Electricity Demand Variability due to climatic factors

Praneeth Valiveti
Sai Kiran Hanumakonda
Vineeth Chigarangappa Rangadhamappa

1. Introduction

Ambient environment plays an important role in determining the quantum of electricity consumed by a household or industrial unit, in any given year. A major share of this energy has been expended for spatial conditioning. Therefore, as the population and economic well-being of a community increases, the demand for spatial conditioning also increases, which in turn enhances the electricity demand. Further, the effects of climate change, which are already felt in different part of the United States of America, leads to an increase in frequent occurrences of extreme weather events. For instance,
"The state of Florida is predicted to see an increase in cooling degree days in the range of 750 to 1150 and a sharp increase in the number of extreme temperature (>95°F), quadrupling to over 60 days per year by the middle of the century."
Such an increase coupled with improved material wealth heightens the electricity demand, placing a severe strain on the country's energy infrastructure. At the same time, the ongoing population shift to the south and west of USA and consequent increase in use of electricity for space cooling reinforces the electricity consumption trend caused by climate warming.

Major electricity consumers are divided into four segments, based on their built environment. They are: residential, commercial, transportation and industrial. Among these segments, it was observed that residential and commercial sector exhibit greater sensitivity to vagaries of climate. The earlier study, by Sailor and Munoz (1997), focused on these two sectors because it was found out that the greatest increase in electricity consumption, since 1980, originated from the residential and commercial sectors. Hence the focus of the current study is residential and commercial sectors.

The goal of the current study is to develop statistical models that explains the relation between electricity and climatic variables on regional scales. As a case study, Mukhopadhyaya and Nateghi (2017) predicted the climate-sensitive electricity consumption in the state of Florida. Therefore, to obtain the spatio-temporal heterogeneity in climate-sensitive electricity demand in USA, this analysis has been extended to five other states of USA. They are: California, Illinois, New York, Texas and Washington. These states were chosen because of the following characteristics:

- Most energy intensive states representing significant fraction of total energy consumption in USA
- Diverse geographical distribution

2. Literature Review

Several studies have focused on modelling the effect of climate variability and climate change on residential and commercial energy consumption. A number of studies focused on establishing the link between energy demand and the climate to help with electric utility adequacy planning and load forecasting. However, these studies are specific to a given utility and not generalizable for regional planning. Region-specific predictive models are needed to estimate climate sensitive load and Implement effective policy analysis due to the following reasons:

- Global climate change has geographically distinct impacts.
- Regional assessment of energy consumption minimizes the effect of unobserved heterogeneities arising from regional differences in policy; characteristics of the energy sectors, the built-environment, and finally end-users' lifestyles.
- Energy consumption for different sectors are recorded at the region-level. Thus regional analyses will facilitate assessment of sectorial end-use electricity consumption sensitivity to climate.

Although there is a dearth of literature focusing on the regional energy demand sensitivity to weather and climate change, there has been some interesting studies focusing on sectorial energy demand sensitivity at regional level. Badri (1992) analysed the energy demand for all the 50 states in the U.S. using a two-stage least square methodology applied to cross-sectional data collected for the year 1988, for which energy demand and price elasticities were calculated for the three sectors – residential, commercial and industrial.

Sailor and Muñoz (1997) analysed the sensitivity of electricity and natural gas demand to climate for the top eight energy intensive states viz., California, Louisiana, Texas, Florida, Washington, Illinois, Ohio and New York. Amato et al. (2005) and Ruth & Lin (2006) estimated the regional energy demands to climate change for the Commonwealth of Massachusetts and the state of Maryland. Mirasgedis et al. (2007) conducted a regional energy demand analysis for Greece, with a multiple linear regression model. They accounted for climatic and socio-economic factors as well.

However, regional studies that analyse the climate sensitive electricity demand within states are few and far in between. As mentioned before, majority of such studies were undertaken by utility companies that were narrower in scope and were published only in their internal reports. Hence, there is a need to create appropriate models for each state and thus build a comprehensive model database for the entire U.S.A.

3. Data Source and Description

The dataset consists of climatic predictors, weather predictors and economic predictors. The type of predictors considered in each of these segments, their data source and pre-processing steps has been described below:

Electricity sales data was obtained from the database published by the United States Energy Information Administration (EIA). The monthly electricity sales data of the residential and commercial sectors for the period of 1990–2015 was extracted from the database. It consists of:

- Electricity sales (in GWHr) {**Note** : Electricity consumption is per million population}
- Electricity price (in cents/KWHr)

Climate data was obtained from the National Digital Forecast Database maintained by National Oceanic and Atmospheric Administration (NOAA), starting from 01-Jan-1990 to 01-Jan-2016. The climate data contains the following predictors:

- Total precipitation (tenth of mm precision) (predictor)

Weather data ranging from 01-Jan-1990 to 01-Dec-2016 was requested from National Climatic Data Center (NCDC). The weather data consists of following predictors:

- Mean dew point temperature (in ^{0}F) (predictor)
- Mean wind speed (in mph) (predictor)
- Maximum wind gust (in mph) (predictor)

The data pertaining to economic indicators consist of share of unemployed among total labour force. It was obtained from the U.S. Department of Labor, Bureau of Labor Statistics. The following table summarizes the final data set:

Table: 1

Predictors	**Response Variable**	**Data Sample Size**
Electricity Price	Electricity sales	313 observations pertaining to 313 months between 01/01/1990 to 01/01/2016
TPCP: Total precipitation		
MDPT: Mean dew point temperature		
Mean wind speed		
Maximum wind gust		
Unemployment		

3.1. Pre – Processing

The Electricity consumption data were adjusted for the present study by simply removing the population trend by dividing the monthly data by state-wide population interpolated from the census data. In order to isolate the influence of climate factors on electricity and natural gas consumption the per capita consumption was then trend-adjusted. The method to adjust electricity sales is described in the paper (Sailor & Muñoz – 1997). We calculate a yearly average electricity sales consumption $\bar{E}(y)$ from the monthly data over the entire period of study. The adjustment factor F_{adj} for each year was calculated from

$$F_{adj} = \bar{E}(y)^{-1} \sum_{m=1}^{12} E(m, y)$$

Each month of electricity data was adjusted by dividing it by adjustment factor for that year i.e.

$$E_{adj}(m, y) = E(m, y)/F_{adj}(y)$$

From the above two equation E represent electricity consumption. This adjustment was implemented to remove trend effect such as technology proliferation with time. In fact, it inherently contains all necessary adjustment factors. In the above equations E( ) represent the expectation of the variables, y denotes years and m denotes months. Our analyses were performed with the adjusted Electricity sales data.

Below figures show the raw and de-trended residential and commercial electricity consumption levels respectively of New York State:

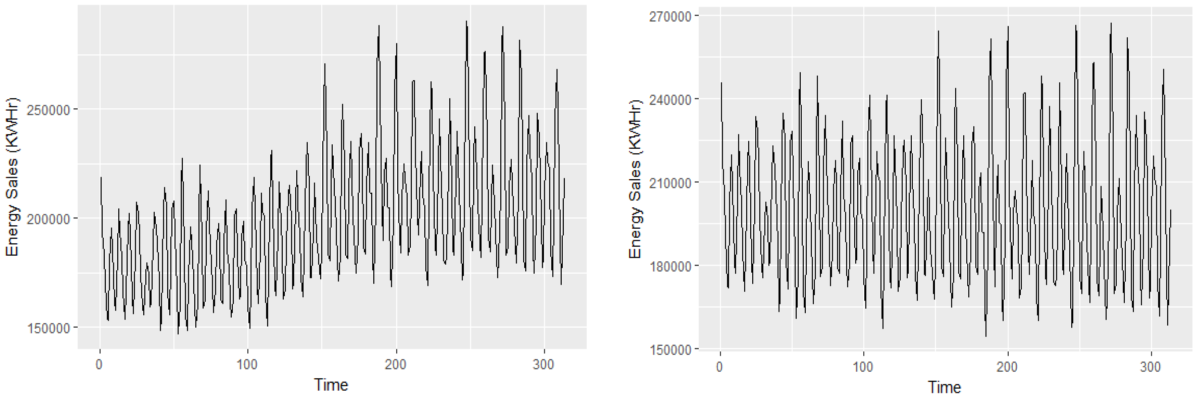


Fig.1 Raw and de- trended plots are for the New York State – Residential Sector

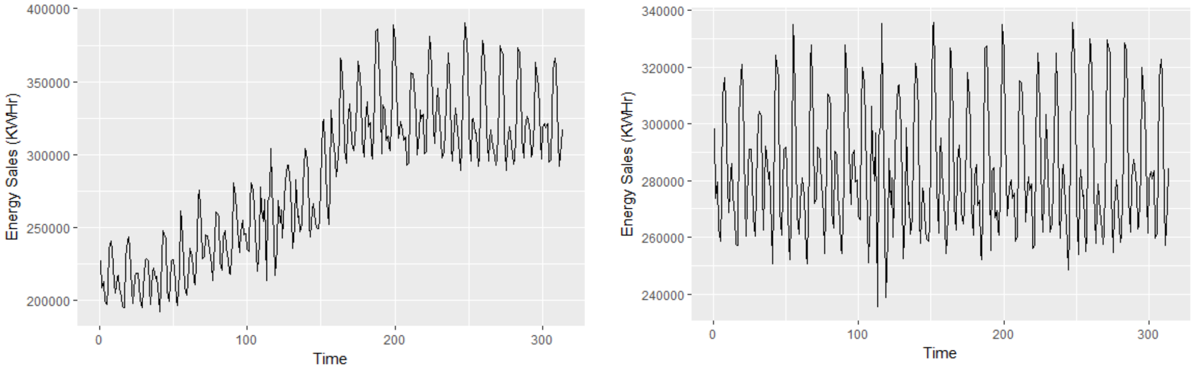


Fig.2 Raw and de- trended plots are for the New York State – Commercial Sector

3.2. Data Visualization

We see that overall, in the entire sales revenue data, there is very high correlation between the Sales and Revenue, which is predictable since the revenue depends directly upon the sales.
The price, however is not strongly correlated with the Revenue.
By this, we can conclude that there seems to be no linear association between the price of electricity (Cents/K.W.hour) and the Revenue (Dollars).

From the sales revenue data, we assume the “Sales” data to be our response variable. The other important variable from this dataset is the “Price”.

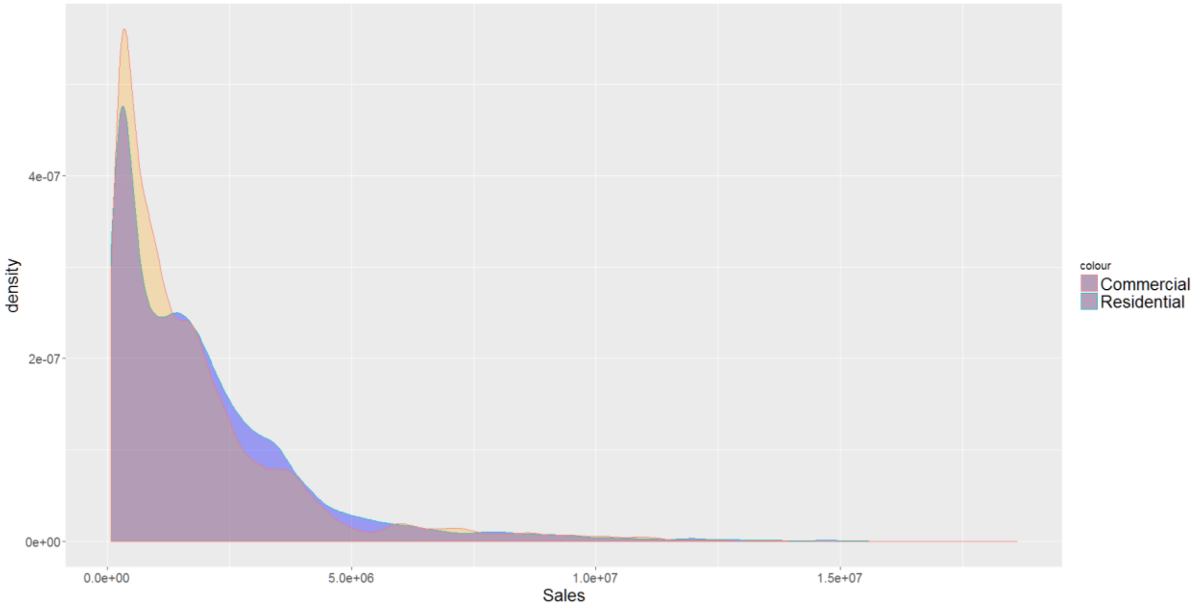


Fig: 3 Sales Densities for Residential (Blue) and Commercial (Red) Sectors

Above we compare the density distributions of the Sales of Residential (Blue) and the Commercial (Green) sectors. It is evident that both are highly right skewed, and both center around the value of

- Residential = 2001236 MegaWatts
- Commercial = 1811824 MegaWatts.

Hence, we can see clearly that we are dealing with a highly skewed Response Variable.

3.2.1. Sales Trend

This section deals with exploration of trends shown by “Sales” of electricity.

- Sales Trend over Entire Period of Study

From the below plots, we see that the Sales of electricity has a significant upward slope indicating that during the past 27 years, there has been a steady increase in the demand and sales of electricity in the United States.
Below are plots for both the Residential and Commercial Sectors, and from these we can see that there is a steep increase in electricity consumption in both the sectors.

