MSDS 6372

Applied Statistics: Inference and Modeling

Section 403

Data Reduction Project

Predicting Home Energy Use

July 16, 2017

Dr. Jack K. Rasmus-Vorrath
Ashwin Thota
Laurie Harris

Predicting Home Energy Use

In a 2016 study of residential appliance energy, researchers collected several data points for a residence in Stamburges, Belgium over a four-and-a-half-month period. The purpose of their study was to identify a model that could be used to predict appliance energy usage based on temperature and humidity readings inside the home as well as external weather conditions recorded from the nearest weather station in Chievres Airport, Belgium. For their approach, the researchers measured internal conditions in nine specific areas of the home using temperature and humidity sensors. In addition to the weather condition measurements and internal monitoring data, lighting energy usage was used as an explanatory variable to indicate occupancy within the home. (*Luis M. Candanedo, Veronique Feldheim, Dominique Deramaix, Data driven prediction models of energy use of appliances in a low-energy house, Energy and Buildings, Volume 140, 1 April 2017, Pages 81-97, ISSN 0378-7788*)

Problem Statement

We will analyze the total energy of the home by combining the appliance and light energy measurements. We will examine whether principal component analysis (PCA) can be useful as a data reduction technique to account for multicollinear relationships between the explanatory variables.

Constraints and Limitations

These data are restricted to what has been gathered within one single residence. Although interesting, we can make no assumptions that energy use would follow similar patterns in other residential buildings in the same area or in other localities. The data was collected using sensors within a single residential dwelling and are observational. Therefore, no causal inference can be made.

The data set includes a total of 19,735 observations of measurements recorded at 10 minute intervals for a period of approximately 4.5 months. It should be noted that all measurements were not available at precise 10-minute intervals. Internal conditions were recorded more frequently than every ten minutes, and external weather conditions were recorded less frequently. These values were imputed to align measurements with the energy measurements, which were collected at 10-minute intervals.

Data Set Description

The original data set contains 27 measurements recorded at 10-minute intervals over the period of January 11, 2016 through May 27, 2016. All measurements are numerical, with energy measurements recorded in watt-hours, temperature recorded in degrees Celsius, and humidity recorded as a percentage. External weather conditions were recorded using industry standards for the area. The original dataset also contained two random variables that were introduced by the researchers for use in their selected algorithm; however, the researchers acknowledge that these variables do not have any predictive power.

The original study considers appliance energy as the response variable and light energy usage as one of the explanatory variables to indicate room occupancy. For the purposes of this analysis, we have combined the appliance energy and light energy variables and will examine the relationship between the internal and external measurements with respect to a single total energy response. In our analysis, we also parse the "Date Time" variable into two attributes: "Day of Week" and "Time of Day". Table 1 contains a description of the variables in the total data set.

Variable	**Unit of Measure**	**Description**
Date Time	Date and Time (Hour and Minute)	Time at 10 minute intervals
Total Energy	Watt-hour	Combination of Appliance Energy and Light Energy
Appliances	Watt-hour	Energy usage for Appliances
Lights	Watt-hour	Energy usage for Lights
T1	Degrees Celsius	Temperature in Kitchen
RH_1	Percentage	Humidity in Kitchen
T2	Degrees Celsius	Temperature in Living Room
RH_2	Percentage	Humidity in Living Room
T3	Degrees Celsius	Temperature in Laundry Room
RH_3	Percentage	Humidity in Laundry Room
T4	Degrees Celsius	Temperature in Office
RH_4	Percentage	Humidity in Office
T5	Degrees Celsius	Temperature in Bathroom
RH_5	Percentage	Humidity in Bathroom
T6	Degrees Celsius	Temperature in Outside Building
RH_6	Percentage	Humidity in Outside Building
T7	Degrees Celsius	Temperature in Ironing Room
RH_7	Percentage	Humidity in Ironing Room
T8	Degrees Celsius	Temperature in Teenager Room
RH_8	Percentage	Humidity in Teenager Room
T9	Degrees Celsius	Temperature in Parents Room
RH_9	Percentage	Humidity in Parents Room
To	Degrees Celsius	Outside Temperature
Pressure	Mm HG	Outside Pressure
RH_out	Percentage	Outside Humidity
Wind Speed	m/s	Outside Wind speed
Visibility	Km	Outside Visibility
Dew point	Degrees Celsius	Outside Dew point (2)
Rv1	Numeric	Random Variable 1
Rv2	Numeric	Random Variable 2
Day of Week	Date	Parsed from Date-Time Variable
Time of Day	Hour and Minute	Parsed from Date-Time Variable

Table 1: Data variables and descriptions

Exploratory Data Analysis

Given that the dataset contains similar measurements (temperature and humidity) in several rooms within the house, we suspect there is correlation between the variables. Although we presume that many of the variables will be correlated, it is prudent to perform exploratory data analysis to visually identify those relationships. Initially, we examine the summary statistics for the raw variables in the dataset in Figure 1. We note that all non-numeric variables are appropriately converted.

