Rexon Carvalho
Final Project

Analysis of Regional Energy Consumption Survey (RECS) Dataset

Introduction

The rapidly growing world energy use has already raised concerns over supply difficulties, exhaustion of energy resources and heavy environmental impacts (ozone layer depletion, global warming, climate change, etc.). The global contribution from buildings towards energy consumption, both residential and commercial, has steadily increased reaching figures between 20% and 40% in developed countries, and has exceeded the other major sectors: industrial and transportation. Growth in population, increasing demand for building services and comfort levels, together with the rise in time spent inside buildings, assure the upward trend in energy demand will continue in the future [(4)].

The residential sector in the United States of America consumed about 21% if the total primary energy consumption [(3)]. In 2015 alone the residential sector used 20869 Trillion BTUs of primary energy. According to EIA the average energy consumption of a home in United States is decreasing. The homes are becoming larger and still consuming less energy. Having said that, the total energy consumption of the residential has been almost unchanged. In order to address this ambiguity, the RECS data set was considered for analysis. The RECS data set is a dataset which represents the US population and its characteristics related to energy consumption fairy well, as its samples represent the population.

This analysis includes some general trends in the characteristics like variation in consumption of energy as the year in which a housing unit was built changes, variation in energy consumption patterns with respect to where a housing unit is located and the change in the average square footage of housing units with time. It also includes a time series analysis of number of housing units built each year and uses the time series model to forecast the trend in housing units to 2040. Moreover, it also includes a regression analysis to model the total energy used for space heating.

Information About the Dataset [(1)]

Regional Energy Consumption Survey (RECS) is a periodic survey which is sponsored by the U.S. Energy Information Administration (EIA). RECS provides detailed information about energy usage in U.S. homes. RECS consists of a Household Survey phase, data collection from household energy suppliers, and detailed consumption and expenditures estimation. The Household Survey collects data on energy-related characteristics and usage patterns of a nationally representative sample of housing units. For renters that do not directly pay for their energy usage, a supplementary Rental Agent Survey is conducted. The Energy Supplier Surveys (ESS) collect data on how much electricity, natural gas, propane/LPG, fuel oil, and kerosene were consumed in the sampled housing unit during the reference year. It also collects data on actual dollar amounts spent on these energy sources. EIA uses a non-linear statistical model to produce consumption and expenditures estimates for heating, cooling, refrigeration and other end uses in all housing units occupied as a primary residence in the United States. Originally conducted with paper and pencil, RECS now uses a combination of Computer-Assisted Personal Interview (CAPI), internet, and mail to collect data for the Household and Energy Supplier Surveys.

RECS samples homes occupied as a primary residence, which excludes secondary homes, vacant units, military barracks, and common area in apartment buildings. RECS estimates, therefore, do not represent sector level estimates, but are best suited for comparing across different characteristics of homes within the residential sector. EIA instituted the following survey design revisions, content changes, and variable updates for the 2009 RECS:

- A nearly threefold increase in the sample size. This allowed EIA to release estimates for household characteristics and energy use for 16 States. The total number of responding households increased from 4,382 in the 2005 RECS to 12,083 in the 2009 RECS.
- A new sample frame development methodology. The majority of the sample frame for the 2009 RECS was constructed using the U.S. Postal Service mail address database, the Delivery Sequence File (DSF).
- Expanded data collection on the type and usage of consumer electronics, including televisions and related devices, computers, and personal electronic devices.
- Finer resolution on age of appliances to align with benchmark years for efficiency standards. The "10-19 years old" equipment and appliance age range from previous surveys was split into two responses; it is now split into two age groups, 10-14 and 15-19 years.
- New data items for recent energy efficiency actions taken by the household including caulking, weather-stripping, insulation, and home energy audits.
- Standardized, Internet-based data collection for the Energy Supplier Surveys that allowed companies to respond via online forms, structured spreadsheet, or paper forms via mail.

The final sample of housing units was randomly selected from the housing unit list constructed from the selected area segments. This type of design is called multi-stage area probability sampling. Its proper application ensures that the selected sample statistically represents the entire population of occupied housing units in the United States.

Primary Sampling Units	430
Segments	3898
Sample Size	12083
Number of Variables	943

Table 1: About the Dataset

Questionnaire content included the type and number of energy-consuming devices, usage patterns, structural characteristics of the home, household demographics, and energy supplier information. The sections of the questionnaire were as follows:

- Section A: Housing Unit Characteristics
- Section B: Kitchen Appliances
- Section C: Home Appliances and Electronics
- Section D: Space Heating
- Section E: Water Heating
- Section F: Air Conditioning
- Section G: Miscellaneous (lighting, windows and efficiency-related improvements)

- Section H: Fuels Used
- Section I: Housing Unit Measurements (when permission was provided)
- Section J: Fuel Bills
- Section K: Residential Transportation
- Section L: Household Characteristics
- Section M: Energy Assistance (for a subsample of housing units)
- Section N: Scanning of Fuel Bills

Basic Statistics

The following table (Table 2) gives some information and basic statistics of the various variables used in for modeling and analysis.

