



Log of population and vehicle counts were well correlated with energy burden, but also correlated with each other, which raised concerns about multicollinearity.

Interaction relationships between continuous and categorical variables were assessed via the difference in slopes when graphing against energy burden. Two interaction terms were deemed viable: minority indicator x poverty flag and less than high school indicator x population loss flag (although linear relationships with energy burden appeared weak).

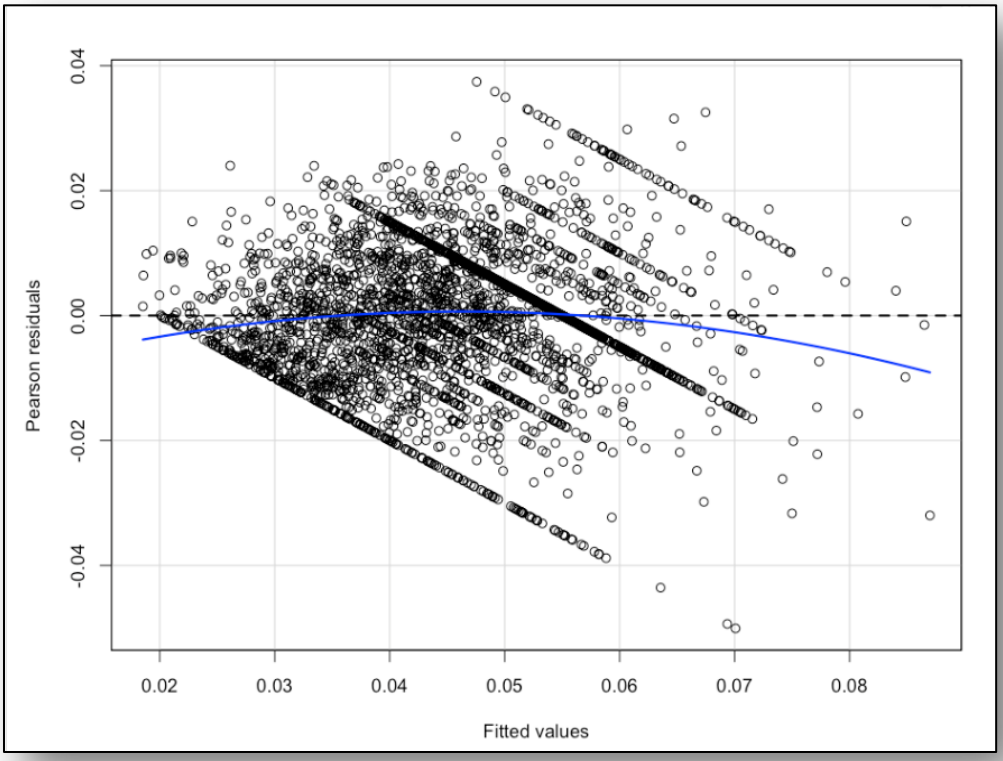


Original linear model, including all predictor variables plus interactions regressed upon the average energy burden revealed significant multicollinearity, where VIF for log population was 167 and VIFs for log vehicle counts were above 10. Additionally, when graphing energy burden against log population and vehicle counts, it became obvious they had similar relationships. After multiple steps of trying different variable combinations, it was determined to remove all vehicle counts and only keep the log population variable, so that VIFs stay in a reasonable range.

	GVIF <dbl>	Df <dbl>	GVIF^(1/(2*Df)) <dbl>
state	32.962126	48	1.037081
log_pop	3.308895	1	1.819037
pct_unemployed	2.450803	1	1.565504
pct_mining_oil_gas	1.676760	1	1.294898
rural_urban_flag	4.278368	8	1.095103
farming_flag	1.808307	1	1.344733
pop_loss_flag	4.048427	3	1.262451
poverty_flag	3.426759	3	1.227856
minority_indicator	3.426759	3	1.227856
lowincome_indicator	2.788414	1	1.669855

To select the final model, we also removed statistically insignificant variables (farming flag, cancer indicator) and ran a "backwards" iteration process to find model with the lowest AIC (incorporates a penalty for additional variables, which do not add new information). The lowest ("best") AIC model omitted variable state and resulted in R^2 of 47.35%. It was surprising, given that state moderates geographic correlation between adjacent counties. State-wide policies can also affect the level of energy burden.