The MEANS Procedure

Variable	N	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Range	N Miss	Std Dev
Appliances	19735	10.0000000	50.0000000	60.0000000	97.6949582	100.0000000	1080.00	1070.00	0	102.5248905
Press_mm_hg	19735	729.2999000	750.9333000	756.1000000	755.5225583	760.9333000	772.2999000	43.0000000	0	7.3994415
RH_1	19735	27.0233300	37.3333300	39.6566600	40.2597348	43.0666600	63.3599900	36.3366600	0	3.9792988
RH_2	19735	20.4633300	37.8999900	40.5000000	40.4204160	43.2599900	56.0266600	35.5633300	0	4.0698125
RH_3	19735	28.7666600	36.8999900	38.5300000	39.2424957	41.7599900	50.1633300	21.3966700	0	3.2545765
RH_4	19735	27.6600000	35.5300000	38.3999900	39.0268994	42.1566600	51.0900000	23.4300000	0	4.3413206
RH_5	19735	29.8150000	45.3999900	49.0900000	50.9492781	53.6633300	96.3216600	66.5066600	0	9.0220342
RH_6	19735	1.0000000	30.0233300	55.2899900	54.6090794	83.2266600	99.9000000	98.9000000	0	31.1498053
RH_7	19735	23.1999900	31.5000000	34.8633300	35.3881959	39.0000000	51.3999900	28.2000000	0	5.1142079
RH_8	19735	29.6000000	39.0666600	42.3750000	42.9361608	46.5360000	58.7800000	29.1800000	0	5.2243607
RH_9	19735	29.1666600	38.5000000	40.8999900	41.5523963	44.3428500	53.3266600	24.1600000	0	4.1514973
RH_out	19735	24.0000000	70.3333300	83.6666600	79.7504158	91.6666600	100.0000000	76.0000000	0	14.9010877
T1	19735	16.7899900	20.7600000	21.6000000	21.6865675	22.6000000	26.2600000	9.4700100	0	1.6060657
T2	19735	16.1000000	18.7899900	20.0000000	20.3412154	21.5000000	29.8566600	13.7566600	0	2.1929736
T3	19735	17.1999900	20.7899900	22.1000000	22.2676070	23.2899900	29.2360000	12.0360100	0	2.0061105
T4	19735	15.1000000	19.5300000	20.6666600	20.8553308	22.1000000	26.1999900	11.0999900	0	2.0428845
T5	19735	15.3300000	18.2749900	19.3900000	19.5921025	20.6250000	25.7950000	10.4650000	0	1.8446233
T6	19735	-6.0650000	3.6266660	7.2999990	7.9109379	11.2560000	28.2899900	34.3549900	0	6.0903445
T7	19735	15.3900000	18.6999900	20.0333300	20.2671023	21.6000000	26.0000000	10.6100000	0	2.1099932
T8	19735	16.3066600	20.7899900	22.1000000	22.0291025	23.3900000	27.2300000	10.9233400	0	1.9561618
T9	19735	14.8900000	18.0000000	19.3900000	19.4858243	20.6000000	24.5000000	9.6100000	0	2.0147122
T_out	19735	-9.8333300	1.4066660	3.5833330	3.9323662	6.5999990	9.9833330	19.8166630	0	3.4063051
Tdewpoint	19735	-9.8333300	1.2800000	3.2999990	3.1166666	5.9666660	9.9999990	19.8333290	0	3.8582974
Visibility	19735	1.0000000	29.0000000	40.0000000	38.3308321	40.0000000	66.0000000	65.0000000	0	11.7947188
Windspeed	19735	0	2.0000000	3.6666660	4.0397515	5.5000000	14.0000000	14.0000000	0	2.4512205
lights	19735	0	0	0	3.8018748	0	70.0000000	70.0000000	0	7.9359876
rv1	19735	0.0053210	12.4970500	24.8976500	24.9880294	37.5860100	49.9965200	49.9911990	0	14.4966326
rv2	19735	0.0053210	12.4970500	24.8976500	24.9880294	37.5860100	49.9965200	49.9911990	0	14.4966326
datetime	19735	1768150800	1771110600	1774071000	1774071000	1777031400	1779991200	11840400.00	0	3418288.87

Figure 1: Summary statistics for data set variables

Next, we look for visual evidence to determine if variables are correlated. Multicollinearity would suggest PCA as an appropriate step prior to regression analysis. Because the raw variables are numerous, we relied on three separate scatter plots and a heat map to show the relationships among the variables. The scatterplot representations are available in the Appendix of this document.

The heat map in Figure 2 allows one to easily identify that temperature values in the different rooms of the home are strongly correlated with each other, as are the humidity values. Furthermore, the humidity measure outside the home and the date-time of the observation are strongly correlated with temperature measurements inside the home.