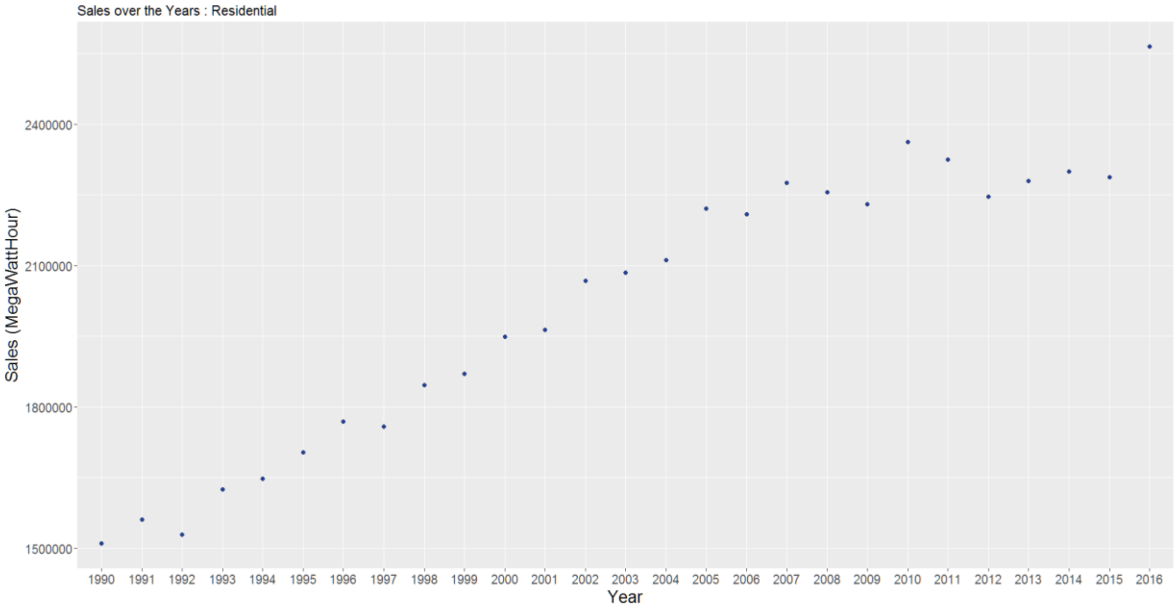


Fig: 4 Nationwide Monthly Electricity Consumption over the years - Residential

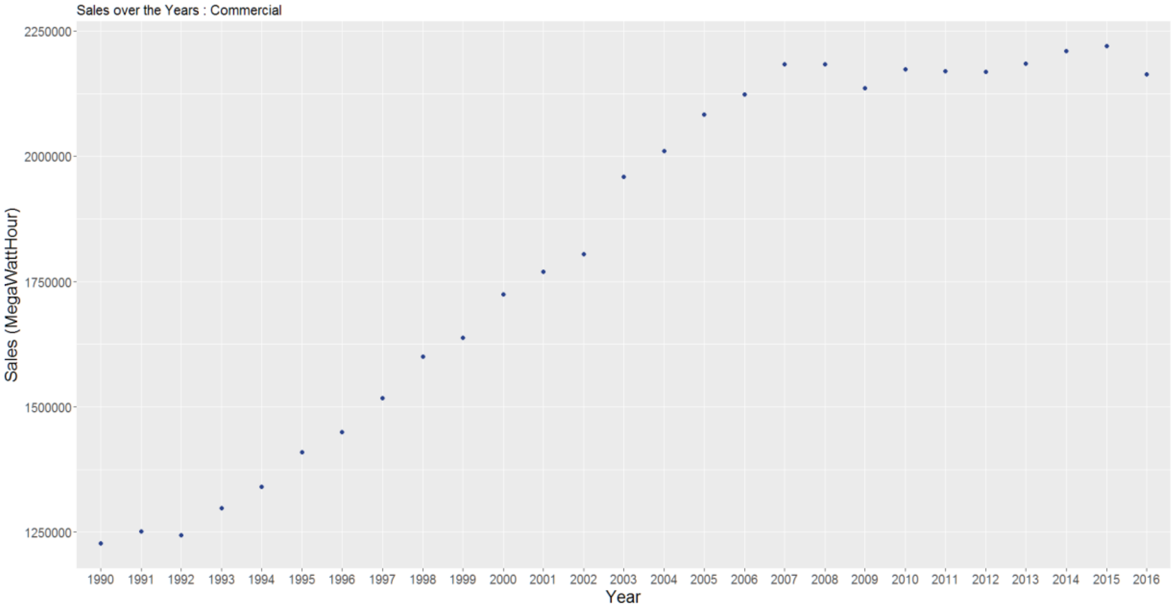


Fig: 5 Nationwide Monthly Electricity Consumption over the years - Commercial

From the two plots, we see the following:

a) There is an almost monotonous increase in the Sales for both Commercial and Residential sectors.

b) We notice a sharp jump in the sales between 2009 and 2010. This may be due to the Economic Depression in the United States in that year.

c) The last data point (year 2016) only represents the month of January, and hence is much varied than the previous year.

- Sales Trend on an average for a Year

On an aggregate, we now try to find the sales trend during the span of one year.
In order to do this, we plot the average values of Sales for each state, for each month of a year for the entire span of 27 years.

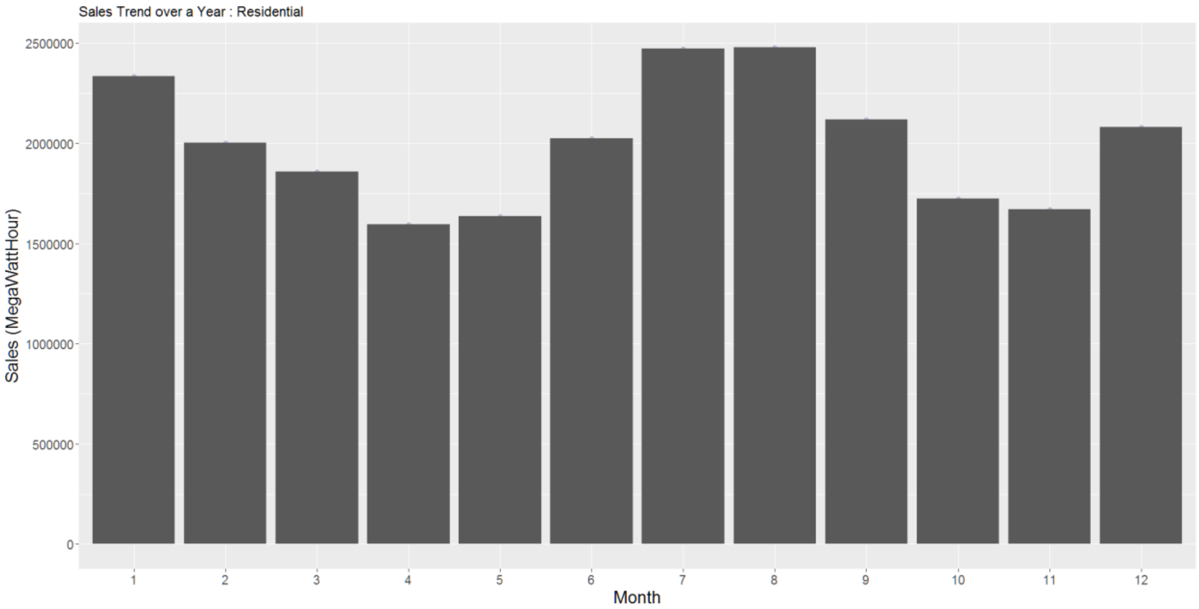


Fig: 6 Combined States' Monthly Electricity Consumption over the years - Residential

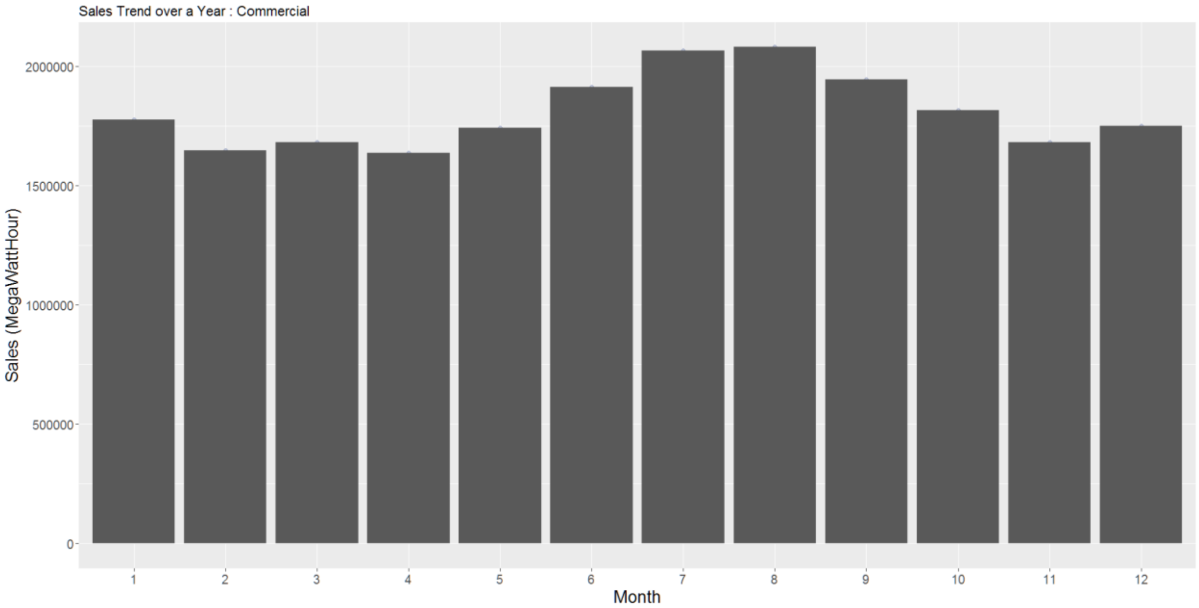


Fig: 7 Combined States' Monthly Electricity Consumption over the years - Commercial

Interestingly, we can notice a very similar pattern in both the residential and commercial sectors, which may have been caused by climate or weather patterns during the year. Since the patterns above are almost perfectly correlated, we can assume that there is at least one variable that affects the Sales over the Year.

3.2.2. Price Trend

In this section, we observe the trends in Price of Electricity over the 27 years of study:

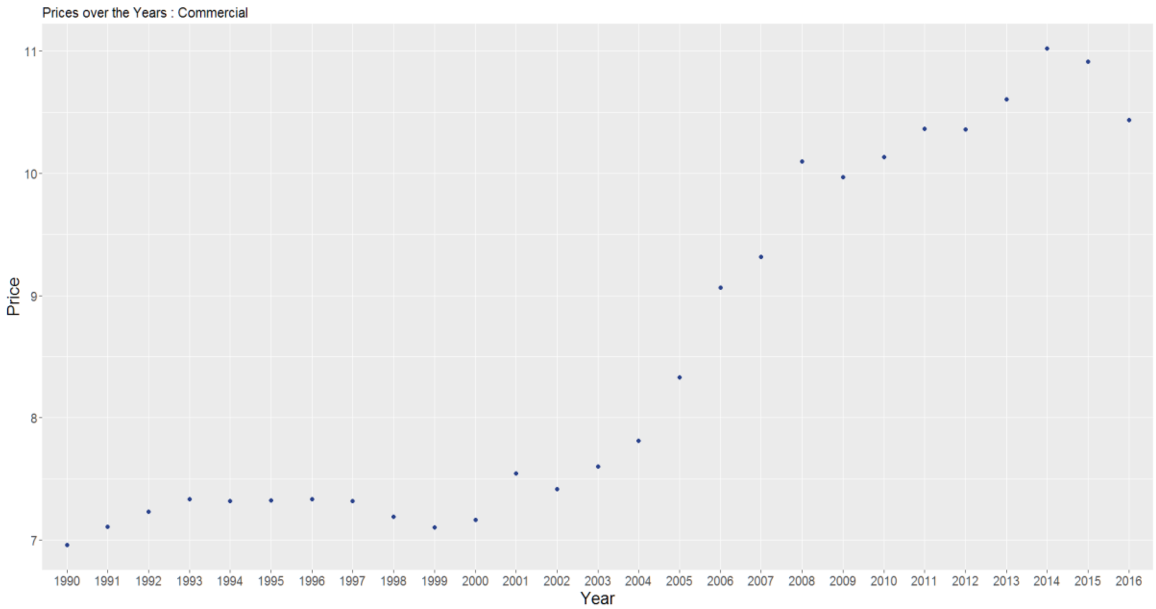


Fig: 8 National Electricity Price Trend over the years - Residential

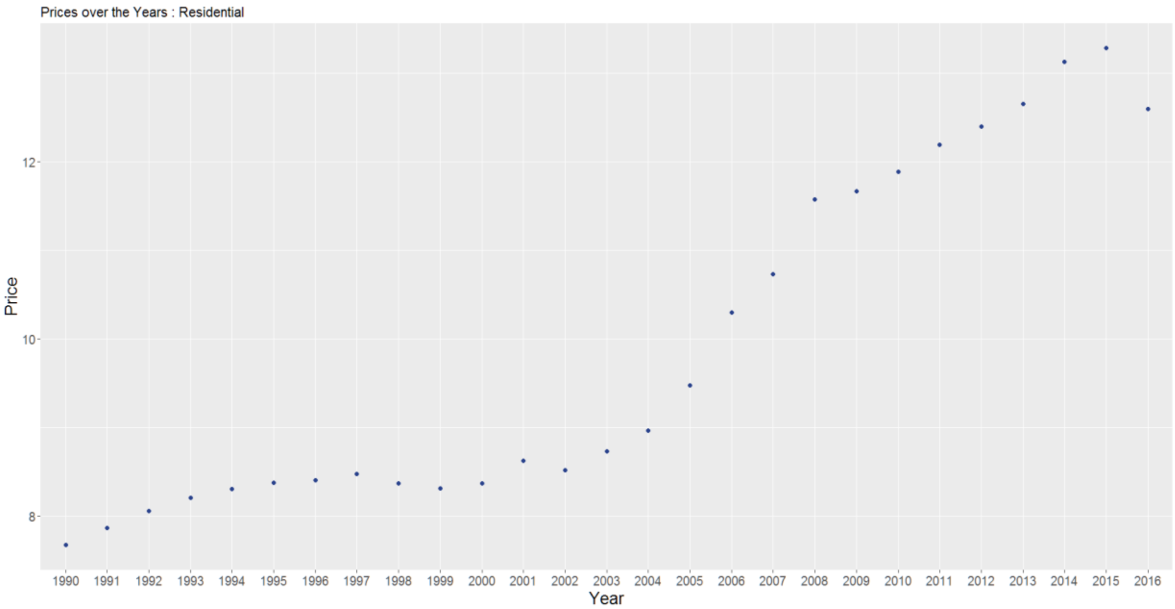


Fig: 9 National Electricity Price Trend over the years - Commercial

Like the Sales Trends, the Prices of Electricity also increases during the last 27 years. A point to be noted here is that there is a significantly large increase in the Price of electricity during the years 2002 to 2015.