The final step of the process was checking assumptions for a linear model. Unfortunately, the residual plot revealed a negative relationship between fitted values and residuals – an indication of significant bias in the model. We went back to some earlier versions of the model (with state included, with new interactions) and still obtained the same results. This is not surprising, given that our continuous energy burden response variable was artificially manufactured with a mode of 5.5%. In this case, if the model predicts a low energy burden (like 3%), it will have a high residual, because most counties are biased towards 5.5%. For this reason, we decided to move onto logistic regression.



Results – Logistic Regression

Logistic Regression is used to predict a binary outcome: yes/no. In this type of model, the log odds of the outcome being 1 have a linear relationship with the predictor variables.

$$ln(\frac{p}{1-p}) = X\beta$$

Since our data had percent of census tracts having a certain energy burden, it was convenient to calculate a binary response of "Average Energy Burden >=4%". 4% was

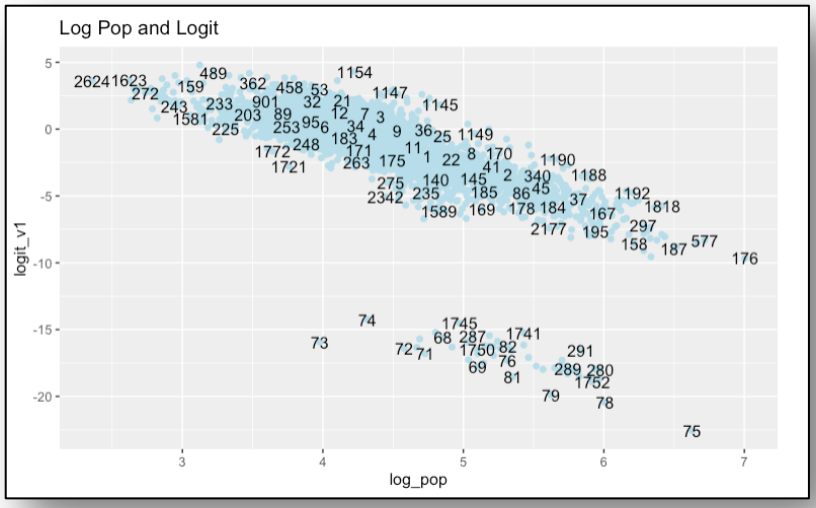
energy_burden_greater_4 <chr>	cnt <int>	pct <dbl>
< 4%	2004	0.65
>= 4%	1101	0.35

determined to be a reasonable benchmark for high vs low, because national median energy

burden is 3.1% (Drehobl, 2020). In the table above, one can see that there is a sufficient sample of "highly burdened" counties (35%) to run a credible model.

Logistic regression has fewer assumptions to meet than a traditional linear model: (1) linear relationship between the logit (log odds of 1) and the predictor variables, (2) no significant multicollinearity, and (3) no influential outliers. We started with a full model, including all variables and hypothesized interactions, except for vehicle counts, because we already determined their correlation with population. The output showed the following variables as statistically insignificant based on alpha level = 0.05: less than hs indicator x population loss flag, minority indicator x poverty flag, cancer indicator, poverty flag, population loss flag, farming flag and percent employed in mining and oil and gas.

Influential outliers can be determined based on Cook's Distance. It measures the model change when one of the observations is removed. To find extreme outliers, we searched for counties with 10x the mean Cook's Distance, this resulted in 33 observations. Additional outliers were observed in the graphs between continuous predictors and the logit, such as the graph below.