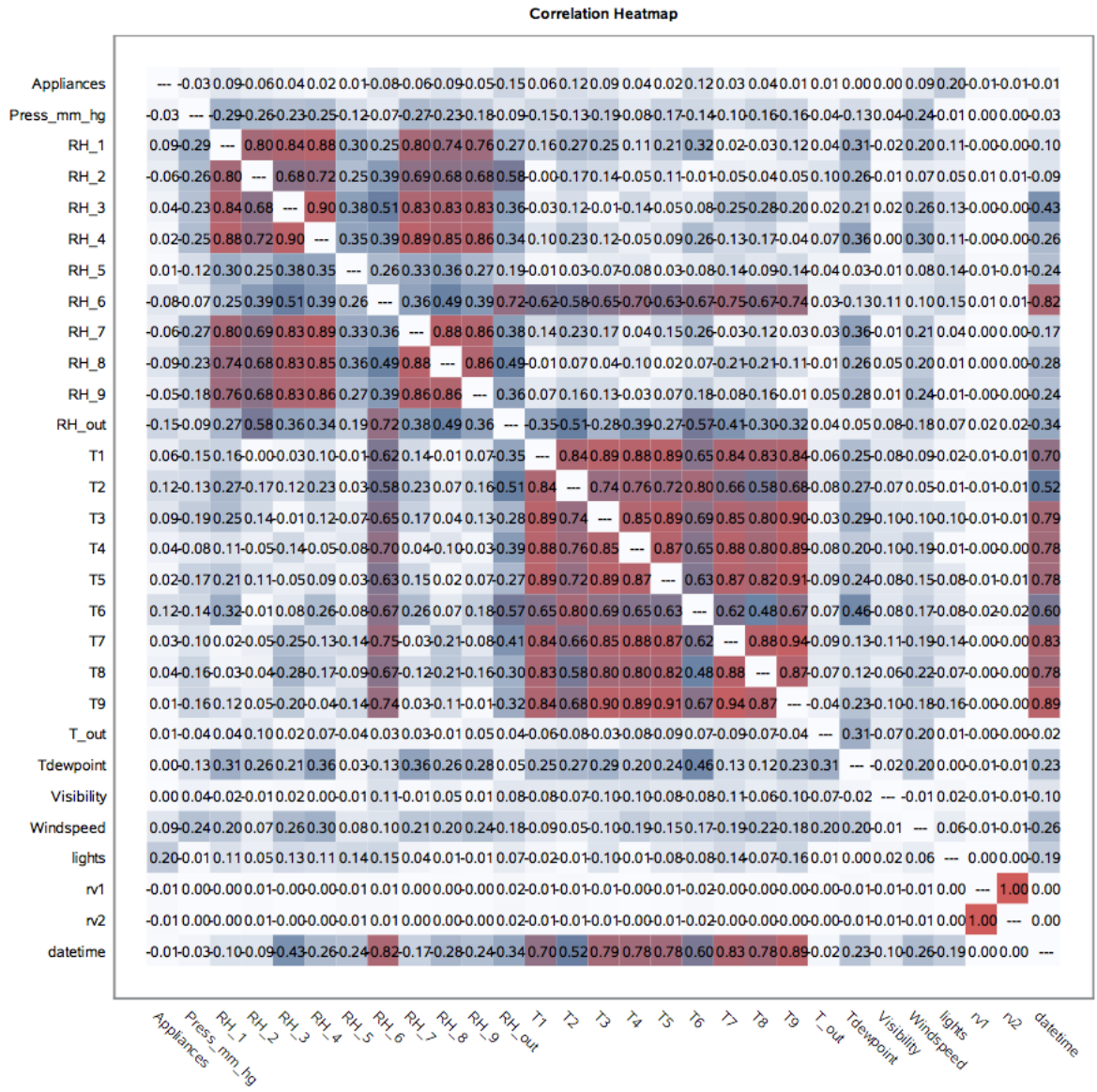


Figure 2: Heat map for variable correlations

To get a better idea of how the energy measurement fluctuates during the week, Figure 3 shows a view of energy usage according to day of the week and hour of the day. Here we note that energy use seems to peak daily during the evening and especially on weekends.

