Energy poverty occurs when a household experiences inadequate energy services for one’s dwelling place (Boardman, 1991). In Europe, there is understanding amongst researchers that energy poverty is a result of high energy prices, low incomes, inefficient buildings, and individual household practices and needs (Boardman, 2010; Thomson, Bouzarovski, & Snell, 2017). Energy poverty is especially acute in Texas where, like many other states, residential electricity prices have increased over the past two decades and electricity demand has increased due to increasingly severe weather (Yun & Steemers, 2009; DOE, 2016; Wible & King, 2016). These trends often create a compounded burden for vulnerable individuals living in older, less energy efficient housing (Valenzuela, et al., 2014).

The most common way of measuring energy poverty is energy burden; individuals who are living in energy burden spend more than a certain percentage of their income on energy bill. Typically people who spend more than twice the median amount of their income on energy. For example, the 4% is the median amount households spend on energy bills in Texas. Therefore, 8% is the threshold at which a household is considered energy burdened…..Roughly 22% of Texas households are energy burdened (Wible & King, 2016).

One study estimated that individuals living in rural counties along the Texas-Mexico border may be spending 18.9% of income on energy bill, compared to the 4% state median (CITE OLMEDO, 2013). …….Measuring energy burden at a regional level is necessary to inform legislators and policy makers of how widespread the issue is, the causes of the issues, and how much funding to allocate to programs. While literature by European scholars have used both qualitative methods to analyze the regional, social, and economic indicators of energy poverty, much of the quantitative literature in the U.S. focused on energy efficiency. ………….

Recent studies have also shown a link between health & food insecurity and energy burden (Tuttle & Beatty, 2017). Since geography is a strong predictor of these disparities (Bouzarovski & Simcock, 2017). Thus, the objective of the study is twofold. First, this study investigates how much of energy burden can be described by socio-demographic, economic, and health indicators at the county level. Secondly, this study seeks to investigate whether being a border county is a significant predictor of energy burden.

The results of this study hold implications for policy makers on what regions to focus energy burden programs. Additionally, what qualified…….….Findings show that socio-demographic, economic, and health indicators explain nearly half of the county level variance in energy burden. Results also show that border counties are a significant predictor of energy burden..

**Literature Review**

Over the past decade there has been increasing evidence that socio-demographic, geographic, and economic indicators are causes of energy poverty in the United States. Hernandez and Bird interviewed low income individuals in Boston revealing that lack of financial resources, housing instability, health issues, and energy inefficiency compounded into a heightened burden for low income households (CITE)…..,and….For instance, the authors assert residents who have critical health conditions are more likely to have higher energy bills and individuals who cannot afford energy bills are likely to suffer from extreme weather conditions (CITE).

The authors of that study also posit that individuals living in homes they rented as opposed to owned homes, are less likely to be able to access energy poverty alleviating program like energy efficiency because their landlord must make the decision to make structural upgrade (i.e. principal agent problem) (Bird & Hernández, 2010). However, the authors do not provide any empirical evidence quantifying how many individuals may be facing this issue or its quantifying its impact on energy poverty. Similar to Hernandez and Bird another study posits that energy poverty should be thought of as a network of factors particularly socioeconomic, economic, regional, and structural factors (Bird & Hernández, 2010) (Harrison & Popke, 2011). The study used case study interviews in eastern North Carolina and furthermore the focus was on heating bills. However, there was no quantitative……

More recently, another study used census blocks in Kansas City to identify socio-demographic and economic predictors of energy inefficiency (Reames, 2016). The results showed that census block level poverty percentage, percentage of individuals over the age of 65, and race/ethnicity as predictors of energy inefficiency (Reames, 2016). However, the study didn’t include health indicators or address regional variation in energy burden. ……. have…………

Finally, on a national scale, an ACEEE 2016 report showed that amongst the largest XX cities in the U.S. the percent of income varied XXX from city to city (CITE).

………………

Finally in Texas, Olmedo show that in Colonias ( informal settlements within XXX miles of the Texas-Mexico border) individuals may be spending 18-19% of income (CITE TEPRI std)

This study seeks to fill the gap in energy poverty research by providing an empirical quantitative analysis to investigate the relationship between energy burden (a metric of energy poverty) and social indicators. Specifically, we linear OLS model to assess the power of selected economic, socio-demographic, economic, health, and geographic (i.e. Texas-Mexico border counties) to explain low-moderate income (LMI) energy burden.