Upon closer examinations these bottom outliers were counties with high measures in more than one variable. For example, Santa Cruz, AZ had high unemployment, high minority indicator, high less than high school indicator and high low income indicator. It was determined to remove the high Cook's D observations and certain outliers spotted in the graphs (there was some overlap there). Below is the final model, without the statistically insignificant variables and influential outliers. We rechecked all logistic model assumptions to a satisfactory degree.

$$Odds\ of\ Energy\ Burden\ Being >= 4\% = EXP\ [Coef * I(State \neq AL) - 2.839 * LogPopulation + 18.25 * PctUnemployed + 0.08429 * I(RuralUrbanFlag = 2) + 0.34225 * I(RuralUrbanFlag = 3) - 0.15090 * I(RuralUrbanFlag = 4) - 2.39445 * I(RuralUrbanFlag = 5) + 0.90670 * I(RuralUrbanFlag = 6) + 0.79336 * I(RuralUrbanFlag = 7) + 1.32809 * I(RuralUrbanFlag = 8) + 0.80638 * I(RuralUrbanFlag = 9) - 1.7414 * MinorityIndicator + 2.06682 * LowIncomeIndicator + 0.53167 * LessThanHSIndicator]$$

Because this is an explanatory, not a predictive model, we did not put much effort into appropriately evaluating model performance. Accuracy of 80% and F1 score of 84.8% were obtained based on the whole data set, because no train/test split was performed. As such, these measures are not entirely reliable, because the model is overfit.

Interpretation of the model is facilitated by taking a natural log of the coefficients and the confidence intervals. For categorical variable state, Alabama was used as a reference group. Based on the model, if a county is in Colorado, the odds of it having average energy burden higher than 4% is 81% lower than for a county in Alabama. 95% confidence interval tells us that the true odds for a CO county are between 49% lower and 93% lower than AL. If log10 population increases by 1, then the odds of a county having high energy burden decrease by 94.1%. 95% confidence interval shows that the true odds fall between 91.4% lower and 96% lower. If unemployment rate goes up by 0.01 (1%), then the odds of a high energy burden go up by a factor of 1.2. 95% confidence interval spans a factor of 1.12 and

1.29. Compared to a metro county with a metro population of 1 million or more (baseline, urban flag = 1), a metro county with a metro population of less than 250K (urban flag = 3) residents has 40% higher odds of having average energy burden over 4%. 95% confidence interval spans 17% lower odds and 136% higher odds. When minority indicator goes up by 1, the odds of having a high energy burden go down by 83%. 95% confidence interval includes 67% lower odds and 91% lower odds. This interpretation applies because of the negative coefficient associated with minority indicator, which is a counterintuitive result, not consistent with broader literature. Perhaps, this is happening due to a missing variable, or this is an expected result for some minority groups, but not for others. When low income indicator goes up by 1, the odds of the county having energy burden over 4% go up by a factor of 7.9. 95% confidence interval spans factor of 4 to factor of 15. Finally, when less than high school indicator goes up by 1, the odds of the county having energy burden over 4% go up by a factor of 1.7. The 95% confidence interval spans factor 1.1 to 2.6.

Conclusions

Based on the model interpretation, we can conclude that overall, rural counties have higher odds of high energy burden, compared to urban counties, this is also consistent with the log population result. Energy burden varies by state, however most states have lower odds of a higher energy burden than Alabama. Low income indicator was especially prominent in generating higher odds of energy burden over 4%. Low education levels and high unemployment rate also play a significant role. Overall, under-privileged communities in rural areas fare the worst in energy affordability.

Before we discuss the implications of these findings, it's important to understand the project limitations. As touched on before, this data contains spatial correlation because

it's at the county level – where adjacent counties may share similar characteristics. This is partially corrected by the variable state. To go a step further, we could have ran an unsupervised learning algorithm to identify clusters of similar counties and added that as a predictor variable. Indicator variables (e.g., low income, minority, less than high school) represent calculated scores based on the percent of census tracts that were within the 60-80th and 80-100th percentiles of the national results. Basically, they indicate high concentrations of disadvantaged residents. However, they are difficult to interpret, so it would have been more practical to include actual percentages of "x" type populations. Finally, an appropriate train/test split would provide better insight about model performance and assist in the best model selection.