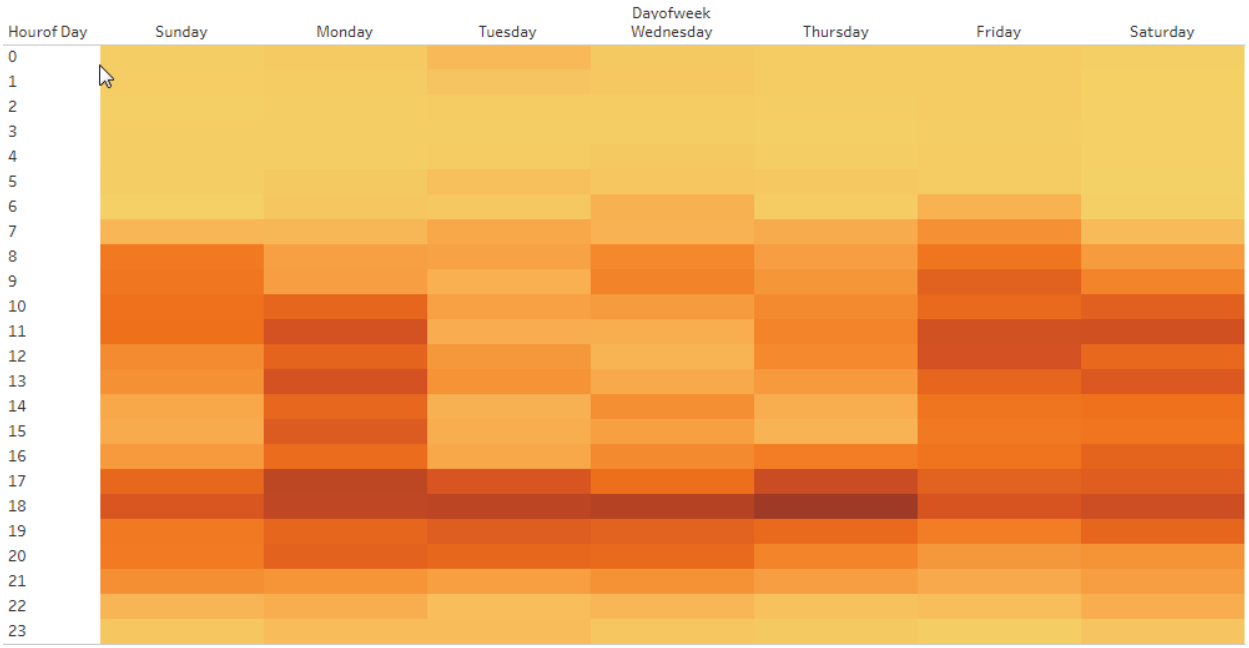


Figure 3: Heat map of energy by day of the week and hour of the day

Box plots of temperature and humidity measurements over the time period visualize variability in Figures 4 and 5. Of note are measures taken outside the building, as these appear to have extreme values over the period of measurement. Humidity is also shown to have some extreme value measurements in the bathroom. These measurements are consistent with what one would expect of activity in these areas of the home.

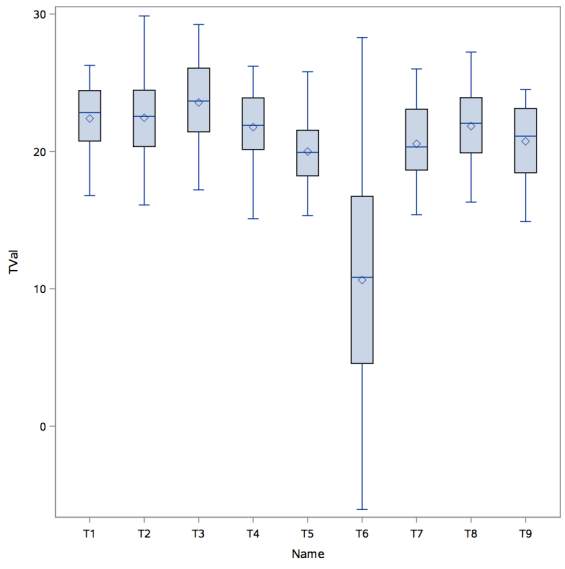


Figure 4: Distribution of temperature values by room

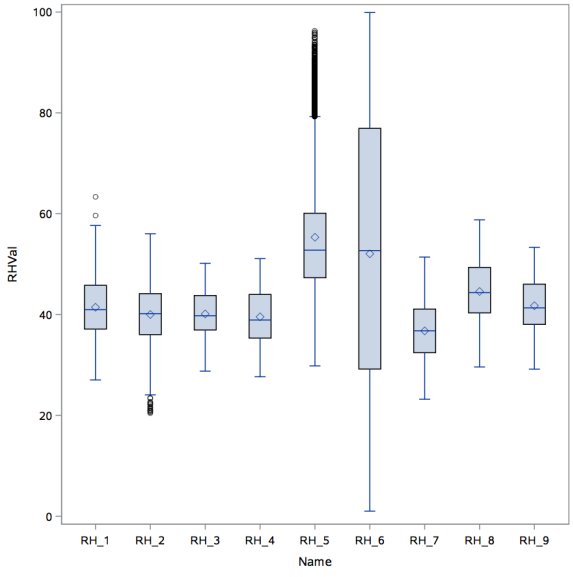


Figure 5: Distribution of humidity values by room

PCA results

Based on the results of exploratory data analysis and the evidence of variable multicollinearity, we proceed with PCA to reduce the variables used in regression analysis.

For purposes of validation, we split the data between training and test sets before proceeding with our analysis. We use a random method of splitting 70% of the observations into the training set and the remaining 30% into the test set.

We begin the Principal Component Analysis using the training data set. The scree plot indicates that variance explained appears to level off somewhat after the first seven components. However, there could be some additional benefit from including more components, as seen in Figure 6.