**2. METHODS**

The methods for this study include (1) data cleaning and merging, (2) regression models on a analysis and validation data set that mirror each other, (3) regression robustness checks and (4) GIS visualization. The general work flow of the analysis and validation models is shown in Appendix XXXX

**Analysis and Validation Regression Methods**

The relationship between energy burden and socio-demographic, health, and economic indicators are tested using a linear regression. Linear model was chosen because all of the variables in the study are continuous, the aim is to understand the relative impact of the variables on energy burden…. (i.e. compare magnitude of the coefficients), and finally linear models have high interoperability (CITE ISLR txtbk). Median age of the houses built in each county is included as a control variable to represent relative energy efficiency (CITE).

The independent variables for the study are shown in the table below( FIG XX).

In order to test the level of variation in energy burden that is attributed to region, a dummy variable was included to represent whether or not a county was a border county. Here we use the TX DHS definition of a border county: a county that is with 69 miles of the Texas-Mexico variable (CITE TX DHS). 32 counties are considered border counties (Fig XXX).

In order to validate the results, a dataset with the same independent variables but from different data sources than the analysis set is used. The same was done for the linear model with the dummy variable. A full list of the data sources can be seen in the next section (TABLE/FIGUREXX).

**Data Sources:**

The data used for this study comes from:…….. a list of data sources for each variable is listed below (TABLE/FIGURE XXX).

Most of the variables were a percentage on the scale of 0-100. This was done to ease interoperability of coefficients. The only variables that weren’t on a 0-100 % scale were the border county dummy variable and the median age of household in a given county.

**Research Workflow:**

The workflow of the study is as follows: (1) data downloaded and saved, (2) data sets are cleaned merged to create an analysis and validation data set, and (3) linear models were created iteratively. The work flow for each of these stages are shown below (FIGURE XX)

The data and documentation were downloaded from their respective website and saved into folder on a desktop that was linked to a Github repository. This raw data was then loaded into a R script that was used to clean the data. In the script, the datasets that included observations other than Texas were removed. Additionally, irrelevant variables were deleted. The variables that were relevant were renamed.

Most data sets included a unique county ID number called the “Federal Information Process ID (FIP)” and/or the county name. The FIP was used as the unique identifier to merge that data sets in to two new data sets: an analysis data set and a validation data set. All continuous variable were changed to the datatype “numeric”.

The same procedure described above was used to create a the datasets with the border region dummy variable: analysis data with a dummy and validation data set with a dummy. The dummy variable “brdr\_cnty” was created using an if statement where is if the name of the county and FIP matched the DHS list of border counties, brdr\_cnty=1. Otherwise, the variable brdr\_cnty=0.

The models were run in a script written specifically for linear modelling. After a regression was run a summary table with Adj. R2 , F statistic, and p-values were developed to assess explanatory power of the model and significance of variables. Robustness check were done to check for typical issues linear models have including: linearity between energy burden and each independent variable, multicollinearity, heteroscedasticity, error dependence were checked. In order to reduce these typical linear model issues, variables were removed in an iterative process.

The map visualizations were done using two datasets that were joined based on FIP. One dataset was a .shp file and the other was a .csv. The datasets were uploaded to ArcGIS and joined in ArcGIS.

**Data Management Methods**

Data management occurred through: (1) saving the entire study in two locations, (2) folder structuring, (3) file and variable naming conventions. All of the files used for this study where saved in a publicly available Github repository cloud. This repository was linked to the authors desktop and the cloud was updated after any changes were made to any of the files. The folder structure is pictured below (FIG XX). The folders were named after major functionalities: (1) data (raw data and clean data), (2) clean and merge code, (3) regression code, (4) outputs, (5) mapping, (6) writing, literature and presentations.

Within these folders there was a “README” file that described each file in the folder. For folders that had more folders within them, there was another “README” within that folder describing the files. For example, the raw data folder had folders in it that represented each data source (e.g. the raw data from USDA was inside a folder called “USDA”), and each of these folders had their own “README” that described the dataset, listed the URL, and date accessed.

In addition to the raw datasets having their own folders, documentation from the websites were downloaded as well. For instance, the County Health Ranking dataset had documentation explaining the meaning and units for each variable. Some websites had better documentation was better than others.

File naming convention was based on functionality and drafts.

**Results:**

**Discussion and Conclusion:**

**Limitations:**

**Replication Limitations**

Overall, this study is highly replicable because the datasets are all publicly available and R and R Studio are publicly available programming software. The only software that is not publicly available is ArcGIS which was used to create the maps. Additionally, the literature that was used in the literature and to develop this study, is not publicly available. As such, validating the assertions made in the introduction and literature review will be difficult for researchers without access to an academic library.

**Study Limitations:**

The study was done at a county level, which does not have the same statistical power as a study done at the census tract level (i.e. here n=254 compared to census tract level where n>>1000). Additionally, the variables used in this study were primarily limited to variables where more than one source for the data was found. The response variable, county median LMI energy burden, was only available from one data source which could lead to validation issues.