Variable Name	Description	Response Code	Mean	Range	Standard Deviation
TOTALBTU	Total usage, in thousand BTU	-	90000	58-196000	54468.7
REGIONC	Census region	1-Northeast 2-Midwest 3-South 4-West	-	-	-
TOTSQFT_EN	Total square footage in Square Feet	-	2172	100-16120	1350.8
TOTALBTUSPH	Total usage for space heating in thousand BTU	-	36860	0-548700	35583.1
YREARMADE	Year housing unit was built	1600-2009	1971	1920-2009	24.8
HDD30YR	Heating days 30 year average	-	4135	0-13350	2260.5
MONEYPY	Household income	1-24	-	-	-
HHAGE	Age of householder	16-95	50	16-85	16.7
TOTALBTUCOL	Total usage for air con in thousand BTU	-	5750	0-208100	8458.8
TOTALBTUOTH	Total usage for appliances in thousand BTU	-	26870	33-428100	19867.07
TOTALBTUWTH	Total usage for water heating in thousand BTU	-	16270	0-284100	13552

Table 2: Information about variables used

Trends

Energy Consumption of Houses Built Before 2000 and After 2000:

To phot the graph of energy consumption of houses for space heating, cooling, appliances and water heating the average energy consumption of house built before 2000 and after 2000 for each category was calculated and plotted. The year 2000 was chosen because an EIA article stated that the average consumption of houses built before 2000 was more then that of the houses built after 2000 [2]

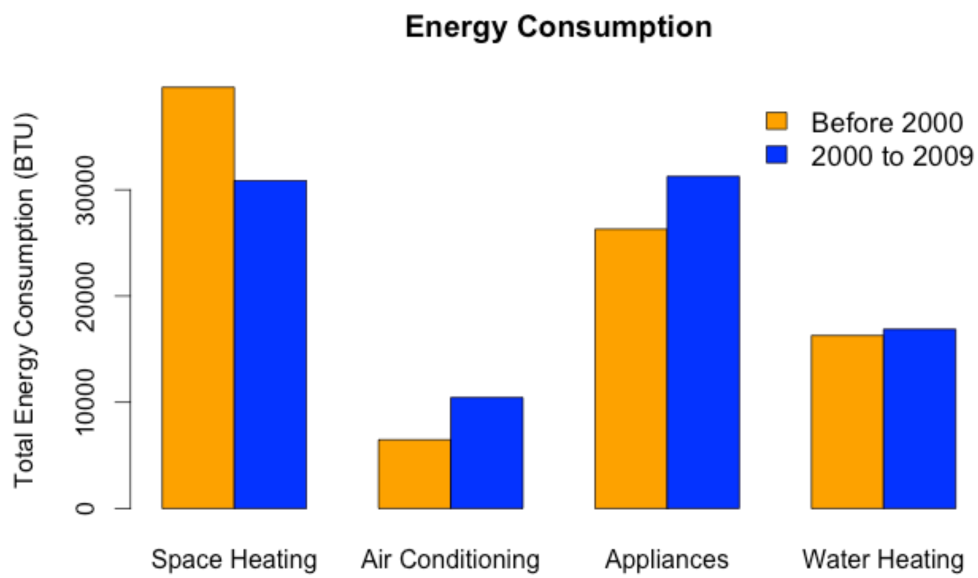

Figure 1:Energy consumption

The figure (Figure 1) shows a 22% decrease in energy used for space heating whereas, 61% increase in the energy used for cooling. Moreover, the increasing use of appliances has increased the energy used for appliances by 19%.

Area wise density plot for energy used for space heating:

The figure below, shows the density plot energy consumption for space heating for various regions. For this plot, the variable REGIONC was used to segregate energy consumption for space heating (TOTALBTUSPH) into different regions.