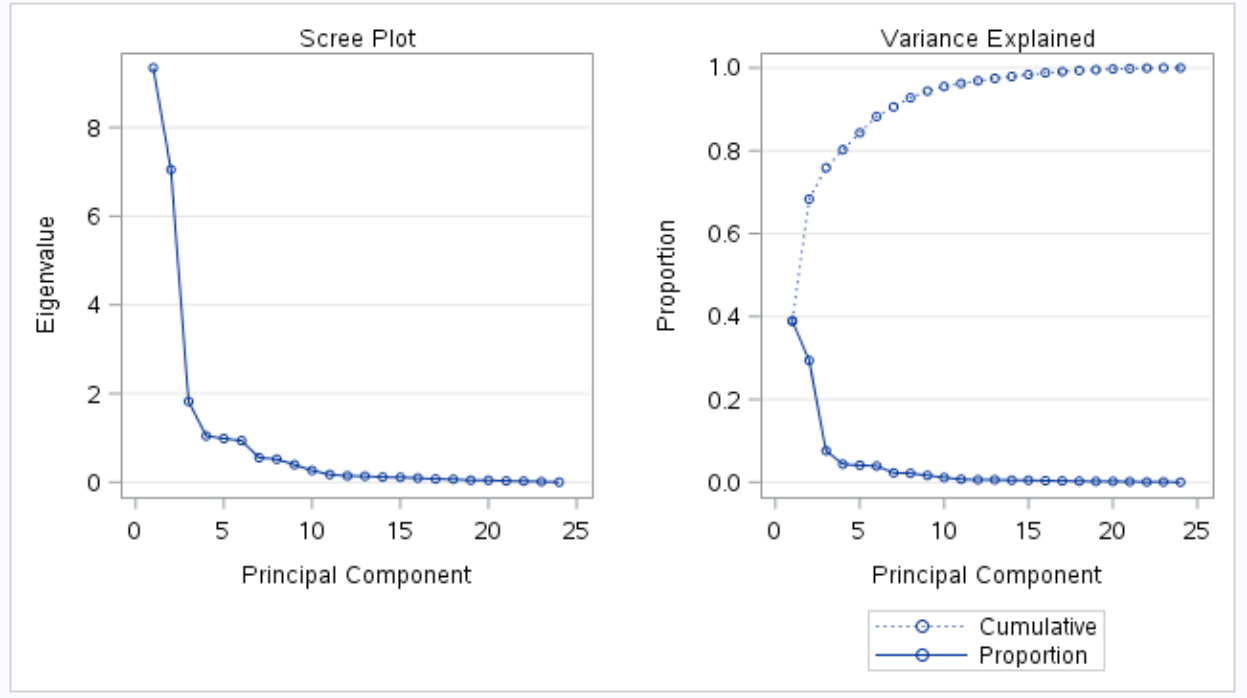


Figure 6: Scree plot and variance explained plot

As shown in Figure 7, by examining the first ten principal components, we can identify meaningful associations between the original variables. Figure 8, specifically, demonstrates the strong associations between the internal temperature measurements (T1-T9) in the first principal component and the internal humidity measurements (RH1-RH9) in the second component.

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
T1	0.300906	-.017668	0.097525	-.009147	0.002990	0.137173	-.041566	0.313358	0.019389	0.247443
RH_1	0.090295	0.329109	-.010202	0.006411	-.020363	-.052289	0.022464	-.169910	-.471683	0.264544
T2	0.280238	0.015724	-.187346	0.075542	-.067774	0.217110	-.329238	0.268558	-.000403	0.310005
RH_2	0.021453	0.303701	0.265852	-.055918	0.049613	-.223327	0.347594	-.284563	-.291859	0.027607
T3	0.303891	0.004573	0.133827	-.022874	0.023892	-.035922	0.068441	0.099272	-.017785	0.033948
RH_3	-.000233	0.349109	-.068633	0.038528	-.063201	0.084789	-.127062	0.148531	-.359226	0.065140
T4	0.296681	-.059826	0.122461	0.033280	-.052337	0.067874	-.025149	0.141116	-.043611	0.001128
RH_4	0.051677	0.356538	-.086237	0.049520	-.037793	0.010563	-.029523	0.024361	-.079918	0.174393
T5	0.298689	-.008829	0.186466	-.043070	0.006885	0.099975	0.102227	0.081432	-.008850	0.003698
RH_5	-.017525	0.152739	0.056474	-.207958	-.206262	0.829019	0.203975	-.341807	0.133424	-.052296
T6	0.276158	0.032089	-.320186	0.097929	-.006845	-.088658	-.086150	-.298845	0.137303	0.030693
RH_6	-.239731	0.208143	0.124258	-.030976	0.027627	0.057709	-.109663	0.243285	0.070499	0.187892
T7	0.293018	-.093014	0.156585	-.030217	0.007944	-.021624	0.120863	0.060578	-.089603	-.269017
RH_7	0.066972	0.345590	-.014650	0.064650	-.043325	-.012280	-.105664	0.075089	0.187998	-.205974
T8	0.266643	-.101431	0.251265	-.112678	0.098595	0.055164	0.128524	0.080437	-.051642	-.014216
RH_8	0.011245	0.346537	0.032612	0.078842	-.017037	0.032408	-.073707	0.155920	0.250071	-.318542
T9	0.303014	-.057803	0.179855	-.034204	0.042849	-.085486	0.120198	-.025402	0.015658	-.133701
RH_9	0.044614	0.336354	-.022746	0.114326	-.056201	-.056547	0.000226	0.218068	0.012621	-.521535
T_out	0.280611	0.046772	-.310977	0.094985	-.006639	-.069966	-.031452	-.254846	0.111801	-.023076
Press_mm_hg	-.064723	-.109799	-.017044	0.710978	-.463779	0.037734	0.464581	0.138629	0.002063	0.107923
RH_out	-.142264	0.194081	0.464641	-.004781	0.067705	-.124531	0.054951	0.051581	0.442993	0.336388
Windspeed	-.012113	0.104049	-.496881	-.337464	0.238920	-.027263	0.623795	0.365716	0.115813	0.108140
Visibility	-.036159	0.009352	0.022084	0.503515	0.803173	0.295285	0.035781	-.053044	-.062926	-.029532
Tdewpoint	0.236663	0.201933	-.039673	0.098150	0.039413	-.186550	0.047911	-.289448	0.421319	0.229946