To focus on energy affordability, policymakers need to address equipment efficiency, bill subsidies and assistance to rural areas. As established earlier, Alabama is an especially vulnerable state when it comes to energy burden. Low income households in AL and a handful of other states in South Easter US use 36% more electricity than those nationwide (US Department of Energy, 2018). This is because cooling is more energy-intensive than heating, especially if households use inefficient equipment. It turns out that having three window A/C units uses more energy than a central A/C, which can cost up to $5,600 to install.

People in colder climates also need efficiency solutions. The best answer so far is a heat pump, which doubles down as an A/C in the summer. It moves energy from one space to another, instead of generating heat or cold. The sticker price for one is ~$5,500, however it reduces down to ~$4,000 after rebates from the Inflation Reduction Act (Glavinskas, 2022). Both heat pump and central A/C are expensive investments, not available to the

most vulnerable populations – low income renters. One way to help is through legislative changes to the Landlord-Tenant Act, which would require landlords to not only provide functional, but also energy efficient HVAC equipment. These upgrades should be subject to rebates to incentivize landlord compliance.

Direct bill subsidy is accomplished through a federal program called Low Income Home Energy Assistance (LIHEAP). Requirements vary from state to state. In CO, for example, one needs have gross income below 60% of the State Median to qualify. There is a separate application and subsidies are paid directly to the energy provider (Colorado Department of Human Services, 2022). The administration of the program is inefficient because it exists outside of other social safety net support: rental subsidies, Medicaid, food stamps, etc. Low income households need multiple forms of assistance – it makes sense to bundle programs together and approve energy subsidies concurrently with housing vouchers, for example.

Rural areas pay more for electricity because the housing stock is more spread out. On top of that, 70% of all manufactured housing, which is notorious for its poor insulation, is situated in rural areas. Department of Energy provides a Weatherization Assistance Program aimed at improving insulation, but it's poorly communicated (Ross et al, 2018). One obvious solution is to improve the quality of manufactured housing at the point of production. It would be helpful for the Department of Energy to conduct energy audits and report the results back to the housing manufacturers.

While large cities rely on investor-owned utility companies (IOUs), such as Xcel Energy, rural areas depend on distribution co-ops for electricity. These entities purchase wholesale electricity either from IOUs or Generation & Transmission Associations (G&Ts)

(Ross et al, 2018). Rural co-ops are often locked into long-term contracts with G&Ts and the exit procedures can be cumbersome and costly. For example, in the last five years, some Colorado and New Mexico co-ops parted ways with the Tri-State Generation and Transmission Association. One of the points of contention was restrictions related to self-generated power and installing energy storage. Relationships between G&Ts are rural co-ops are governed by the Federal Energy Regulatory Commission (FERC), however there is lots of red tape and friction (Powering Cooperatives, 2019). One solution is for individual states to advocate on behalf of rural co-ops at the federal level.

NOTE: Built a KNN model to predict energy burden at the county level – see Jupyter notebook

Sources

Day, Megan; Ross, Liz (2021): Equitable Energy Investment Prioritization Data Set. National Renewable Energy Laboratory. https://data.nrel.gov/submissions/175

US Department of Energy, Office of Energy Efficiency & Renewable Energy (2018). Low-Income Household Energy Burden Varies Among States – Efficiency Can Help in All of Them. https://www.energy.gov/sites/prod/files/2019/01/f58/WIP-Energy-Burden_final.pdf

Glavinskas, Vanessa (2022). Environment Defense Fund. 8 ways the Inflation Reduction Act can save you money. https://www.edf.org/article/8-ways-inflation-reduction-act-can-save-you-money

Colorado Department of Human Services (2022). Low Income Energy Assistance Program. https://cdhs.colorado.gov/leap

Ross, Lauren; Drehobl, Ariel; Stickles, Brian (2018). American Council for Energy Efficient Economy. The High Cost of Energy in Rural America. https://www.aceee.org/sites/default/files/publications/researchreports/u1806.pdf

Powering Cooperatives, A Policy Primer (2019). CSU Center for the New Energy Economy. https://cnee.colostate.edu/wp-content/uploads/2019/03/Powering-Cooperatives-CNEE-Report-on-Colorado-Cooperatives-and-TriState.pdf