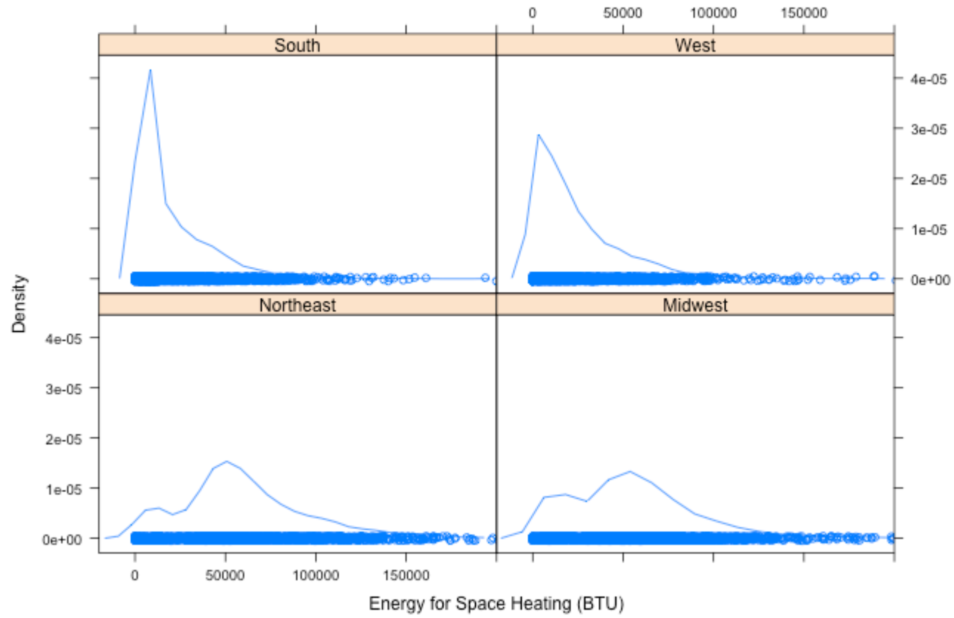

Figure 2:Density Plot of Energy for Space Heating

The plot (Figure 2) shows higher mean for Northeast and Midwest compared to South and West which is intuitive considering the fact that Northeast and Midwest are colder regions.

Energy used for space heating and total square footage over time:

To analyze the trends in energy used for space heating over time, the average amount of energy used was calculated by calculating the total energy used by total number of housing units built in a particular year and then talking a ratio. The variable YEARMADE was used to tag the samples as per years. The square footage over time was calculated in a similar fashion.

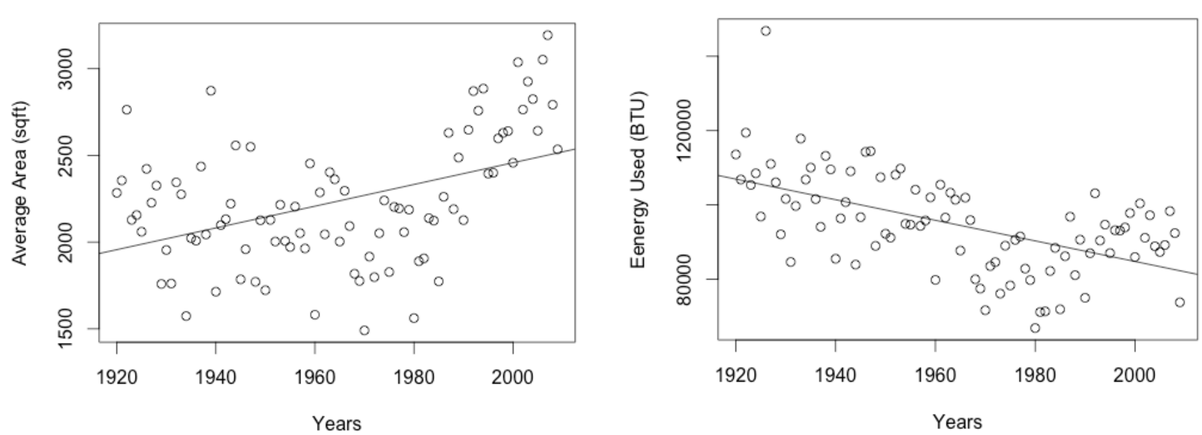

Figure 3: Trend In Area and Energy Usage from 1920 to 2009

Over the years there is a decrease in energy used for space heating in spite of there being an increase in the average square footage as shown in Figure 3.

Time Series Analysis

A time series analysis was performed to model the number of housing units built from 1920 to 2009 and to further forecast it to 2040. The time series data was generated from the RECS dataset by calculating the number of housing units built in each year. The following time series plot shows the number of housing units built each year from 1920 to 2009.

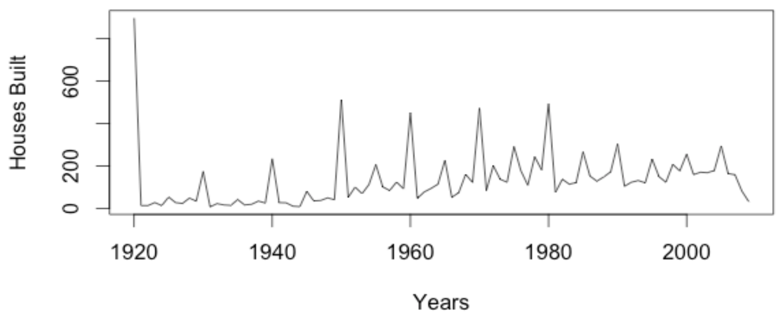

Figure 4: Time Series of Houses Built Every Year from 1920 to 2009

As there was just one data point every year, there was no data about the seasonality and hence, decomposing this time series was not possible. As the mean and variance for all the data points (processes) is not the same, it can be assumed that the time series is not stationary. Moreover, as the variance of processes is increasing initially and then decreasing it can be assumed that the time series has some trend component to it.

To further analyze the time series, the the auto correlation function and the partial auto correlation function were plotted.