Figure 7: Eigenvector loadings for the first 10 principal components

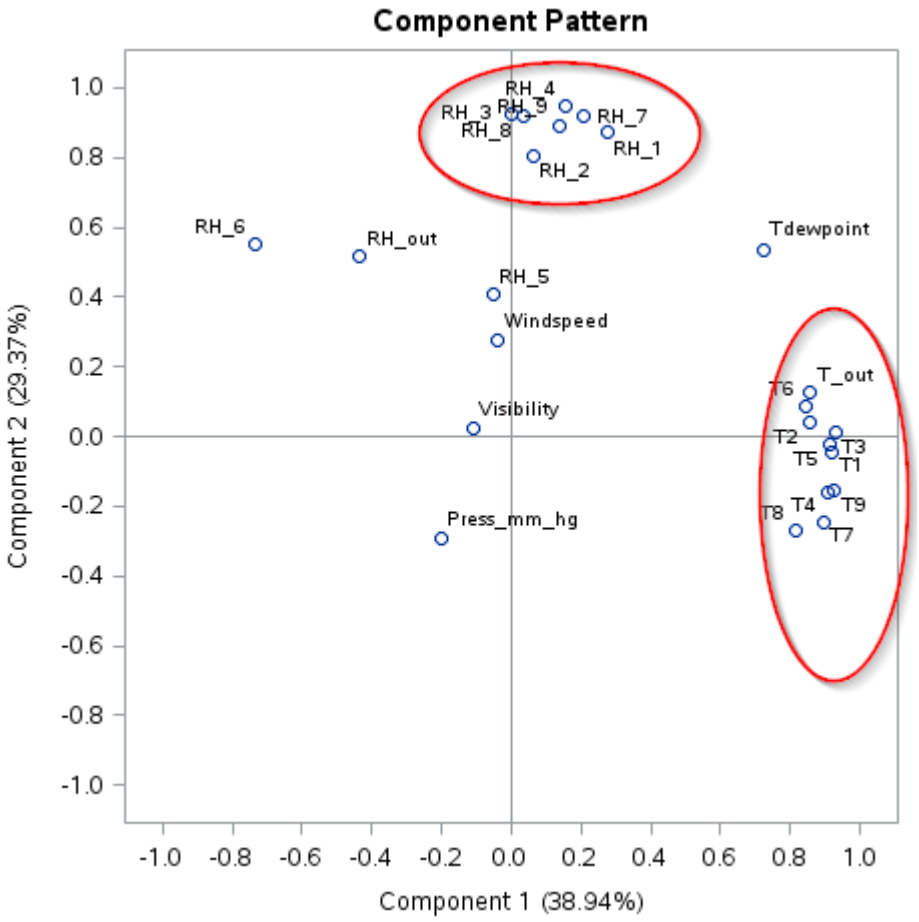


Figure 8: Component pattern for Principal components 1 and 2

Regression Analysis

In proceeding with regression, we observe, in Figure 9, that the response variable, total energy, is right-skewed and does not follow a normal distribution. This distribution is great improved by transforming the response using a base-10 logarithm, as seen in Figure 10.