Note: *The data does not contain the entire price data for 2016, and hence nothing can be concluded about the year 2016 from this chart.*

3.2.3. Sales vs Price

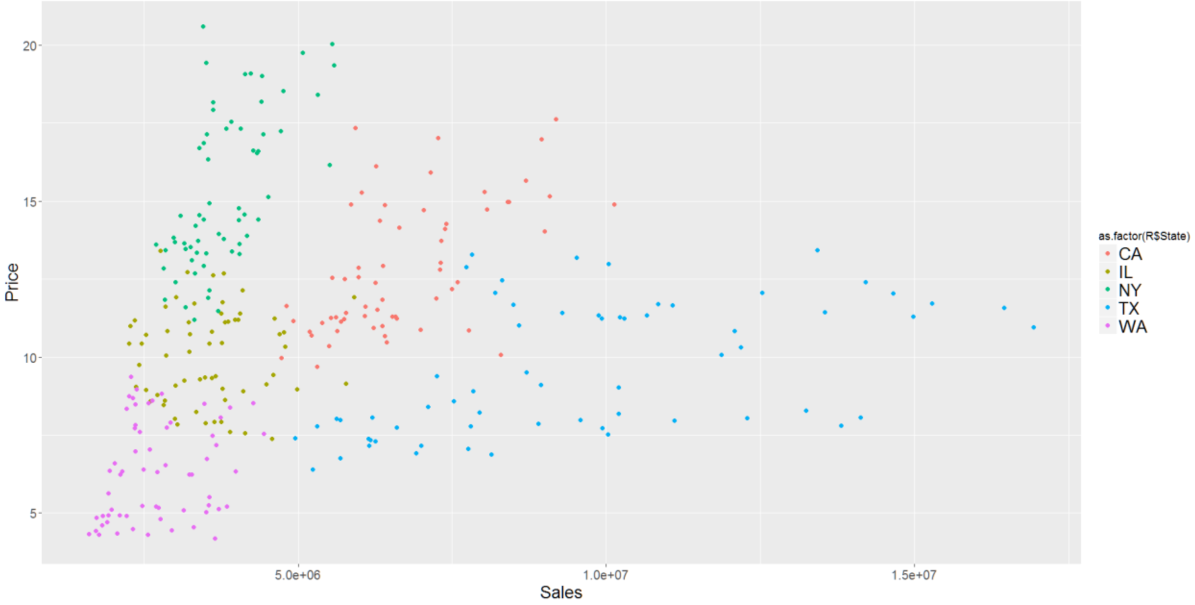


Fig: 10 Clustering Plot - Residential

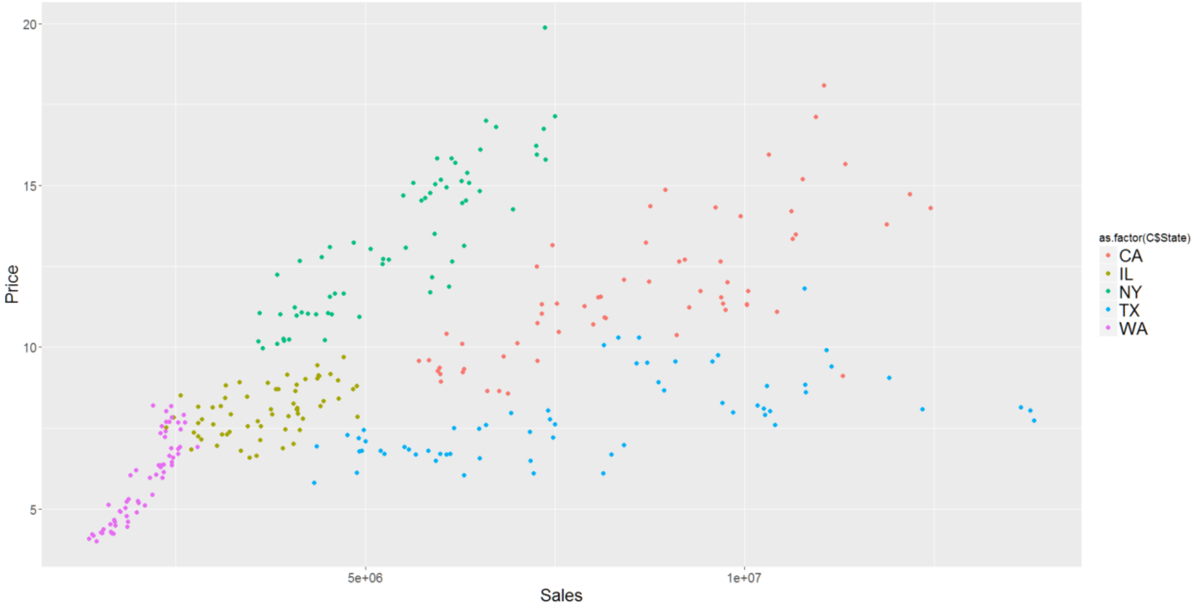


Fig: 11 Clustering Plot - Commercial

We can see an obvious cluster formation based on the states, and we see that:

Washington seems to maintain a low price-sales relationship, and most of it seems to vary linearly. This is expected since the state of Washington is not a very large contributor to the national electricity sales as we will see later in this section.

The relative “Mean” positions of each of the states seem to remain constant with respect to each other. This shows that indeed the per-capita and other factors affect the output of the State towards this graph.

3.2.4. Top States in Electricity Consumption per capita

Below plots show the actual Electricity sales as a per-capita estimate.
We see that in the Residential sector, Illinois top the list, whereas in the Commercial Sector, Texas tops the list

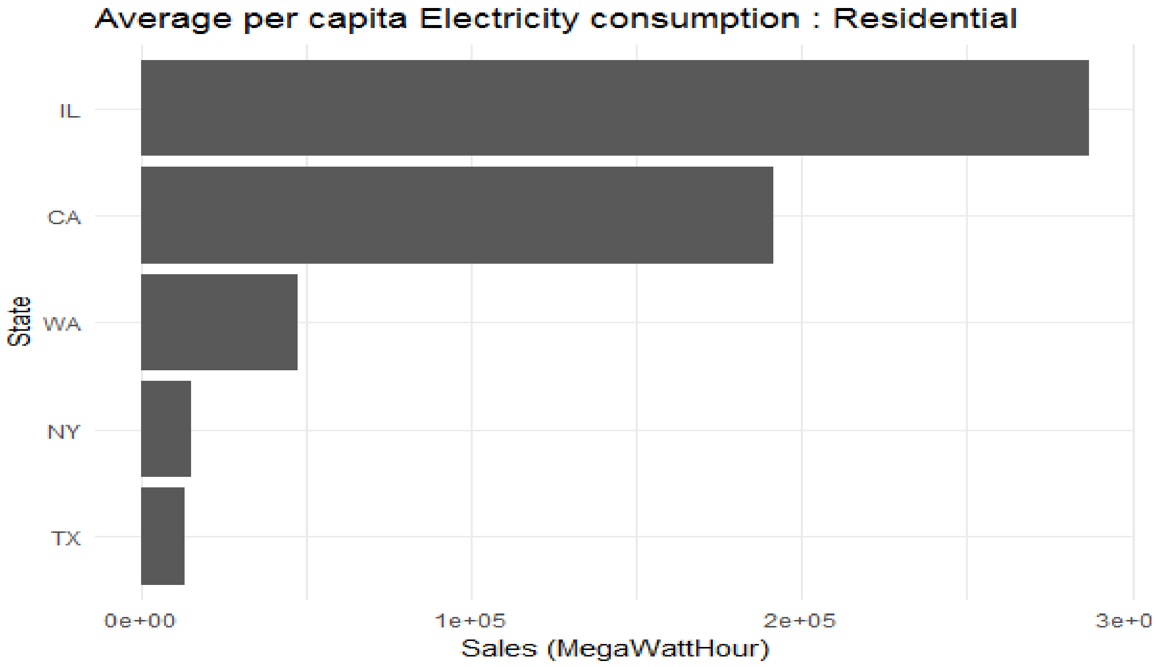


Fig: 12 Per Capita Electricity Consumption - Residential

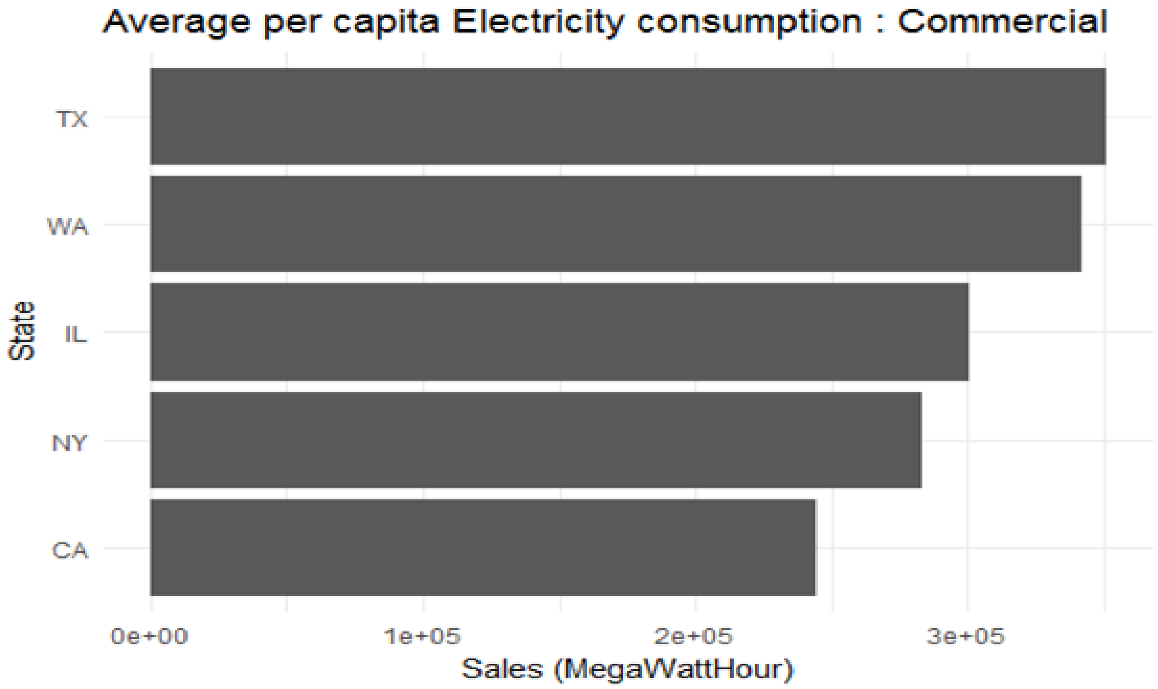


Fig: 13 Per Capita Electricity Consumption - Commercial

4. Methodology

The data is trained with a range of supervised learning models to investigate the electricity consumption in the residential and commercial sectors to climate and weather. Main idea of supervised learning methods is developing a predictive model that best captures the data structure and minimizes the loss. Linear and non- linear supervised statistical learning methods can be parametric, semi-parametric or non- parametric. Parametric models generally assume a particular functional form that relates the input variable to the response. Non-parametric models do not make assumptions about shape of the function relating response to the predictors. Instead they use data in novel ways to approximate the dependencies.

4.1. Supervised Learning Techniques

We trained our data with linear regression models, generalized additive models, multivariate adaptive regression splines and ensemble tree based models including random forest and Bayesian adaptive regression trees. Below, we will provide a brief review of each of the models:

4.1.1. Linear Regression Model

This a parametric model in which the response, Y is estimated as:

$$Y = X\beta + \varepsilon$$

The above Linear Regression Model is built under the following set of assumptions as listed below:
Model is assumed to be linear | Errors have constant variance | The observations are independent of each other | Errors follow normal distribution (normally distributed). The objective in this setting is to find the optimal value of parameters β using the methods of least squares which explains our data. The main disadvantage of linear models is that it is not flexible and we may be unable to capture the actual behavior of the data.
There is another class of methods called shrinkage method in linear regression where regression coefficients are penalized that help us deal with the issue of multicollinearity and helps in selection of best subset of variables as well. These techniques are called ridge regression and the lasso. In these shrinkage methods, our goal is to minimize the function of the form

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \theta\sum_{i=1}^{p} f(\beta_i)$$

where θ is the tuning parameter. The value of $f(\beta_i) = |\beta_i|$ for lasso method and $f(\beta_i) = \beta_i^2$ for ridge regression. The lasso regression helps in selecting the best subset of variables as well by giving us the sparse solution.

4.1.2. Generalized Additive Model (GAM)

GAM is a semi-parametric technique. It relaxes the linearity assumption of GLM, allowing for local non-linarites. This is done by introducing smoothing functions, which can capture which can help provide information that is not revealed using traditional linear models. The general form of the additive models is given as:

$$Y = \beta_0 + \sum_{j=1}^{P} f_j(X_j) + \epsilon$$

where each $f_j(X_j)$ is a smoothing function of a specified class of functions estimated non-parametrically, like regression splines and tensor product splines.

4.1.3. Multivariate Adaptive Regression Splines (MARS)

It is a non-parametric method that is suitable for high dimensional data. It is a method in which piecewise linear basis functions of the form $(x-t)_+ \; and \; (t-x)_+$ are formed at each observation for all the predictors. Each of the basis function is taken into consideration and best set of basis function is chosen that produces the largest decrease in the training error. But this may lead to overfitting in the data. Therefore, in order to avoid overfitting, we penalize the model for every extra basis function and finally chose the model which gives us the best value of general cross validation (GCV) which is defined as,

$$GCV(\theta) = \frac{\sum_{i=1}^{N}(y_i - \hat{f}_\theta(x_i))^2}{(1 - \frac{M(\theta)}{N})^2}$$

where $M(\theta)$ is the effective number of parameters in the model that penalizes the value of GCV

4.1.4. Random Forest (RF)

RF is a non-parametric tree-based ensemble data-miner. The method consists of B bootstrapped regression trees (Tb) with B selected based on cross-validation. Regression trees are low bias high variance techniques. In other words, they can capture the structure of the data really well (low bias), but are highly sensitive to outliers (high variance). RF leverages model averaging as a variance reduction technique. The final estimate is therefore, the average of predictions across all trees as shown in the equation below.

$$\hat{f}_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{T}^b(x)$$

4.1.5. Bayesian Adaptive Regression Trees (BART)

BART is a nonparametric Bayesian regression model and also a sum-of-tree model as shown in the equation below

$$Y = \sum_{j=1}^{P} g(X; T_j M_j) + \epsilon$$

- ***Partial Dependence Plots (PDPs)***

PDPs are efficient methods used for conducting variable inference for non-parametric models. They help in understanding the individual effects of the predictor variables. Mathematically, the estimated PDP is given by:

$$\hat{f}_j(X_j) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_j(X_j X_{-j,i})$$

where, f denotes the statistical model; n denotes the number of observations in the training dataset.