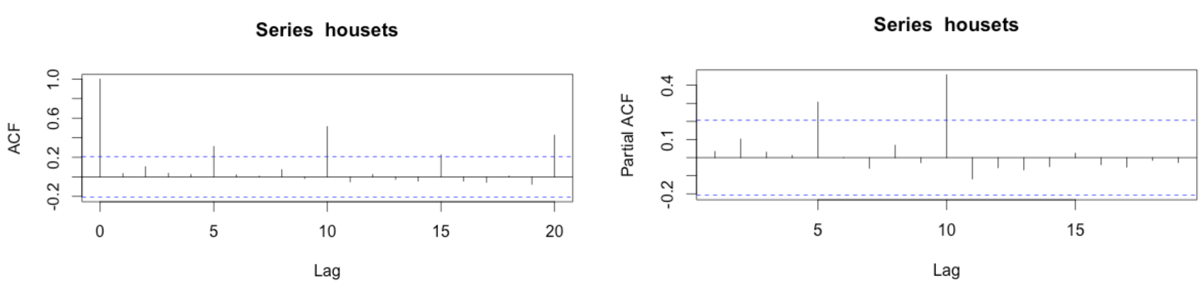

Figure 5: ACF and PACF of the time series

From the above auto correlation function and the partial auto correlation function it can be concluded that the time series is not stationary.

To model the above time series, it has to be stationary and the trend component has to be random (the processes should not follow any trend). Hence the time series has to be corrected. In order to correct the trend component, a series with difference in the processes was considered using values of lag ranging from 1 to 10. Out of all the series, the series with a lag of 1 had the and ACF exponentially tending to zero and PACF cutting off after a lag of around 10.

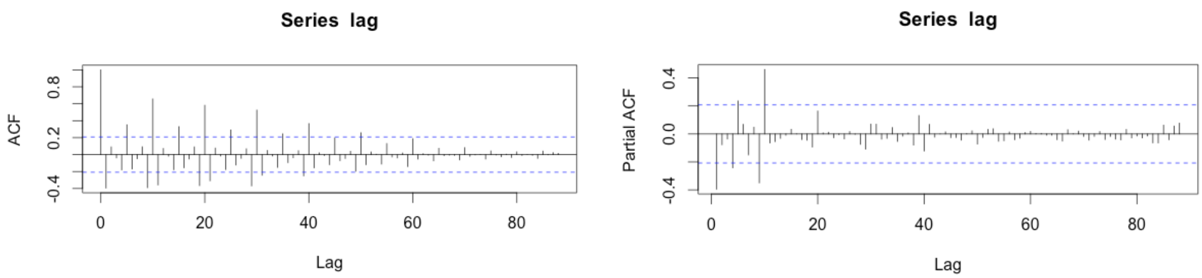

Figure 6: ACF and PACF of the time series with lag 1

From the ACF and PACF it can be concluded that the time series is stationary and autoregressive. Furthermore, taking a log of the processes corrects the increasing variance of the series as it converts the multiplicative effect to additive.

After applying the above corrections, the time series was fit into an Auto Regression Integrated Moving Average (ARIMA) model. In the following model, p is the lag for the auto regressive part and d is the lag for the moving average part.

```
m<-arima(log10(housets),order=c(p,1,d))
```

To find the best possible order the ARIMA model. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) of the model for different values of p and d both ranging from 1 to 10 were tabulated and the values of p and d for which AIC and BIC were the least were chosen. Following table shows some of the values of AIC and BIC. Rest of the values are included in the supplementary information #1.

p	**d**	**AIC(m)**	**BIC(m)**
9	0	26.1	51.0
9	1	12.8	40.2
9	2	8.2	38.1
9	3	10.2	42.5
9	4	11.3	46.1

Table 3: Order of ARIMA model

As the AIC and BIC of the model was least for p=9 and d=2, they were chosen as the values p and d. Hence the ARIMA model was as follows,

```
m<-arima(log10(housets),order=c(9,1,2))
```

To assure that this model represented the time series in a best possible way, a Ljung-Box test was performed. Ljung –Box test tests if the residuals follow any particular pattern or if they are just random noise. If the residuals are random noise, the model represents the data well. The lag for Ljung –Box test was considered to be 1 which is also the degree of freedom. The null and the alternate hypothesis for Ljung –Box test are as follows.

$$Null\ Hyopthesis: \varphi_1 = \varphi_2 = \varphi_3 = \cdots \varphi_{90} = 0 \text{ (i.e. residuals don't follow a pattern)}$$
$$Alternate\ Hypothesis: \varphi_1 \neq \varphi_2 \neq \varphi_3 \neq \cdots \varphi_{90} \neq 0 \text{ (i.e. residuals follow a pattern)}$$

After running the test, the value test statistic χ^2 was found to be 0.97. The critical value of χ^2 for 1 degree of freedom and 97% confidence interval was 5.03. As the critical value was greater than the test statistic the null hypothesis that the residuals are random noise was accepted and it was concluded that the model does not exhibit a lack of fit.