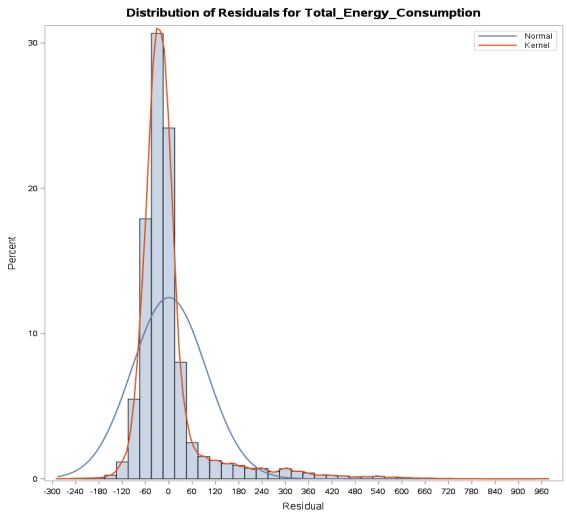


Figure 9: Distribution of Total Energy Raw variable

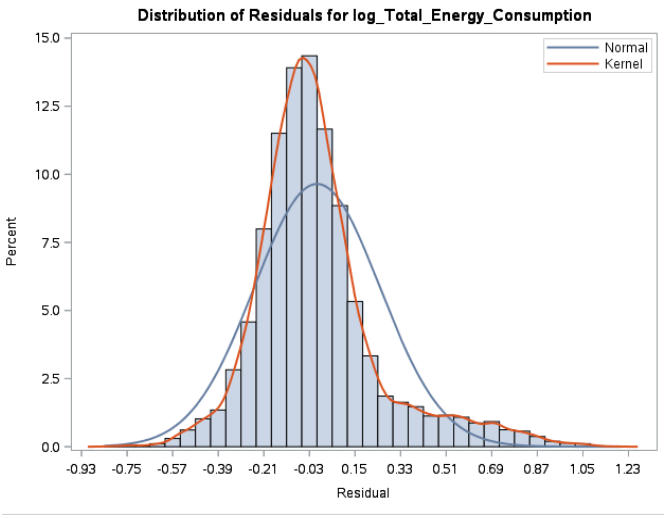


Figure 10: Distribution of Total Energy Raw – base 10 log transformation

Our PCR model for the training set data uses the first 10 principal components in addition to the day of the week and hour of the day. Using this model, we arrive at a p-value of <0.0001 and an R-Square value of 0.319 with Root Mean Square Error of 0.238, as seen in Figure 11.

The GLM Procedure
Dependent Variable: log_Total_Energy_Consumption

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	39	360.411483	9.241320	163.98	<.0001
Error	13632	768.265588	0.056358		
Corrected Total	13671	1128.677071			

R-Square	Coeff Var	Root MSE	log_Total_Energy_Consumption Mean
0.319322	12.58014	0.237397	1.887081

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Prin1	1	27.1010914	27.1010914	480.88	<.0001
Prin2	1	5.1340305	5.1340305	91.10	<.0001
Prin3	1	26.2577602	26.2577602	465.91	<.0001
Prin4	1	8.5272539	8.5272539	151.31	<.0001
Prin5	1	1.2957417	1.2957417	22.99	<.0001
Prin6	1	12.3329009	12.3329009	218.83	<.0001
Prin7	1	0.0635553	0.0635553	1.13	0.2883
Prin8	1	3.2497105	3.2497105	57.66	<.0001
Prin9	1	41.5047055	41.5047055	736.45	<.0001
Prin10	1	76.9536135	76.9536135	1365.45	<.0001
dayofweek	6	10.7623670	1.7937278	31.83	<.0001
HourofDay	23	147.2287521	6.4012501	113.58	<.0001

Figure 11: Regression analysis using training data set on first 10 principal components

We examine the residual plots produced by this model in Figure 12 to confirm that the basic assumptions of multiple linear regression are met. Although we have some reservations regarding residual normality, the large number of observations makes this issue less problematic. To the same extent, satisfaction of the assumption of constant variance in the regression model remains questionable. Moreover, even in making adjustments for serial correlation, adequate independence of the observations remains a matter of concern. Since many of the raw variable values were imputed to align with energy measurements taken at 10-minute intervals, we cannot validate with full confidence the calculations producing these results.