The generalization performance of a predictive model depends on its capability to make good predictions on an independent test sample. Balancing the bias-variance trade-off is key for minimized generalization error. Cross validation is one of the most widely used methods in balancing bias and variance. We use the method of k-fold cross validation to estimate predictive accuracy. K-fold cross-validation involves randomly dividing the data into k equally sized subsets. In each iteration, the model is fitted to the subsets except the kth held out sample and the predictive accuracy is calculated based on the models performance on the kth held-out subset.

$$MSE_{Out\ of\ Sample} = \frac{1}{K}\left[\sum_{k=1}^{n}\frac{1}{m}\left[\sum_{i=1}^{m}\left(y_{i.k}-\widehat{y_{i,k}}\right)^2\right]\right]$$

$$MAE_{Out\ of\ Sample} = \frac{1}{K}\left[\sum_{k=1}^{n}\frac{1}{m}\left|\sum_{i=1}^{m}\left(y_{i.k}-\widehat{y_{i,k}}\right)^2\right.\right]$$

k=number of times cross-validation is performed; m=number of holdouts during each cross-validation; $y_{i,k}$=i^{th} actual observation that was randomly holdout during the k^{th} cross-validation; $y_{i,k}$ =predicted i^{th} observation during the k^{th} cross-validation using the model developed using the training set data during the k^{th} cross-validation.

5. Results

The statistically trained models presented in this section predict monthly state-wide electricity consumption for the following 5 states of USA:

- California
- Illinois
- New York
- Texas and
- Washington

There exist 2 models for each state, pertaining to two sectors: residential and commercial. For each statistical model, the optimal set of model hyper parameters were obtained by minimizing the cross-validation estimates of the prediction error. For instance, the number of trees in the BART model was obtained by minimizing the out-of-sample root mean square error (RMSE) over a 5-fold cross-validation. Further, all the tunable hyper parameters of a model were obtained by grid-searching over the set of all possible hyper parameter combinations. For instance, in the BART model the parameters and their ranges considered are given below:

Table: 2

Parameter	Description	Range of values
m	Number of trees	{50, 200}
nu	Degrees of freedom for the inverse χ^2 prior	{3, 5, 10}
q	Quantile of the prior on the error variance	{0.75,0.99}
k	Prior probability	{1,2}

Taking one combination at a time, the BART model would be built and the set for which out-of-sample RMSE is minimum, that set will be taken as the optimal parameter set. We use the same procedure for choosing model hyper parameters pertaining to other statistical methods like number of trees for random forest model etc.

Prediction results of all the statistical learning models were rigorously assessed using different efficiency measures like RMSE, coefficient of determination (R^2). All the error rates reported for each statistical model refer to the cross-validation error rates. In cross-validation, the data-set is divided into 10-folds and a surrogate model is trained on combined 9-folds and tested upon one hold-out test set. This process is repeated 10 times, so that the model is tested on each observation in the data-set. The corresponding 10 error rates (both train and test error rates) that we obtain were averaged and in doing so, we assume the following:

- All the surrogate models are equivalent to teach other
- All the surrogate models are equivalent to the model that is trained on whole data-set

In each of the following sub-sections, the model performance measures are provided at the start in a table. It presents the key statistics to evaluate the efficiency of all the statistical models in the training and testing phases. It can be observed that all the machines have higher performance in the training phase than in the testing phase. The loss of performance on the testing set indicates the model's vulnerability to the issue of overtraining. The best statistical learning model was obtained after performing a pairwise Wilcox signed-rank test. For all the states, it can be noted that the BART model provides best performance in terms of predictive accuracy. This table is followed by plots showing:

- Model performance via predict vs actual plots and
- Model inference via partial dependency plots for top three predictors

5.1. California

5.1.1. Model Performance

Different statistical model results for both Residual and Commercial has been show below

Table: 3 Model performance measure – RESIDENTIAL

MODEL	TUNING.PARAMETERS	RMSE TRAIN	RMSE TEST	R. SQ
Linear	NA	20.080019	20.8892	0.29
Lasso	Default	20.135625	20.6219	0.29
GAM	Default	15.189296	18.2006	0.59
CART	Default	15.663601	19.9384	0.56
Bagged−CART	nbagg=500	12.418896	16.6278	0.73
MARS.1	pmethod=backward,nfold=10,ncross=5	15.181311	18.2706	0.59
MARS.2	pmethod=cv,nfold=10,ncross=5,degree=2	13.159839	16.0122	0.69
MARS.3	pmethod=cv,nfold=10,ncross=5,degree=3	13.286440	15.9734	0.69
MARS.4	pmethod=backward,nfold=10,ncross=5,degree=3,penalty=2	12.989811	16.6156	0.70
Random Forest	ntree = 300	6.692382	16.1622	0.92
BART	k= 2 num_trees= 50 q= 0.99 nu= 5	10.798137	15.4783	0.89

Table: 4 Model performance measure – COMMERCIAL

MODEL	TUNING PARAMETERS	RMSE TRAIN	RMSE TEST	R. SQ
Linear	NA	15.221356	15.73768	0.57
Lasso	Default	15.248425	15.74913	0.57
GAM	Default	12.409428	13.17244	0.71
CART	Default	12.691537	14.48082	0.70
Bagged−CART	nbagg=500	10.874252	13.20762	0.78
MARS.1	pmethod=backward,nfold=10,ncross=5	12.438437	13.82228	0.71
MARS.2	pmethod=cv,nfold=10,ncross=5,degree=2	12.263922	13.65123	0.72
MARS.3	pmethod=cv,nfold=10,ncross=5,degree=3	12.498903	13.40507	0.71
MARS.4	pmethod=backward,nfold=10,ncross=5,degree=3,penalty=2	11.654035	14.00044	0.74
Random Forest	ntree=300	5.657372	13.01656	0.94
BART	k= 2 num_trees= 50 q= 0.75 nu= 10	9.887708	12.74712	0.81

5.1.2. Model Diagnostics

Figure below shows the scatterplot of predicted vs actual. In case of the residential sector, the 95% credible intervals provide 76.68% coverage for all the observations. In case of the commercial sector, the 95% credible intervals provide 72.52% coverage for all the observations.

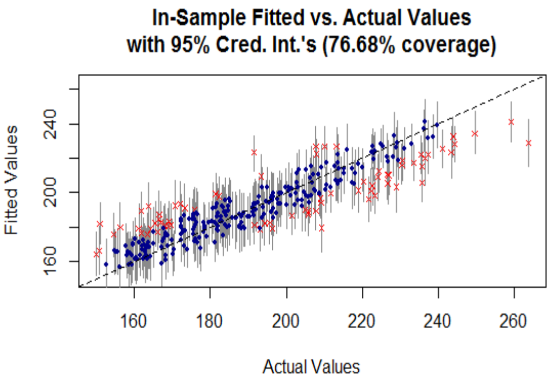


Fig14: In-Sample fitted vs. actual values : RE sector

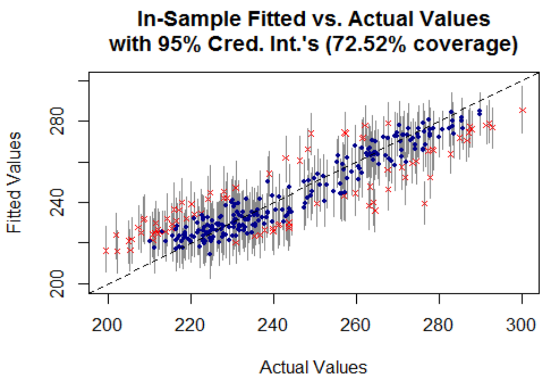


Fig15: In-Sample fitted vs. actual values : CE sector

5.1.3. Model Inference

It can be observed from the variables important plot that the trends for both sectors with Mean Dew Point Temperature (MDPT) found to be the most important predictor followed by Total Monthly Precipitation (TPCP) and Mean Wind Speed (WDSP). The only notable difference is that in the residential sector Electricity Price (data.california.price) is more important than Maximum Wind Gust (GUS) whereas the reverse is true for the commercial sector

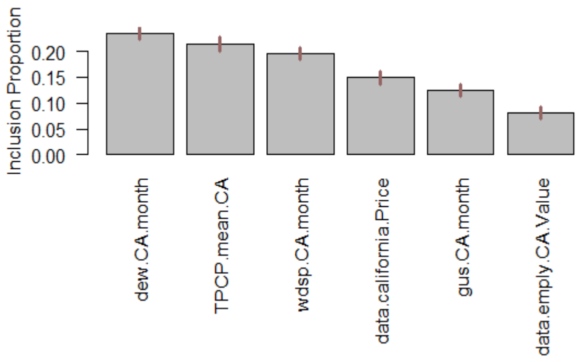


Fig16: Variable importance plot: RE sector

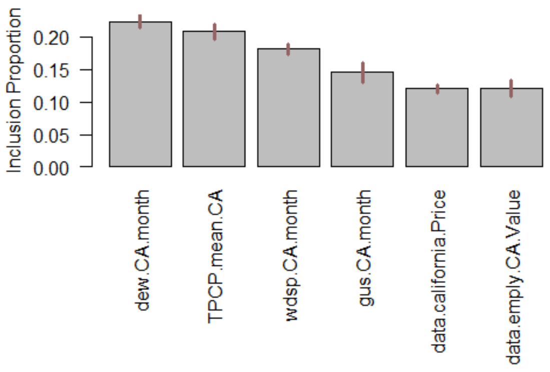


Fig17: Variable importance plots: CE sector

- ***Influence of Mean Dew Point Temperature***

From the Mean Dew Point Temperature plots of both Residential and commercial it can be observed that there is a non- linear relationship with the per capita consumption (Figure 18). Monthly per capita consumption of electricity (Residential) decreases first until the Mean Dew Point Temperature reaches 43°F and then increases till 52°F. It can also be inferred that the increase in per capita consumption between 43°F - 52°F is 30 MWH. Monthly per capita consumption of electricity (Commercial) increases as the MDPT increases more specifically it can be observed that after 43°F steep increase in the consumption and this is reasonable because the RE use is controlled by the households and therefore more influenced by their behavior.

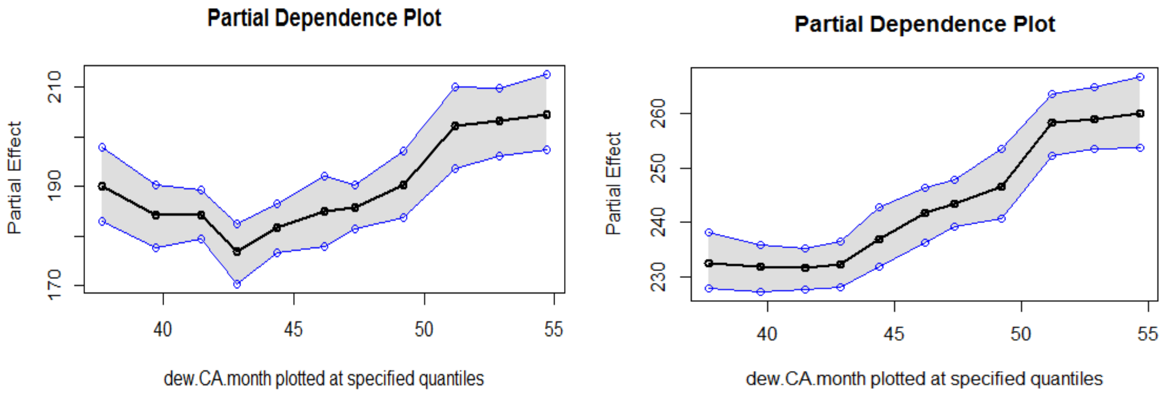


Fig18: Influence of MDPT on Residential and Commercial per capita consumption

- ***Influence of Total Monthly Precipitation***

From the Total Monthly Precipitation plots of both Residential and commercial sector it can be seen that as the TPCP increases the consumption decreases and then the consumption begins to increase for residential after TPCP = 2 this can be inferred from the box plot that the peak consumption occur during the summer months in which quantum of precipitation is considerably low as compared to other months.

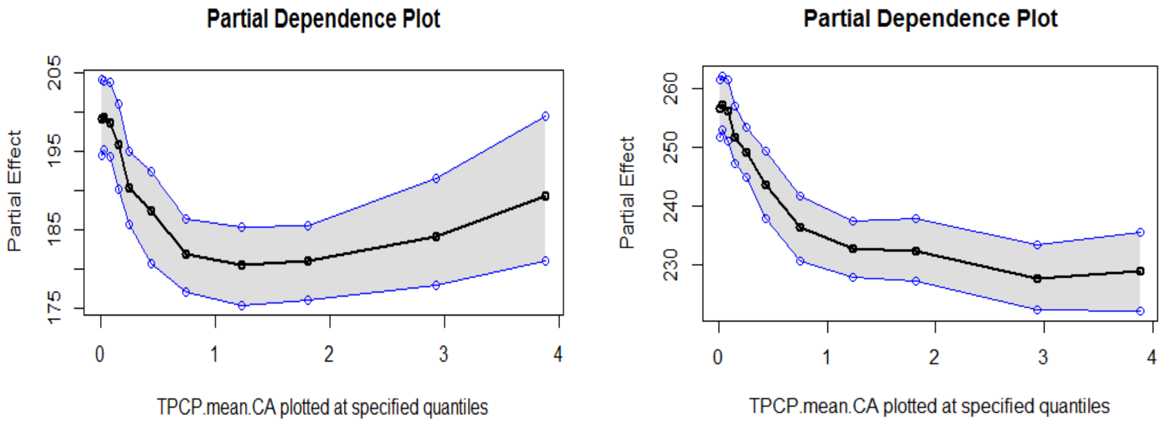


Fig19: Influence of TPCP on Residential and Commercial per capita consumption

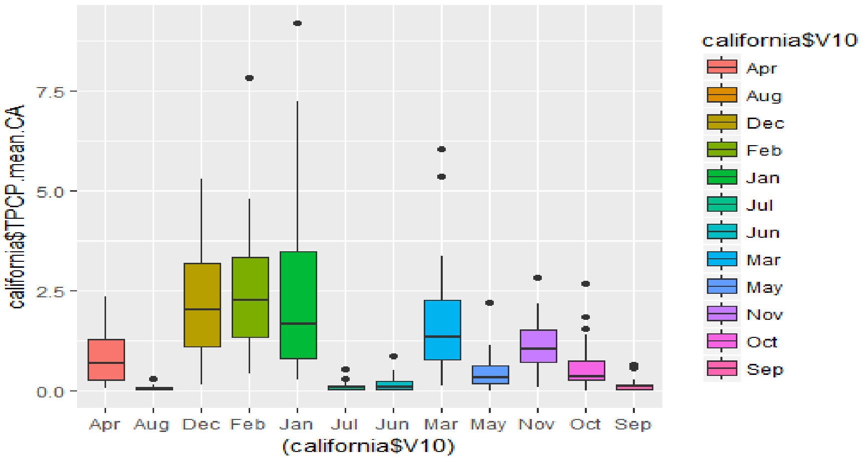


Fig20: Seasonal variation of TPCP

- ***Influence of Mean Wind Speed***

From the Mean Wind Speed plots of both Residential and commercial sector it can be seen that the peak consumption occur when the wind speed is between 4 to 6 MPH and it can be observed from the following box plot that this range of wind speeds pertains to summer months in which per capita electricity consumption is high.