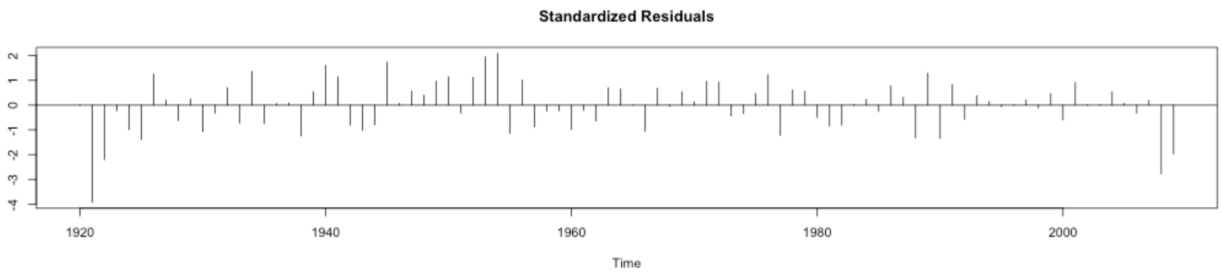

Figure 7: Residuals of processes of time series

The above plot of the residuals of the processes show that they are random.

The above model was used to forecast the data to 2040. The following graph shows the forecasted data.

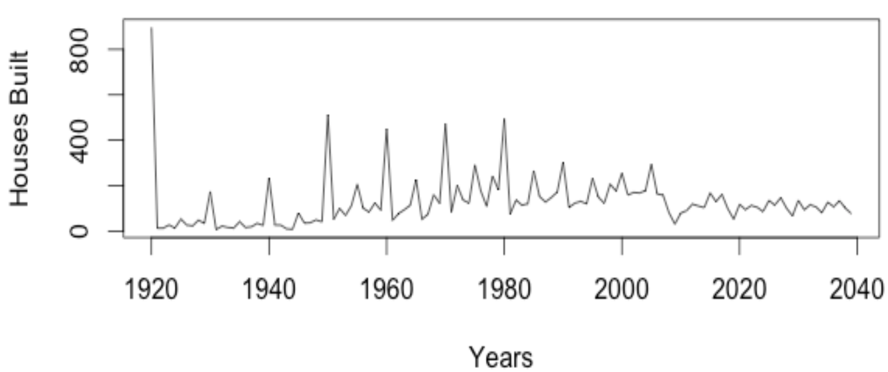

Figure 8: Forecast of time series from 2009 to 2040

The forecasted data was used to find out the cumulative number of housing units from 1920 to 2040. As shown in the plot below (Figure 9), the number of housing units will increase by 27% from 2009 to 2040.

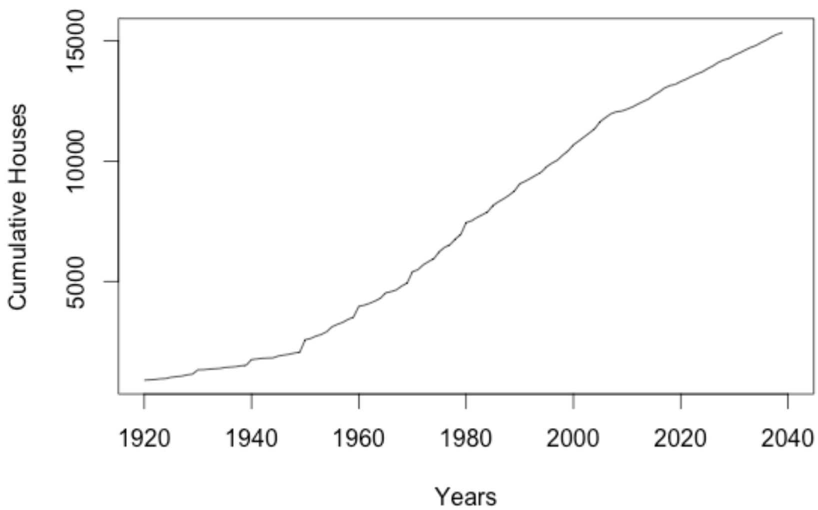

Figure 9: Cumulative Housing Units until 2040

Regression Analysis

As space heating was one of the highest consumer of energy, the energy used for space heating (TOTALBTUSPH) was modelled using regression analysis. To find out which variables had some effect on energy consumed for space heating so that an initial model could be made, a cross tabulation of some variables and energy for space heating was done. This gave the number of observation in for each segment. Tables below show cross tabulation of independent variables area, number of heating days and the year in which the house was made and the dependent variable energy for space heating.

		TOTALBTUSPH (BTU)		
		0-8748	8749-55650	55651-548700
TOTSQFT (Sqft)	100-1052	934	1548	270
	1053-2606	1262	3257	1290
	2607-15470	276	1237	1458

Table 4: Cross Tabulation of Energy for space heating and Area

		TOTALBTUSPH (BTU)		
		0-8748	8749-55650	55651-548700
HDD30YR	0-2224	1559	1045	14
	2225-5854	779	3820	1340
	5855-13350	134	1177	1665

Table 5:Cross Tabulation of Energy for space heating and Number of Heating Days

		TOTALBTUSPH (BTU)		
		0-8748	8749-55650	55651-548700
YEARMADE	1920-1975	940	2975	1988
	1976-1991	827	1490	475
	1992-2009	705	66	29

Table 6: Cross Tabulation of Energy for space heating and Year the housing unit was built

The above three tables show all the three independent variables influence the dependent variable to some extent. However, the trend is not very clear.