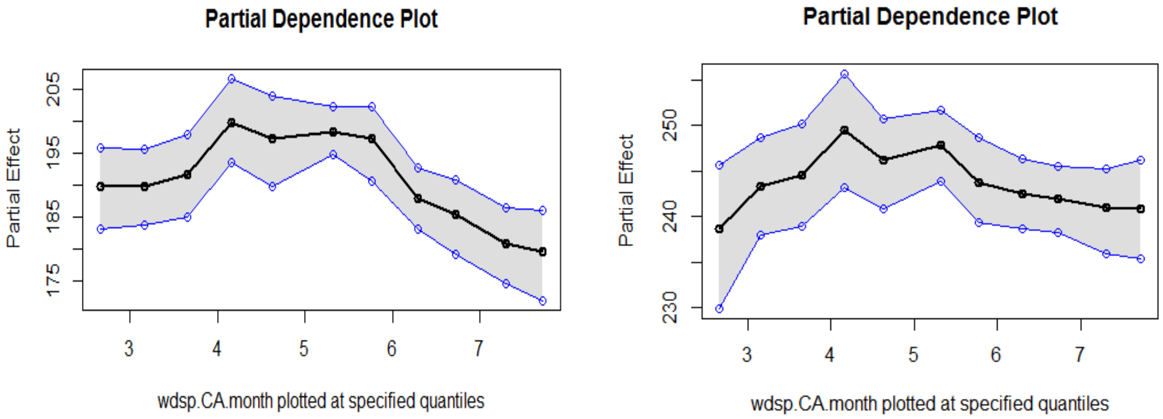


Fig21: Influence of WDSP on Residential and Commercial per capita consumption

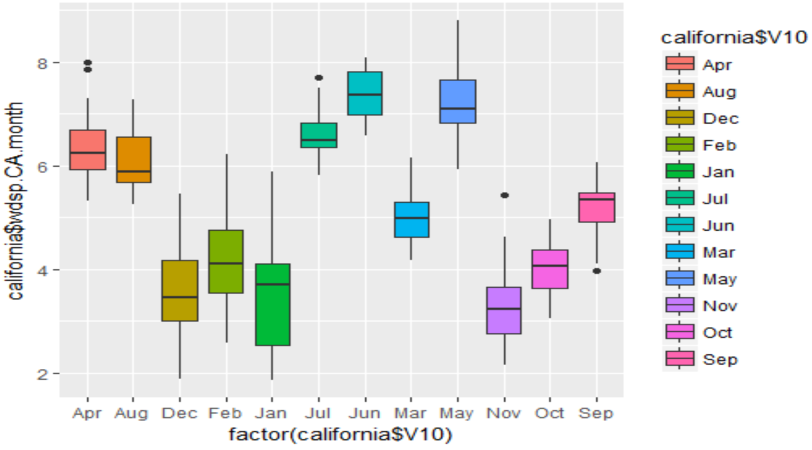


Fig22: Seasonal variation of WDSP

5.2. Illinois

5.2.1. Model Performance

Different statistical model results for both Residual and Commercial has been show below

Table: 5 Model performance measure – RESIDENTIAL

MODEL	TUNING.PARAMETERS	RMSE TRAIN	RMSE TEST	R.SQ
Linear	NA	51.408	52.418	0.19
Lasso	Default	51.505	52.847	0.18
GAM	Default	26.429	28.855	0.78
CART	Default	25.898	30.330	0.79
Bagged−CART	nbagg=500	23.169	27.244	0.83
MARS.1	pmethod=backward,nfold=10,ncross=5	26.077	29.773	0.79
MARS.2	pmethod=cv,nfold=10,ncross=5,degree=2	25.801	27.361	0.79
MARS.3	pmethod=cv,nfold=10,ncross=5,degree=3	25.455	27.670	0.80
MARS.4	pmethod=backward,nfold=10,ncross=5,degree=3,penalty=2	24.045	27.059	0.82
Random Forest	ntree=400	11.869	27.946	0.95
BART	k= 2 num_trees= 50 q= 0.75 nu= 10	20.3091	28.028	0.87

Table: 6 Model performance measure – COMMERICAL

MODEL	TUNING.PARAMETERS	RMSE TRAIN	RMSE TEST	R.SQ
Linear	NA	20.230921	19.77556	0.35606
Lasso	Default	20.232880	19.77681	0.35593
GAM	Default	14.763408	13.87859	0.65699
CART	Default	14.626473	14.97528	0.66328
Bagged−CART	nbagg=500	12.955017	14.21946	0.73597
MARS.1	pmethod=backward,nfold=10,ncross=5	14.467946	14.61004	0.67025
MARS.2	pmethod=cv,nfold=10,ncross=5,degree=2	15.681457	14.30641	0.61281
MARS.3	pmethod=cv,nfold=10,ncross=5,degree=3	15.702734	14.35458	0.61187
MARS.4	pmethod=backward,nfold=10,ncross=5,degree=3,penalty=2	13.249998	16.65847	0.72365
Random Forest	ntree=400	7.005007	14.84262	0.92271
BART	k= 2 num_trees= 50 q= 0.75 nu= 10	10.150626	14.82285	0.83755

5.2.2. Model Diagnostics

Figure below shows the scatterplot of predicted vs actual. In case of the residential sector, the 95% credible intervals provide 81.47% coverage for all the observations. In case of the commercial sector, the 95% credible intervals provide 90.1% coverage for all the observations.

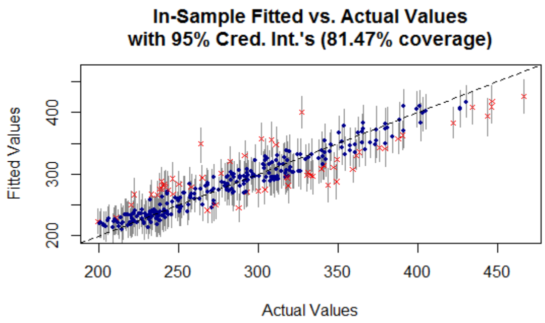


Fig23: In-Sample fitted vs. actual values: RE sector

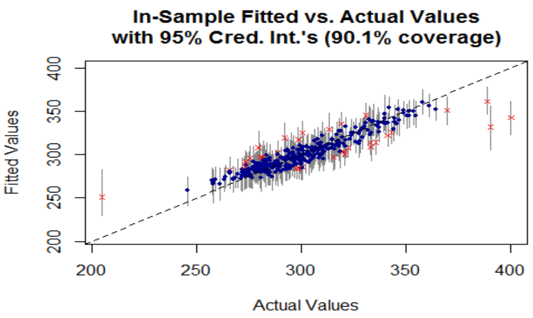


Fig24: In-Sample fitted vs. actual values: CE sector

5.2.3. Model Inference

It can be observed from the variables important plot that the trends for both sectors with Mean Dew Point Temperature (MDPT) found to be the most important predictor followed by Maximum Wind Gust (GUS) and Electricity Price (data.illinois.price). The notable difference is that in the residential sector Mean Wind Speed (WDSP) is more important when compared to commercial sector. Since Illinois lies away from the coast, we observe that maximum wind gust plays a significant role in determining monthly per capita electricity consumption.

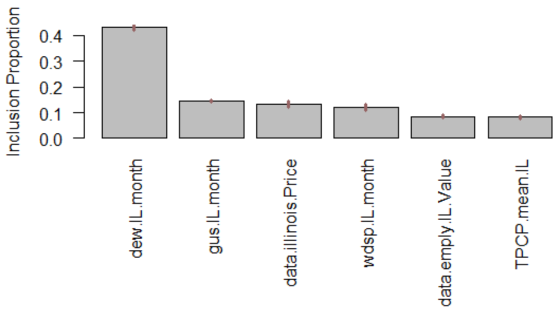


Fig25: Variable importance plot: RE sector

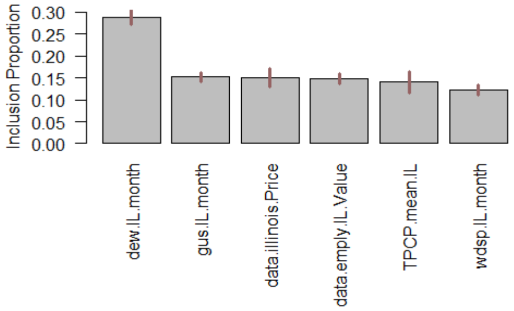


Fig26: Variable importance plots: CE sector

- ***Influence of Mean Dew Point Temperature***

From the Mean Dew Point Temperature plots of both Residential and commercial it can be observed that there is a non- linear relationship with the per capita consumption (Figure.27). For both sectors, with increasing dew point temperature, monthly per capita consumption of electricity (Residential) decreases initially and then increases after a threshold. These thresholds are different (45°F for Residential and 30^0F for Commercial) for both the sectors. This corresponds to difference in the behavior of the occupants, where in residential occupants are more concerned with the usage of the electricity for air conditioning.

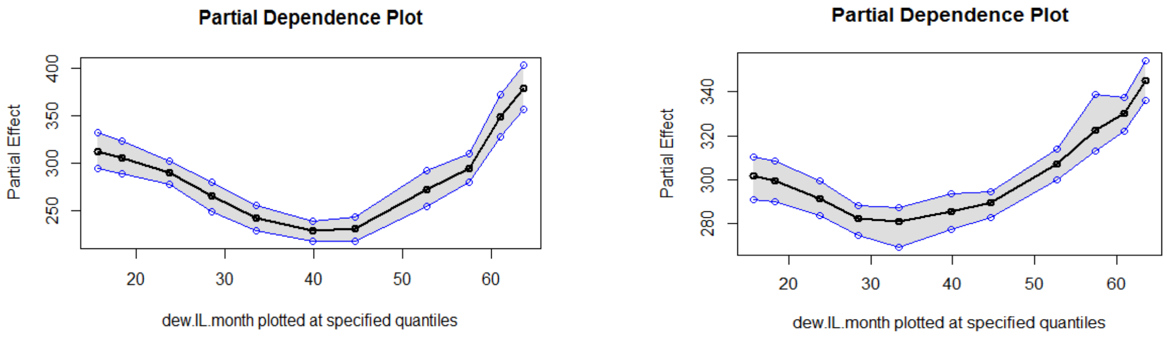


Fig27: Influence of MDPT on Residential and Commercial per capita consumption

- ***Influence of Maximum Wind Gust***

From the Maximum Wind Gust plots of both Residential and commercial sector, it can be noticed that the consumption more or less remains constant with increasing wind gust.

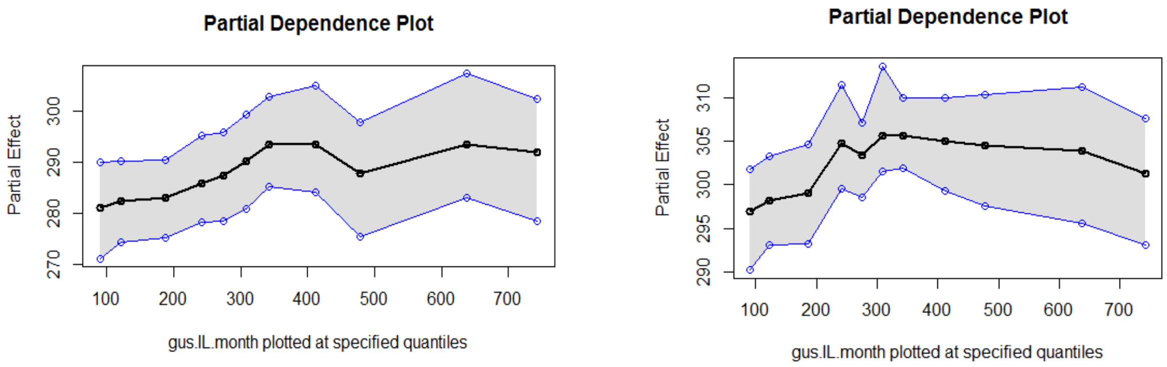


Fig28: Influence of GUS on Residential and Commercial per capita consumption

- ***Influence of Electricity Price***

From the Electricity Price plots of both Residential and commercial sector, it can be noticed that the consumption slightly decreases with increasing Price which is very intuitive.