To model the energy used for space heating, the initial model (Model 1) made was energy for space heating as a function of area, number of heating days and year in which house was built.

Model 1: TOTALBTUSPH~TOTSQFT_EN+HDD30YR+YEARMADE

To check if household income (MONEYPY) and age of the house holder (HHAGE) have an effect on total energy for space heating, they were added to the model (Model 2) and a single parameter test was done.

Model 2: TOTALBTUSPH~TOTSQFT_EN+HDD30YR+YEARMADE+MONEYPY+HHAGE

The null and the alternate hypothesis for the single parameter test are as follows,

$$Null\ Hypothesis\colon\ \beta_{MONEYPY} = 0,\ \beta_{HHAGE} = 0$$

$$Alternate\ Hypothesis\colon\ \beta_{MONEYPY} \neq 0,\ \beta_{HHAGE} \neq 0$$

The t values for MONEYPY and HHAGE were 8.08 and 11.09 respectively, which are higher than the critical t value of 1.96 (summary of model table is supplementary information #2). Hence the null hypothesis that the variables do not have a significant effect on the total energy used for space heating was reject and model 2 was chosen.

The average energy demand for the houses built before 2000 was higher than that for the house built after. There could be a possibility that the rate for decrease was different for before 2000 and after 2000. To check check if the decrease in energy consumption was different the the variable YEARMADE was split into two parts, one houses built before 2000 (YR_2000) and the other was the houses built after 2000 (YR_2009).

The new model in which these changes were incorporated was as follows,

Model 3:
TOTALBTUSPH~MONEYPY+TOTSQFT_EN+HDD30YR+YR_2009+YR_2000+HHAGE

To conclude which of the model out of the two (model 2) or (model 3) is better a multi parameter hypothesis testing was done. The null and the alternate hypothesis for the test are as follows,

$$Null\ Hypothesis\colon\ \beta_{YEARMADE} = \beta_{YR.2000} = \beta_{YR.2009} \ldots . = 0$$

$$Alternate\ Hypothesis\colon\ \beta_{YEARMADE} \neq \beta_{YR.2000} \neq \beta_{YR.2009} \ldots . \neq 0$$

The null hypothesis states that the decrease from 1920 to 2009 was the same and hence the coefficients of all the variables are equal. The null hypothesis states otherwise. The F test statistic for this hypothesis was calculated by the following formula

$$\frac{[RegSS(reg,\ sum\ of\ square)]/J}{RSS/(N-J-1)} = F \sim F(J, N-J-1, \alpha)$$

As value of the F statistic which 8.34, was greater than the critical value of F for 12077 and 5 degrees of freedom which was 5.25 (summary and anova table in supplementary information #3), the alternate hypothesis that the decrease from 1920 to 2000 was different than the decrease after 2000 was accepted. Hence, splitting the year made made the model better.

The samples in the database are categorized into two, population living in the urban area and the population living in the rural area. There is a possibility that the people living in the urban and the rural areas behave differently and many of their characteristics may affect the energy used for space heating differently. To test if where people live affects the energy the model was split into two unpooled models by splitting the dataset into urban and rural.

Model 4 (upooled):
TOTALBTUSPH~MONEYPY+TOTSQFT_EN+HDD30YR+YR_2009+YR_2000+HHAGE,data=ds_urban

Model 5:
TOTALBTUSPH~MONEYPY+TOTSQFT_EN+HDD30YR+YR_2009+YR_2000+HHAGE,data=ds_rural

Both the unpooled models were tested against the pooled model (model 3) to find out which approach was better. The null and alternate hypothesis for the test are as follows.

$$Null\ Hypothesis:\ \beta_{pooled\ 3} = \beta_{unpooled\ 4} = \beta_{unpooled\ 5}$$

$$Alternate\ Hypothesis:\ \beta_{pooled\ 3} \neq \beta_{unpooled\ 4} \neq \beta_{unpooled\ 5}$$

The test statistic was calculated using the following formula;

$$\frac{[RSS_P - (RSS_1 + RSS_2)]/(J+1)}{(RSS_1 + RSS_2)/[(N - J_U - 1)]} = F \sim F(NR, N - J_U - 1, \alpha)$$

As the F value for the unpooled models (40.04) was greater than the F critical (4.04) for 97% confidence and 12076 degrees of freedom (summary and anova table in supplementary information #4 and #5) the alternate hypothesis that both the unpooled models are different from and better than the pooled model was accepted. Hence, splitting the dataset into urban and rural and then modelling energy consumption explains the energy consumption for space heating better.

Conclusion and Future Scope

From the analysis it can be concluded that, the houses in the USA are becoming bigger and the average energy consumption is decreasing. However, the total energy consumption does not decrease because there are more number of houses coming up every year. The energy used for space heating depends on household income, age of the householder, square footage of the house and the number of heating days. Splitting the population into urban and rural gives better models for energy used for space heating. Lastly, the number of houses in the USA will increase by 27% from 2009 to 2040.