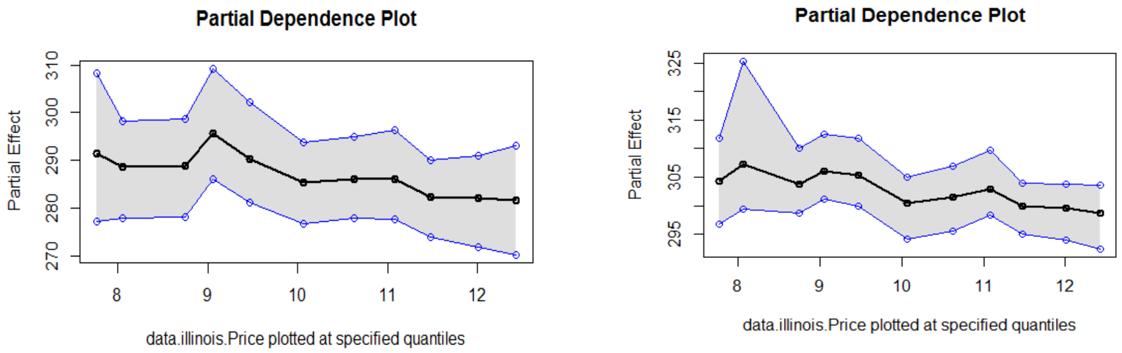


Fig29: Influence of *Electricity Price* on Residential and Commercial per capita consumption

5.3. New York

5.3.1. Model Performance

Different statistical model results for both Residual and Commercial has been show below

Table: 7 Model performance measure – RESIDENTIAL

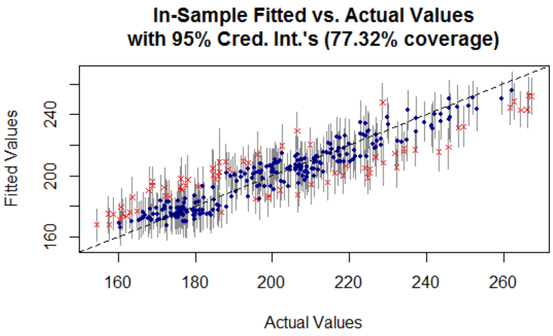
MODEL	TUNING.PARAMETERS	RMSE TRAIN	RMSE TEST	R.SQ
Linear	NA	24.968450	25.34486	0.03751
Lasso	Default	25.079134	25.07974	0.02896
GAM	Default	20.03212	19.86733	0.19675
CART	Default	12.575676	15.10129	0.75521
Bagged−CART	nbagg=500	11.044426	14.23906	0.81153
MARS.1	pmethod=backward,nfold=10,ncross=5	12.982952	15.50917	0.73947
MARS.2	pmethod=cv,nfold=10,ncross=5,degree=2	13.145904	14.63400	0.73261
MARS.3	pmethod=cv,nfold=10,ncross=5,degree=3	13.527975	14.98675	0.71738
MARS.4	pmethod=backward,nfold=10,ncross=5,degree=3,penalty=2	12.102645	14.81880	0.77367
Random Forest	ntree=300	5.887232	14.12495	0.94641
BART	k= 2 num_trees= 50 q= 0.75 nu= 10	10.050113	14.00352	0.84387

Table: 8 Model performance measure – COMMERCIAL

MODEL	TUNING.PARAMETERS	RMSE TRAIN	RMSE TEST	R.SQ
Linear	NA	18.406320	18.64012	0.31289
Lasso	Default	18.468266	18.51621	0.30827
GAM	Default	12.324409	15.32768	0.52121
CART	Default	10.006755	11.17202	0.79674
Bagged−CART	nbagg=500	9.158278	10.56436	0.82984
MARS.1	pmethod=backward,nfold=10,ncross=5	9.969322	11.46413	0.79806
MARS.2	pmethod=cv,nfold=10,ncross=5,degree=2	10.179514	25.48397	0.78949
MARS.3	pmethod=cv,nfold=10,ncross=5,degree=3	10.308528	25.51043	0.78429
MARS.4	pmethod=backward,nfold=10,ncross=5,degree=3,penalty=2	9.677852	28.59840	0.80988
Random Forest	ntree=400	4.500594	10.49863	0.95889
BART	k= 2 num_trees= 50 q= 0.75 nu= 3	7.743459	10.18539	0.87821

5.3.2. Model Diagnostics

Figure below shows the scatterplot of predicted vs actual. In case of the residential sector, the 95% credible intervals provide 77.32% coverage for all the observations. In case of the commercial sector, the 95% credible intervals provide 78.27% coverage for all the observations



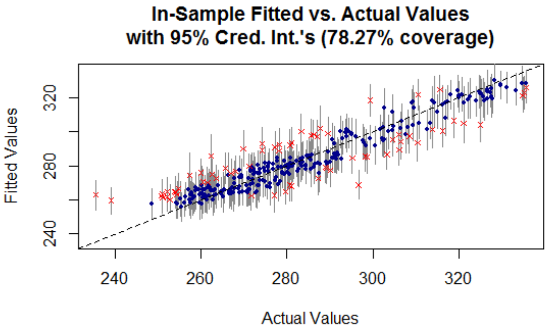
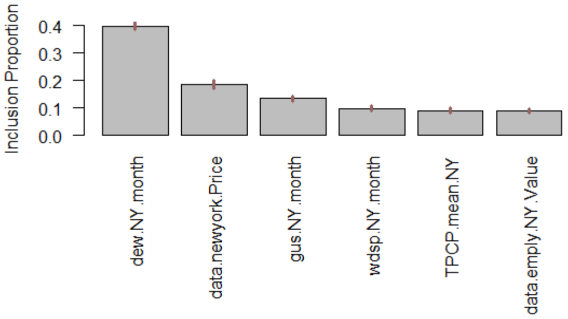


Fig30: In-Sample fitted vs. actual values: RE sector **Fig31: In-Sample fitted vs. actual values: CE sector**

5.3.3. Model Inference

It can be observed from the variables important plot that the trends for both sectors Mean Dew Point Temperature (MDPT) and Electricity price are found to be the most important predictors. The only notable difference is that in the commercial sector Total Monthly Precipitation (TPCP) is more important than Maximum Wind Gust (GUS) whereas the reverse is true for the residential sector



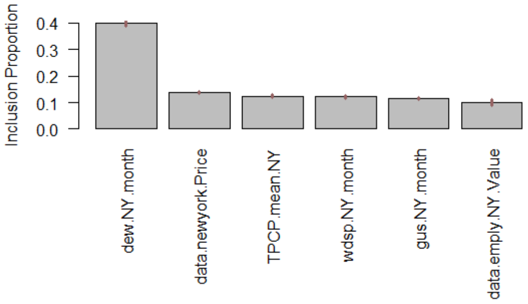


Fig32:Variable importance plot: RE sector **Fig33: Variable importance plots: CE sector**

- ***Influence of Mean Dew Point Temperature***

From the Mean Dew Point Temperature plots of both Residential and commercial it can be observed that there is a non- linear relationship with the per capita consumption (Figure.34). For both sectors, monthly per capita consumption of electricity decreases initially with increasing mean dew point temperature till about 48^{0}F and then sharply increases. Therefore comfortable range of Dew Point Temperature ranges between 35^{0}F and 48^{0}F. Beyond this range the increased consumption can be correlated with increase in cooling due to summer months and can be observed from the following dew point temperature box plot.

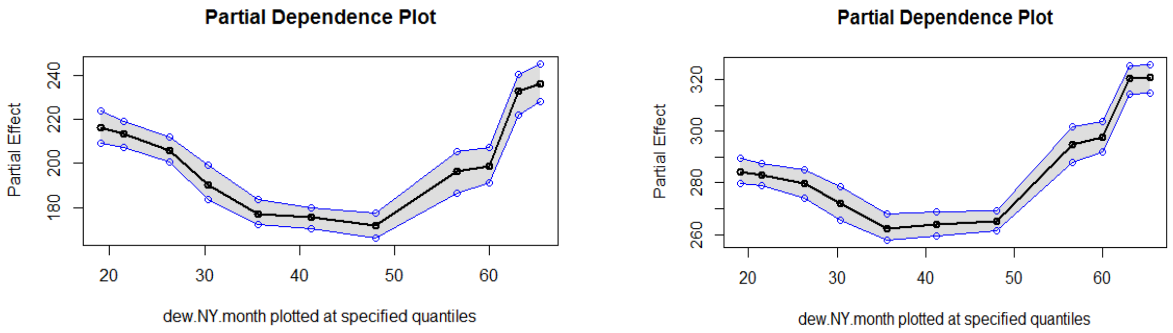


Fig34: Influence of MDPT on Residential and Commercial per capita consumption

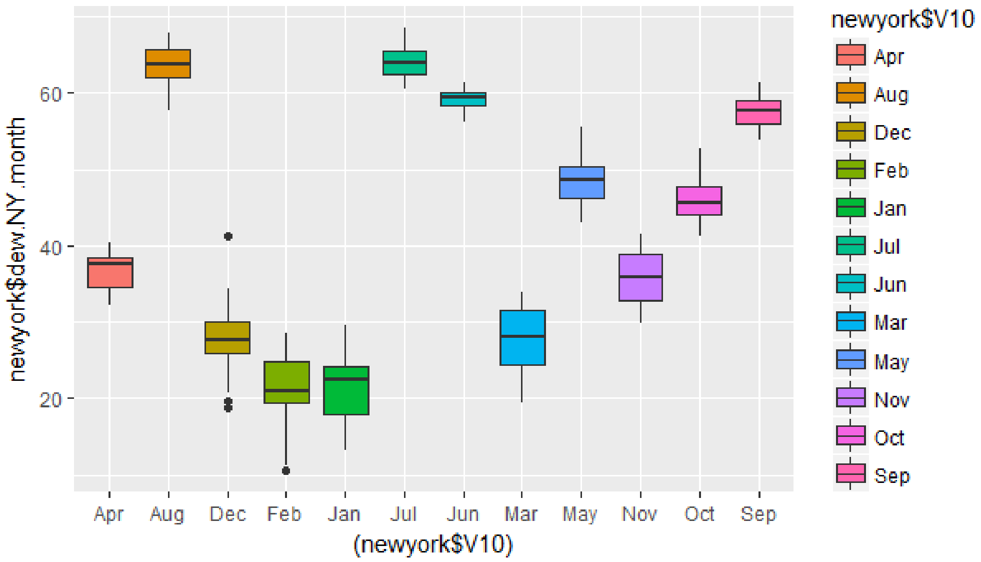


Fig35: Seasonal variation of MDPT

- ***Influence of Electricity Price***

From the Electricity Price plots of both Residential and commercial sector, it can be noticed that the consumption doesn’t change with increase in price. This may be attributed to relative affluence of New York State

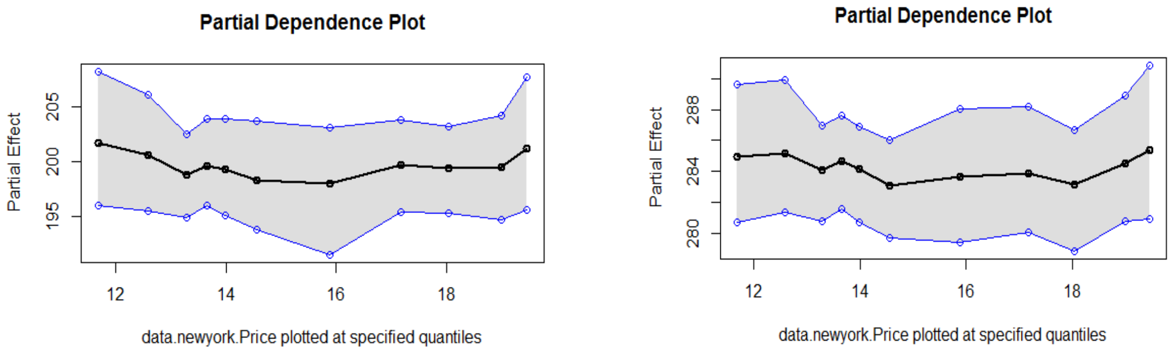


Fig36: Influence of Electricity Price on Residential and Commercial per capita consumption

- ***Influence of Total Monthly Precipitation***

From the Total Monthly Precipitation plots of both Residential and commercial sector it can be seen that, as the TPCP increases the consumption slightly increases. This can be explained by observing from the box plot that the peak consumption occur during the summer months in which quantum of precipitation is considerably high as compared to other months.

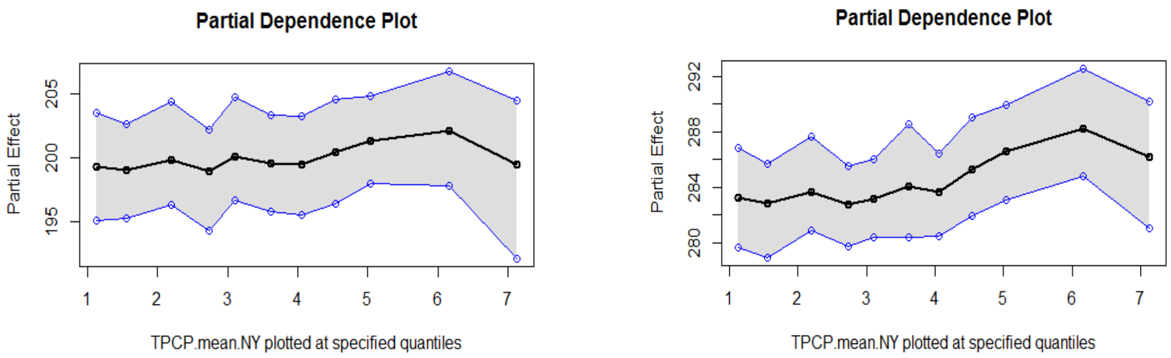


Fig37: Influence of TPCP on Residential and Commercial per capita consumption

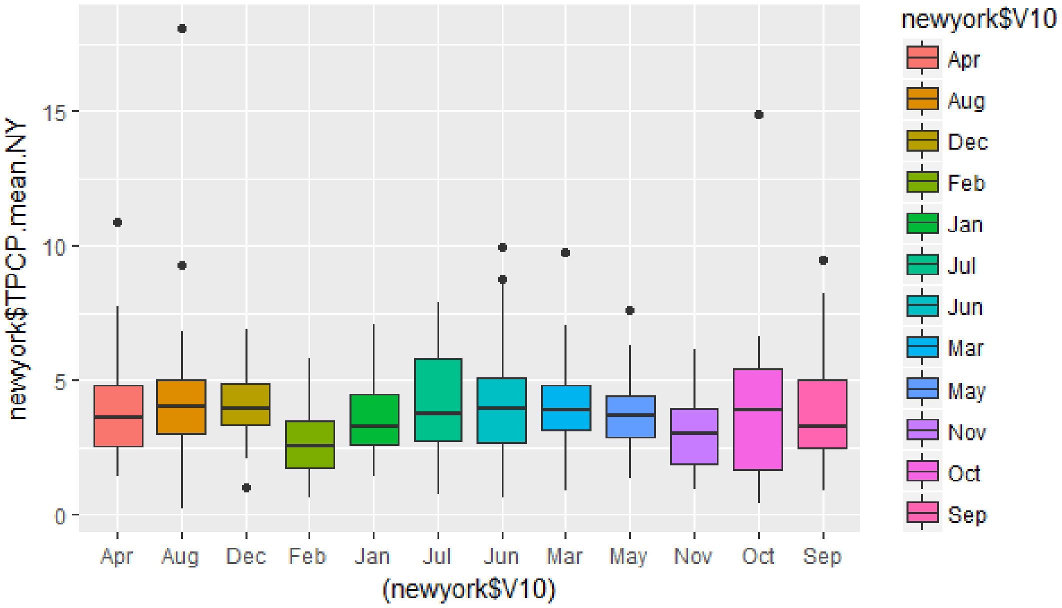


Fig38: Seasonal variation of TPCP

5.4. Texas

5.4.1. Model Performance

Different statistical model results for both Residual and Commercial has been show below