The future work based on this model can be using time series analysis on energy required for space hearing to find a minimum threshold energy required for space heating as the energy cannot go below a certain point. An analysis of pooled models based o the four regions can be done which will give a better model for energy used for space heating.

References

1. https://www.eia.gov/consumption/residential/methodology/2009/pdf/techdoc-summary010413.pdf
2. http://www.eia.gov/todayinenergy/detail.cfm?id=9951&src=%E2%80%B9%20Consumption%20%20%20%20%20%20Residential%20Energy%20Consumption%20Survey%20(RECS)-f2
3. http://www.eia.gov/consumption/data.cfm
4. Pérez-Lombard, Luis, José Ortiz, and Christine Pout. "A review on buildings energy consumption information." *Energy and buildings* 40.3 (2008): 394-398.

Supplementary Information

1. Order of ARIMA model

p	d	AIC	BIC
0	0	115.7	118.2
0	1	75.2	80.2
0	2	76.4	83.9
0	3	76.7	86.6
0	4	70.6	83.0
0	5	66.0	81.0
0	6	66.5	83.9
0	7	72.3	92.2
0	8	60.4	82.8
0	9	62.0	86.8
0	10	53.7	81.1
1	0	89.9	94.9
1	1	76.4	83.8
1	2	78.3	88.3
1	3	76.2	88.6
1	4	70.2	85.2
1	5	66.7	84.2
1	6	68.5	88.4
1	7	72.0	94.4
1	8	61.8	86.7
1	9	55.7	83.1
1	10	49.4	79.3
2	0	86.5	94.0
2	1	78.3	88.2
2	2	79.4	91.9
2	3	63.5	78.4
2	4	43.2	60.7
2	5	41.1	61.0
2	6	70.5	92.9
2	7	31.3	56.2
2	8	65.1	92.4
2	9	58.6	88.5
2	10	51.2	83.6
3	0	88.3	98.2
3	1	80.0	92.4
3	2	80.3	95.2
3	3	45.8	63.2
3	4	44.3	64.3
3	5	71.6	94.0
3	6	61.0	85.9
3	7	74.8	102.1
3	8	56.9	86.8
3	9	55.9	88.3
3	10	50.4	85.2
4	0	55.0	67.4
4	1	56.3	71.3
4	2	53.4	70.8
4	3	38.3	58.2
4	4	55.9	78.3
4	5	31.3	56.2
4	6	31.8	59.2
4	7	31.3	61.1
4	8	33.2	65.6
4	9	23.6	58.4
4	10	15.5	52.8
5	0	55.9	70.8
5	1	57.5	74.9
5	2	60.1	80.0
5	3	25.8	48.2
5	4	55.7	80.6
5	5	30.8	58.2
5	6	32.8	62.6
5	7	34.7	67.0
5	8	34.8	69.6
5	9	36.7	74.1
5	10	15.6	55.5
6	0	55.0	72.5
6	1	33.7	53.6
6	2	46.3	68.7
6	3	32.8	57.7
6	4	28.5	55.9
6	5	32.7	62.6
6	6	33.5	65.8
6	7	28.1	62.9
6	8	32.5	69.8
6	9	32.7	72.5
6	10	17.0	59.3
7	0	46.1	66.0
7	1	46.2	68.6
7	2	44.8	69.7
7	3	27.0	54.4
7	4	20.3	50.2
7	5	21.8	54.2
7	6	21.8	56.6
7	7	13.8	51.1
7	8	14.9	54.7
7	9	18.0	60.3
7	10	21.3	66.1
8	0	48.1	70.5
8	1	48.2	73.1
8	2	32.4	59.8
8	3	26.3	56.2
8	4	28.4	60.8
8	5	23.0	57.9
8	6	21.3	58.6
8	7	14.8	54.7
8	8	16.9	59.2
8	9	18.2	63.0
8	10	20.7	68.0
9	0	26.1	51.0
9	1	12.8	40.2
9	2	8.2	38.1
9	3	10.2	42.5
9	4	11.3	46.1
9	5	13.7	51.0
9	6	16.9	56.7
9	7	14.2	56.5
9	8	14.4	59.2
9	9	14.8	62.1
9	10	15.4	65.1
10	0	7.7	35.1
10	1	9.2	39.1
10	2	10.7	43.1
10	3	11.2	46.0
10	4	10.4	47.7
10	5	11.8	51.7
10	6	13.7	56.0
10	7	15.9	60.7
10	8	15.4	62.7
10	9	17.2	67.0
10	10	19.4	71.6