Table: 9 Model performance measure – RESIDENTIAL

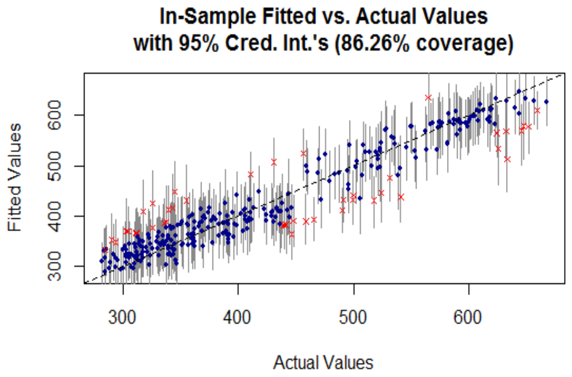
MODEL	TUNING.PARAMETERS	RMSE TRAIN	RMSE TEST	R SQ
Linear	NA	65.82385	67.31209	0.63
Lasso	Default	66.07786	67.01014	0.63
GAM	Default	50.36662	57.02317	0.78
CART	Default	51.65176	60.91207	0.77
Bagged−CART	nbagg=500	44.40210	55.05182	0.83
MARS.1	pmethod=backward,nfold=10,ncross=5	52.30416	59.82378	0.77
MARS.2	pmethod=cv,nfold=10,ncross=5,degree=2	50.96445	55.64150	0.78
MARS.3	pmethod=cv,nfold=10,ncross=5,degree=3	51.72611	53.75982	0.77
MARS.4	pmethod=backward,nfold=10,ncross=5,degree=3,penalty=2	47.99627	57.69320	0.80
Random Forest	ntree=1000	22.28229	53.57223	0.95
BART	k= 1 num_trees= 50 q= 0.9 nu= 10	32.67943	54.58543	0.91

Table: 10 Model performance measure – COMMERCIAL

MODEL	TUNING.PARAMETERS	RMSE TRAIN	RMSE TEST	R.SQ
Linear	NA	22.0236	22.463	0.76466
Lasso	Default	22.0571	22.390	0.76394
GAM	Default	19.7928	21.967	0.80988
CART	Default	19.6456	22.094	0.81192
Bagged−CART	nbagg=500	17.2767	20.713	0.85516
MARS.1	pmethod=backward,nfold=10,ncross=5	19.7947	23.783	0.80978
MARS.2	pmethod=cv,nfold=10,ncross=5,degree=2	20.5255	20.607	0.79551
MARS.3	pmethod=cv,nfold=10,ncross=5,degree=3	20.4267	20.773	0.79752
MARS.4	pmethod=backward,nfold=10,ncross=5,degree=3,penalty=2	19.1223	21.207	0.82250
Random Forest	ntree=1000	8.62154	20.640	0.96391
BART	k= 2 num_trees= 50 q= 0.99 nu= 5	15.3846	22.477	0.8850512

5.4.2. Model Diagnostics

Figure below shows the scatterplot of predicted vs actual. In case of the residential sector, the 95% credible intervals provide 86.26% coverage for all the observations. In case of the commercial sector, the 95% credible intervals provide 75.72% coverage for all the observation



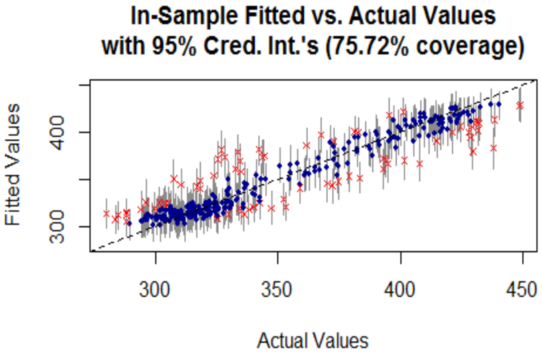
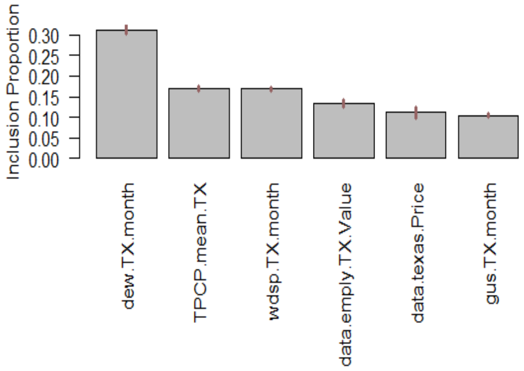


Fig39: In-Sample fitted vs. actual values: RE sector **Fig40: In-Sample fitted vs. actual values: CE sector**

5.4.3. Model Inference

It can be observed from the variables important plot that the trends for both sectors with Mean Dew Point Temperature (MDPT) found to be the most important predictor followed by Total Monthly Precipitation (TPCP) and Wind speed (WDSP).



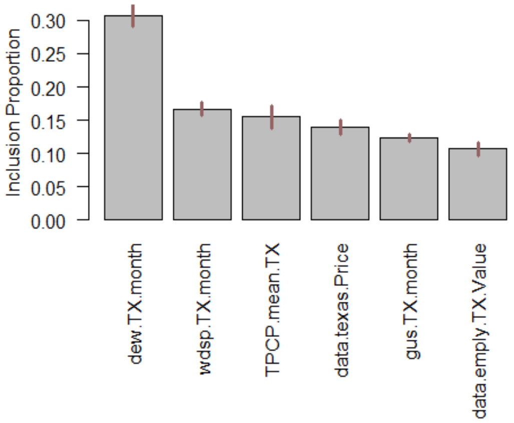


Fig41:Variable importance plot: RE sector **Fig42: Variable importance plots: CE sector**

- ***Influence of Mean Dew Point Temperature***

From the Mean Dew Point Temperature plots of both Residential and commercial it can be observed that there is a non- linear relationship with the per capita consumption (Figure…& Figure…). For residential sector, Monthly per capita consumption of electricity decreases initially with increasing mean dew point temperature till about 32^0F and then sharply increases. For commercial sector, the consumption remains constant till 32^0F and then increases at a much steeper rate when compared to residential sector. Steep increase corresponds to summer months which can be observed from box plot.

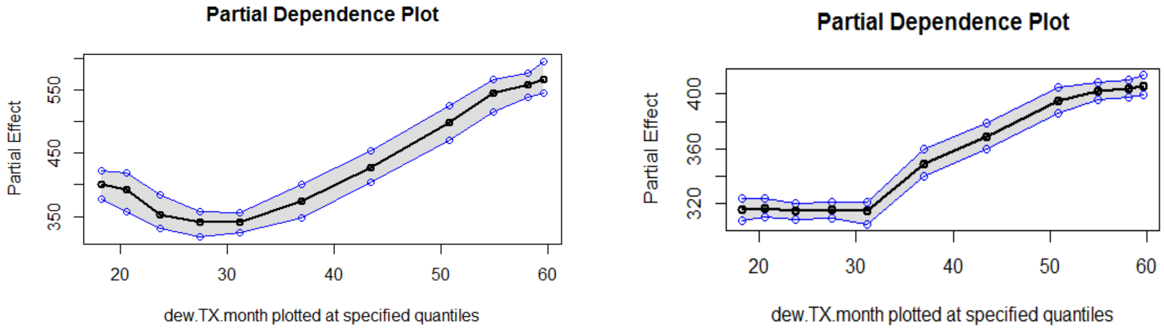


Fig43: Influence of MDPT on Residential and Commercial per capita consumption

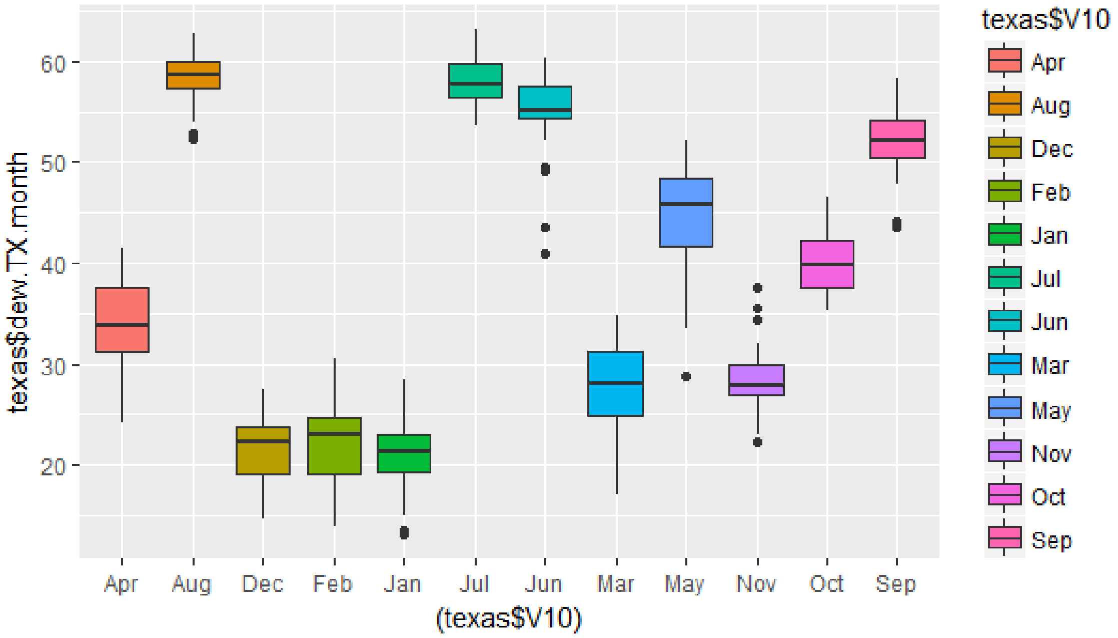


Fig44: Seasonal variation of MDPT

- ***Influence of Total Monthly Precipitation***

From the Total Monthly Precipitation plots of both Residential and commercial sector it can be seen that, as the TPCP increases the consumption decreases continuously, which is very intuitive.

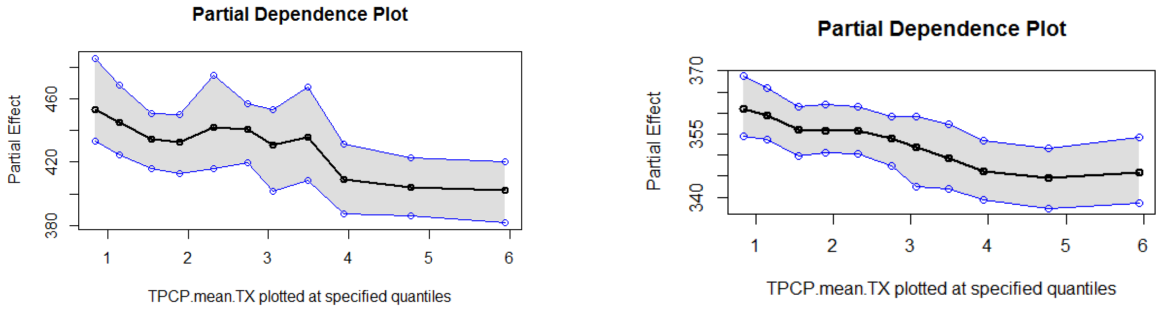


Fig45: Influence of TPCP on Residential and Commercial per capita consumption

- ***Influence of Mean Wind Speed***

From the Mean Wind Speed plots of both Residential and commercial sector, it can be noticed that the consumption decreases with increasing wind speed, especially for residential sector. This is intuitive as increasing wind speeds takes away the need for spatial cooling.

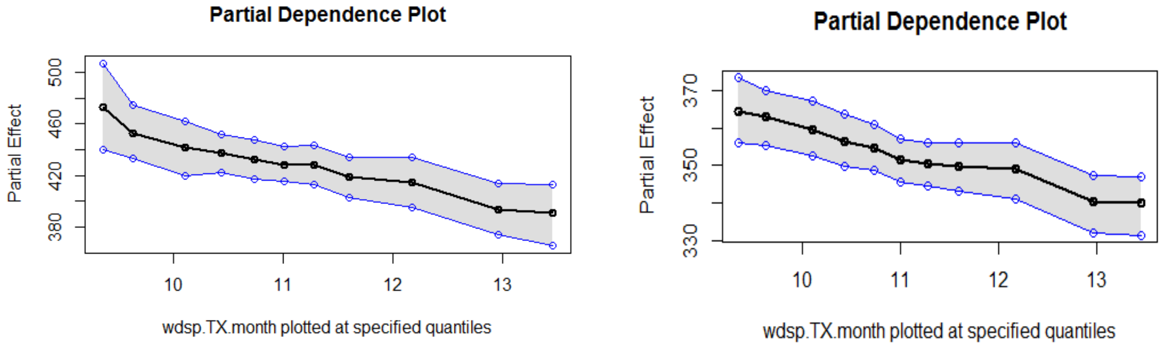


Fig 46: **Influence of WDSP on Residential and Commercial per capita consumption**

5.5. Washington

5.5.1. Model Performance

Different statistical model results for both Residual and Commercial has been show below

Table: 11 Model performance measure – RESIDENTIAL

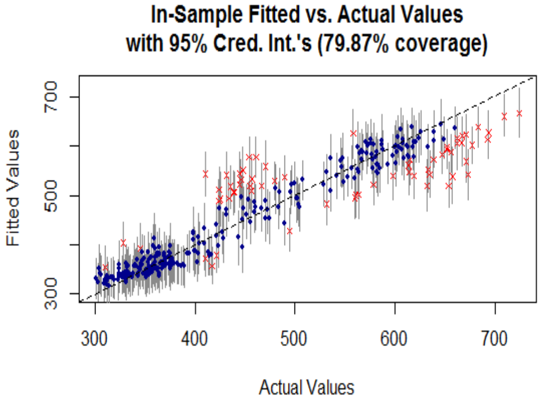
MODEL	TUNING.PARAMETERS	RMSE.TRAIN	RMSE.TEST	R.SQ
Linear	NA	84.71110	163.045	0.45253
Lasso	Default	109.33256	183.359	0.07596
GAM	Default	48.98465	56.45392	0.81915
CART	Default	50.67803	55.88078	0.80638
Bagged−CART	nbagg=500	44.31128	53.37671	0.85207
MARS.1	pmethod=backward,nfold=10,ncross=5	50.25639	67.24626	0.80965
MARS.2	pmethod=cv,nfold=10,ncross=5,degree=2	48.98797	77.72440	0.81861
MARS.3	pmethod=cv,nfold=10,ncross=5,degree=3	52.11149	53.66315	0.79523
MARS.4	pmethod=backward,nfold=10,ncross=5,degree=3,penalty=2	43.25214	78.29311	0.85902
Random Forest	ntree=1000	22.09374	51.78306	0.96321
BART	k= 2 num_trees= 50 q= 0.75 nu= 5	38.12958	50.96467	0.89038

Table: 12 Model performance measure –COMMERCIAL

MODEL	TUNING.PARAMETERS	RMSE.TRAIN	RMSE.TEST	R.SQ
Linear	NA	17.325979	25.43774	0.20975
Lasso	Default	19.154102	25.42354	0.03237
GAM	Default	12.512218	22.91520	0.58810
CART	Default	12.120720	13.28057	0.61195
Bagged−CART	nbagg=500	10.189982	12.46109	0.72674
MARS.1	pmethod=backward,nfold=10,ncross=5	12.626369	20.88146	0.58038
MARS.2	pmethod=cv,nfold=10,ncross=5,degree=2	14.761891	20.70802	0.41691
MARS.3	pmethod=cv,nfold=10,ncross=5,degree=3	14.945516	30.77913	0.40142
MARS.4	pmethod=backward,nfold=10,ncross=5,degree=3,penalty=2	10.845546	27.14060	0.69023
Random Forest	ntree=1000	5.516817	12.43244	0.91986
BART	k= 2 num_trees= 50 q= 0.9 nu= 3	8.928264	12.50752	0.78955

5.5.2. Model Diagnostics

Figure below shows the scatterplot of predicted vs actual. In case of the residential sector, the 95% credible intervals provide 79.87% coverage for all the observations. In case of the commercial sector, the 95% credible intervals provide 80.19% coverage for all the observation



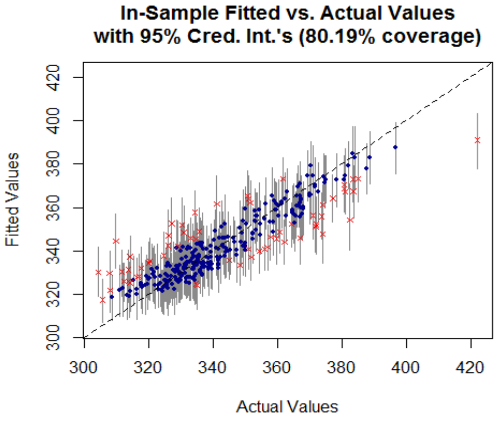
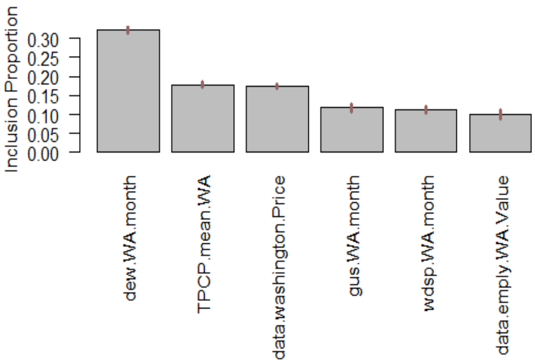


Fig47: In-Sample fitted vs. actual values: RE sector **Fig48: In-Sample fitted vs. actual values: CE sector**

5.5.3. Model inference

It can be observed from the variables important plot that the trends for both sectors with Mean Dew Point Temperature (MDPT) found to be the most important predictor followed by Total Monthly Precipitation (TPCP) and Electricity Price.



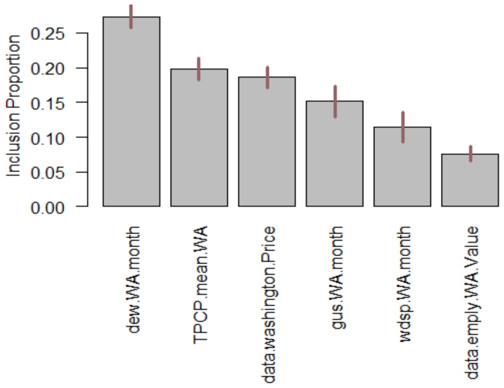
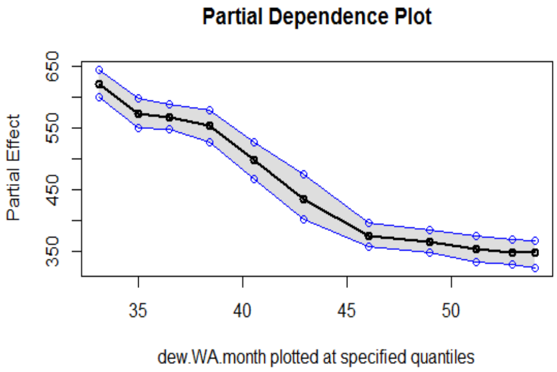


Fig49: Variable importance plot: RE sector **Fig50: Variable importance plots: CE sector**

- ***Influence of Mean Dew Point Temperature***

From the Mean Dew Point Temperature plots of both Residential and commercial it can be observed that there is a non- linear relationship with the per capita consumption (Figure51). Monthly per capita consumption of electricity (Residential) decreases continuously. Since Washington is a state with cold-climate throughout the year, the peak consumption occurs only in Winter. The Mean Dew Point Temperature corresponding to winter season ranges in between 30 and 45°F and hence we seek peak consumption at the start. Similarly, for the commercial sector, consumption decreases until 45°F. Beyond 45°F, we observe that there is a gradual increase in the consumption in commercial sector, since the commercial sector is more sensitive to temperature and humidity as it contains heavily regulated internal climatic conditions.



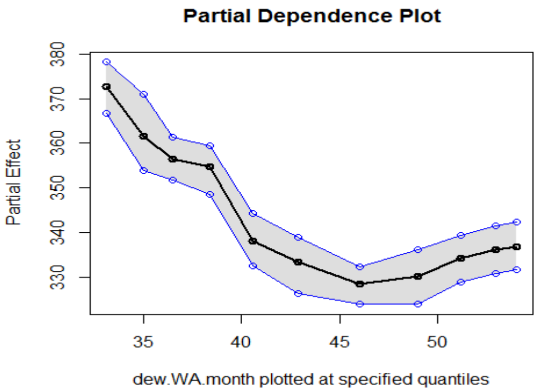


Fig51: Influence of MDPT on Residential and Commercial per capita consumption

- ***Influence of Total Monthly Precipitation***

From the Total Monthly Precipitation plots of both Residential and commercial sector it can be seen that, as the TPCP increases the consumption decreases initially for commercial and then it begins to increase for both residential and commercial after TPCP = 2. This can be explained by observing from the box plot that the peak consumption occur during the winter months in which quantum of precipitation is considerably high as compared to other months.

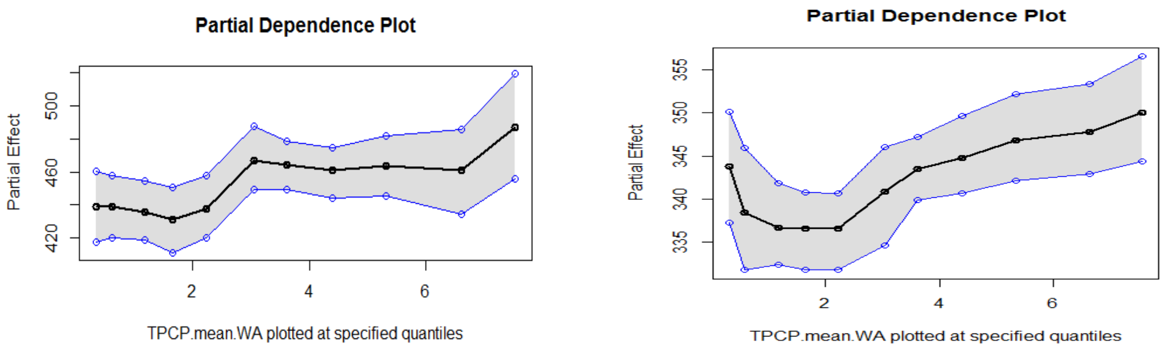


Fig52: Influence of TPCP on Residential and Commercial per capita consumption

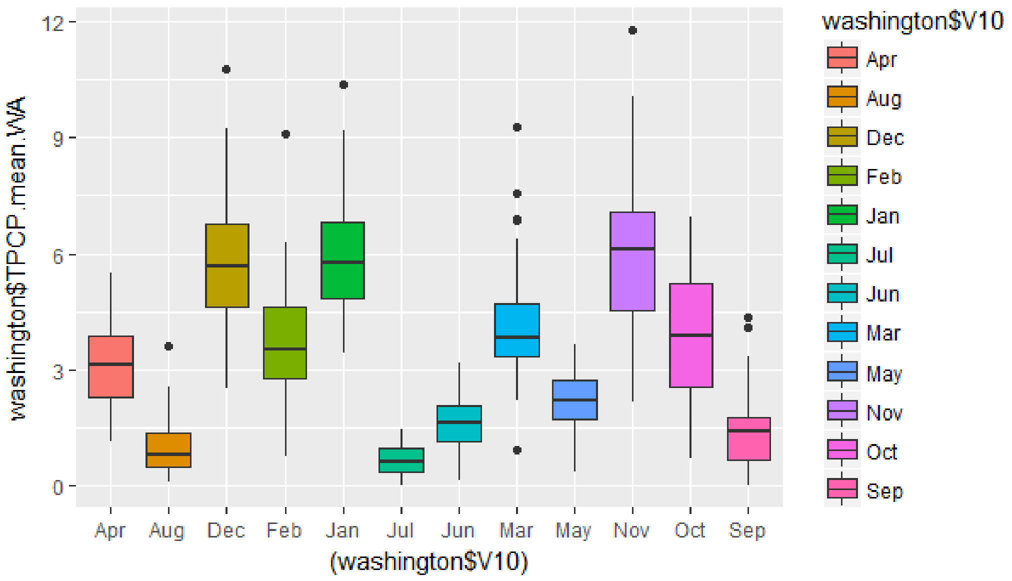


Fig53: Seasonal variation of TPCP

- ***Influence of Mean Wind Speed***

From the Mean Wind Speed plots of both Residential and commercial sector, it can be noticed that the consumption slightly decreases with increasing wind speed, especially for commercial sector. This is intuitive as increasing wind speeds takes away the need for spatial cooling.

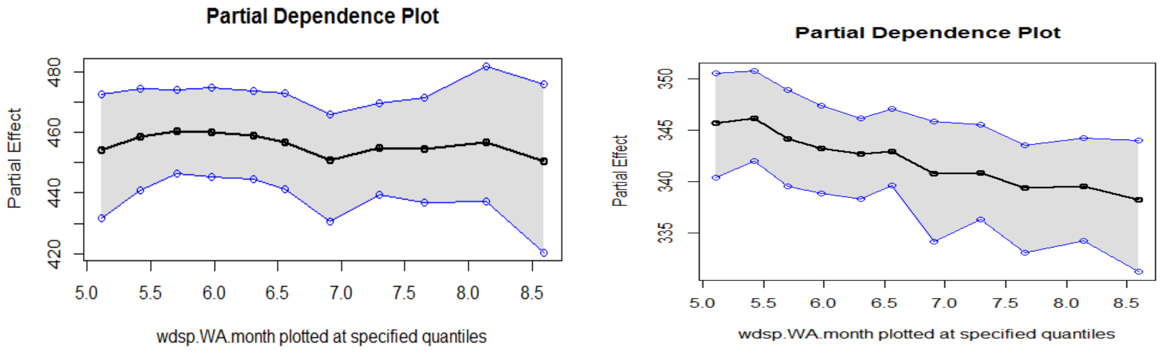


Fig54: Influence of WDSP on Residential and Commercial per capita consumption

6. Conclusions and Future Scope

The supervised machine learning techniques tested here generate accurate models with high predictive capabilities. Through them, we investigated the climate sensitivity of the end-use electricity consumption for both the residential and commercial sector. We notice that the Mean Dew-point Temperature is a significant predictor of Electricity Consumption for both the sectors and for all the states. We conclude that the Bayesian Additive Regression Trees (BART) learning method outperforms all the other statistical methods for this data.

In comparing model results across the states, we notice that no two models have the same variable importance predictors. This indicates the uniqueness pertaining to the state's geography that may not be captured in a national or large-scale regional studies of climatic influences on the electricity consumption.

As an immediate follow-up, the current work can be extended to the rest of the states in USA.

References

[1] Sayanti Mukhopadhyaya, Roshanak Nateghi, *Climate sensitivity of end-use electricity consumption in the built environment: An application to the state of Florida, United States.*

[2] Sailor DJ, Muñoz JR. Sensitivity of electricity and natural gas consumption to climate in the U.S.A.—Methodology and results for eight states. Energy 1997;22:987–98. doi:http://dx.doi.org/10.1016/S0360-5442(97)00034-0

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. *Elements*, *1*, 337–387

[4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *Springer Texts in Statistics An Introduction to Statistical Learning - with Applications in R*. https://doi.org/10.1007/978-1-4614-7138-7

[5] Ruth M, Lin A-C. Regional energy demand and adaptations to climate change: Methodology and application to the state of Maryland, USA. Energy Policy 2006;34:2820–33. doi:http://dx.doi.org/ 10.1016/j.enpol. 2005.04.016

[6] Amato AD, Ruth M, Kirshen P, Horwitz J. Regional Energy Demand Responses To Climate Change: Methodology And Application To The Commonwealth Of Massachusetts. Clim Change 2005;71:175–201. doi:10.1007/s10584-005-5931-2.

[7] Badri MA. Analysis of demand for electricity in the United States. Energy 1992;17:725– 33. doi:http://dx.doi.org/10.1016/0360-5442(92)90080-J.

[8] Mirasgedis S, Sarafidis Y, Georgopoulou E, Kotroni V, Lagouvardos K, Lalas DP. Modeling framework for estimating impacts of climate change on electricity demand at regional level: Case of Greece. Energy Convers Manag 2007;48:1737–50.