2. Summary of Model 2

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.499e+05  1.958e+04   28.08  < 2e-16 ***
TOTSQFT_EN   7.334e+00  2.024e-01   36.23  < 2e-16 ***
HDD30YR      7.558e+00  1.098e-01   68.86  < 2e-16 ***
YEARMADE    -2.898e+02  9.893e+00  -29.30  < 2e-16 ***
MONEYPY      3.172e+02  3.926e+01    8.08 7.09e-16 ***
HHAGE        1.607e+02  1.449e+01   11.09  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26140 on 12077 degrees of freedom
Multiple R-squared:  0.4604,	Adjusted R-squared:  0.4602 
F-statistic:  2061 on 5 and 12077 DF,  p-value: < 2.2e-16
```

3. Summary and Anova of Model 3

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.793e+05  2.207e+04  26.249  < 2e-16 ***
TOTSQFT_EN   7.267e+00  2.037e-01  35.673  < 2e-16 ***
HDD30YR      7.560e+00  1.097e-01  68.896  < 2e-16 ***
YR_2009     -3.036e+02  1.099e+01 -27.634  < 2e-16 ***
YR_2000     -3.049e+02  1.118e+01 -27.266  < 2e-16 ***
MONEYPY      3.154e+02  3.925e+01   8.037 1.01e-15 ***
HHAGE        1.635e+02  1.452e+01  11.266  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26140 on 12076 degrees of freedom
Multiple R-squared:  0.4608,	Adjusted R-squared:  0.4605 
F-statistic:  1720 on 6 and 12076 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: TOTALBTUSPH
              Df     Sum Sq    Mean Sq  F value    Pr(>F)    
TOTSQFT_EN     1 2.3497e+12 2.3497e+12 3439.881 < 2.2e-16 ***
HDD30YR        1 3.9811e+12 3.9811e+12 5828.199 < 2.2e-16 ***
YR_2009        1 9.7601e+10 9.7601e+10  142.885 < 2.2e-16 ***
YR_2000        1 5.0362e+11 5.0362e+11  737.286 < 2.2e-16 ***
MONEYPY        1 3.0231e+10 3.0231e+10   44.258 3.003e-11 ***
HHAGE          1 8.6702e+10 8.6702e+10  126.929 < 2.2e-16 ***
Residuals  12076 8.2488e+12 6.8307e+08                       
```

4. Summary and Anova of Model 4

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.838e+05  2.315e+04   25.22  < 2e-16 ***
MONEYPY      2.700e+02  4.091e+01    6.60 4.32e-11 ***
TOTSQFT_EN   8.364e+00  2.232e-01   37.47  < 2e-16 ***
HDD30YR      8.005e+00  1.158e-01   69.10  < 2e-16 ***
YR_2009     -3.071e+02  1.153e+01  -26.64  < 2e-16 ***
YR_2000     -3.083e+02  1.173e+01  -26.28  < 2e-16 ***
HHAGE        1.717e+02  1.499e+01   11.45  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24480 on 9649 degrees of freedom
Multiple R-squared:  0.5167,	Adjusted R-squared:  0.5164
F-statistic:  1719 on 6 and 9649 DF,  p-value: < 2.2e-16


Analysis of Variance Table

Response: TOTALBTUSPH
             Df     Sum Sq    Mean Sq F value    Pr(>F)
MONEYPY       1 3.5622e+11 3.5622e+11  594.19 < 2.2e-16 ***
TOTSQFT_EN    1 1.7675e+12 1.7675e+12 2948.27 < 2.2e-16 ***
HDD30YR       1 3.4837e+12 3.4837e+12 5810.93 < 2.2e-16 ***
YR_2009       1 7.2470e+10 7.2470e+10  120.88 < 2.2e-16 ***
YR_2000       1 4.2544e+11 4.2544e+11  709.65 < 2.2e-16 ***
HHAGE         1 7.8621e+10 7.8621e+10  131.14 < 2.2e-16 ***
Residuals  9649 5.7846e+12 5.9950e+08
```

5. Summary and Anova of Model 5

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.806e+05  6.059e+04   6.282 3.95e-10 ***
MONEYPY      3.108e+02  1.074e+02   2.893  0.00385 **
TOTSQFT_EN   5.538e+00  4.780e-01  11.586  < 2e-16 ***
HDD30YR      6.547e+00  2.888e-01  22.674  < 2e-16 ***
YR_2009     -2.013e+02  3.011e+01  -6.684 2.87e-11 ***
YR_2000     -2.032e+02  3.055e+01  -6.650 3.60e-11 ***
HHAGE        1.720e+02  4.115e+01   4.181 3.01e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30670 on 2420 degrees of freedom
Multiple R-squared:  0.3157,	Adjusted R-squared:  0.314
F-statistic:   186 on 6 and 2420 DF,  p-value: < 2.2e-16
```

Analysis of Variance Table

Response: TOTALBTUSPH

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
MONEYPY	1	1.0235e+11	1.0235e+11	108.7755	< 2.2e-16	***
TOTSQFT_EN	1	2.6856e+11	2.6856e+11	285.4169	< 2.2e-16	***
HDD30YR	1	6.1536e+11	6.1536e+11	653.9767	< 2.2e-16	***
YR_2009	1	3.6779e+09	3.6779e+09	3.9088	0.04815	*
YR_2000	1	4.3895e+10	4.3895e+10	46.6497	1.070e-11	***
HHAGE	1	1.6445e+10	1.6445e+10	17.4769	3.012e-05	***
Residuals	2420	2.2771e+12	9.4095e+08			