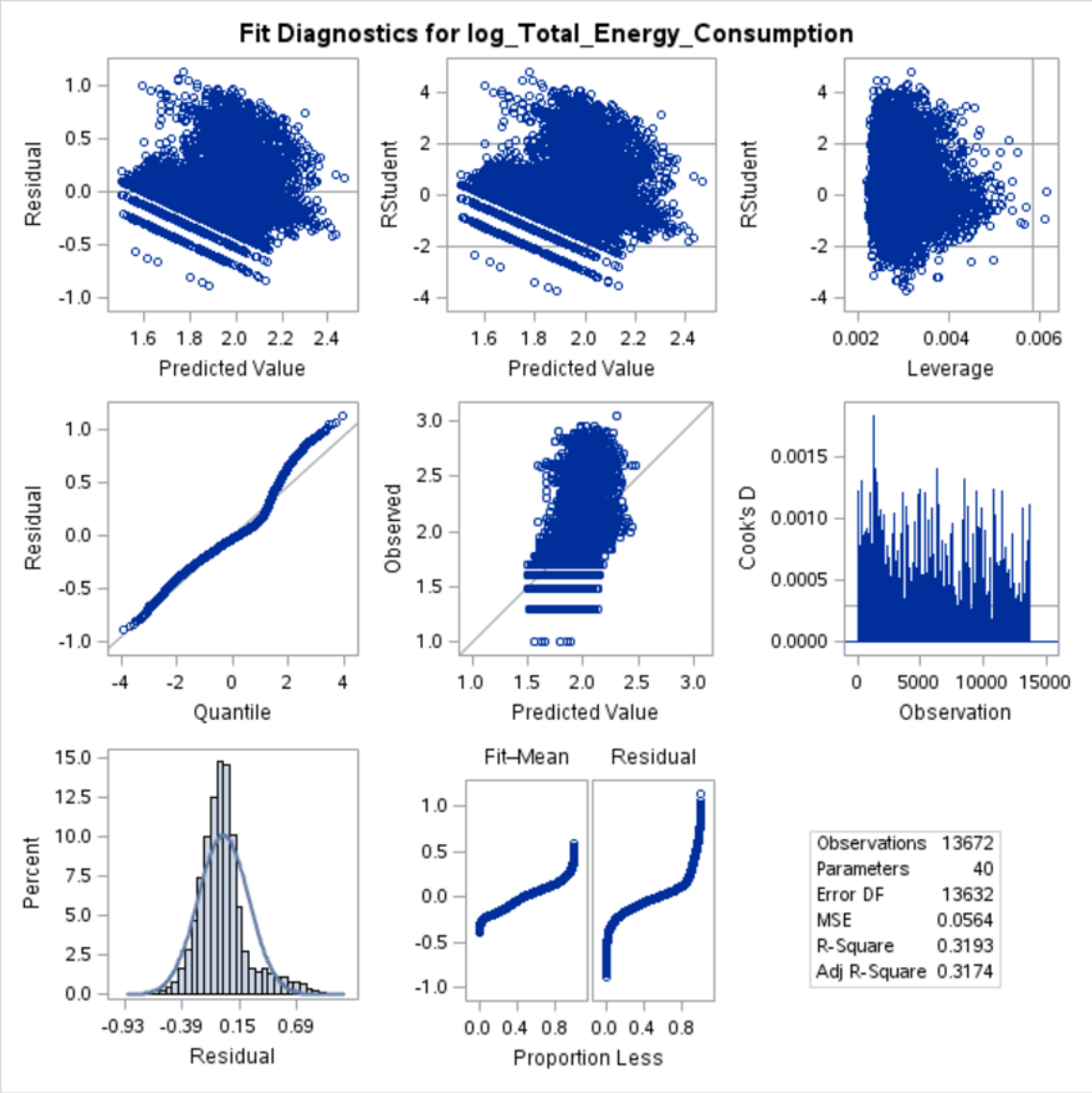


Figure 12: Residuals plots for PCA regression on training data set.

Noting the above reservations regarding model assumptions, we move forward with our analysis by applying the model from the training set to the test data set, which produces similar statistics: an R-Square of 0.29 and root mean square error of 0.245, as seen in Figure 13.

The GLM Procedure
Dependent Variable: log_Total_Energy_Consumption

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	39	148.0533530	3.7962398	63.32	<.0001
Error	6023	361.1163701	0.0599562		
Corrected Total	6062	509.1697232			

R-Square	Coeff Var	Root MSE	log_Total_Energy_Consumption Mean
0.290774	13.00305	0.244860	1.883094

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PC1	1	0.0040829	0.0040829	0.07	0.7941
PC2	1	0.1119839	0.1119839	1.87	0.1718
PC3	1	0.0185087	0.0185087	0.31	0.5785
PC4	1	0.1324229	0.1324229	2.21	0.1373
PC5	1	0.2003840	0.2003840	3.34	0.0676
PC6	1	0.0509271	0.0509271	0.85	0.3568
PC7	1	1.2533445	1.2533445	20.90	<.0001
PC8	1	7.3993353	7.3993353	123.41	<.0001
PC9	1	0.2069755	0.2069755	3.45	0.0632
PC10	1	0.0199369	0.0199369	0.33	0.5642
dayofweek	6	3.1410467	0.5235078	8.73	<.0001
HourofDay	23	135.5144045	5.8919306	98.27	<.0001

Figure 13: Regression analysis using training data set on first 10 principal components

As a final means of examining the data, we employ PCR using time-lagged cross validation to determine if this improves the residuals plots. As shown in Figure 14, there does not appear to be any substantial improvement using this method.

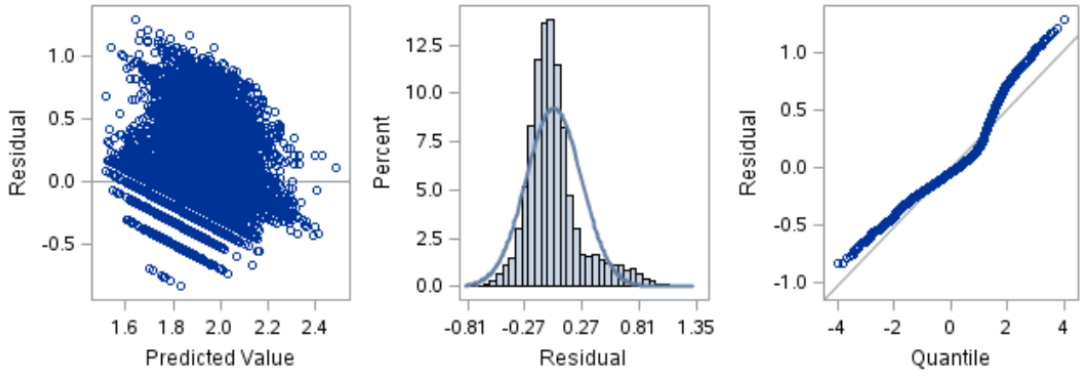


Figure 14: Residual plots using SAS PROC PLS and time lagged cross validation

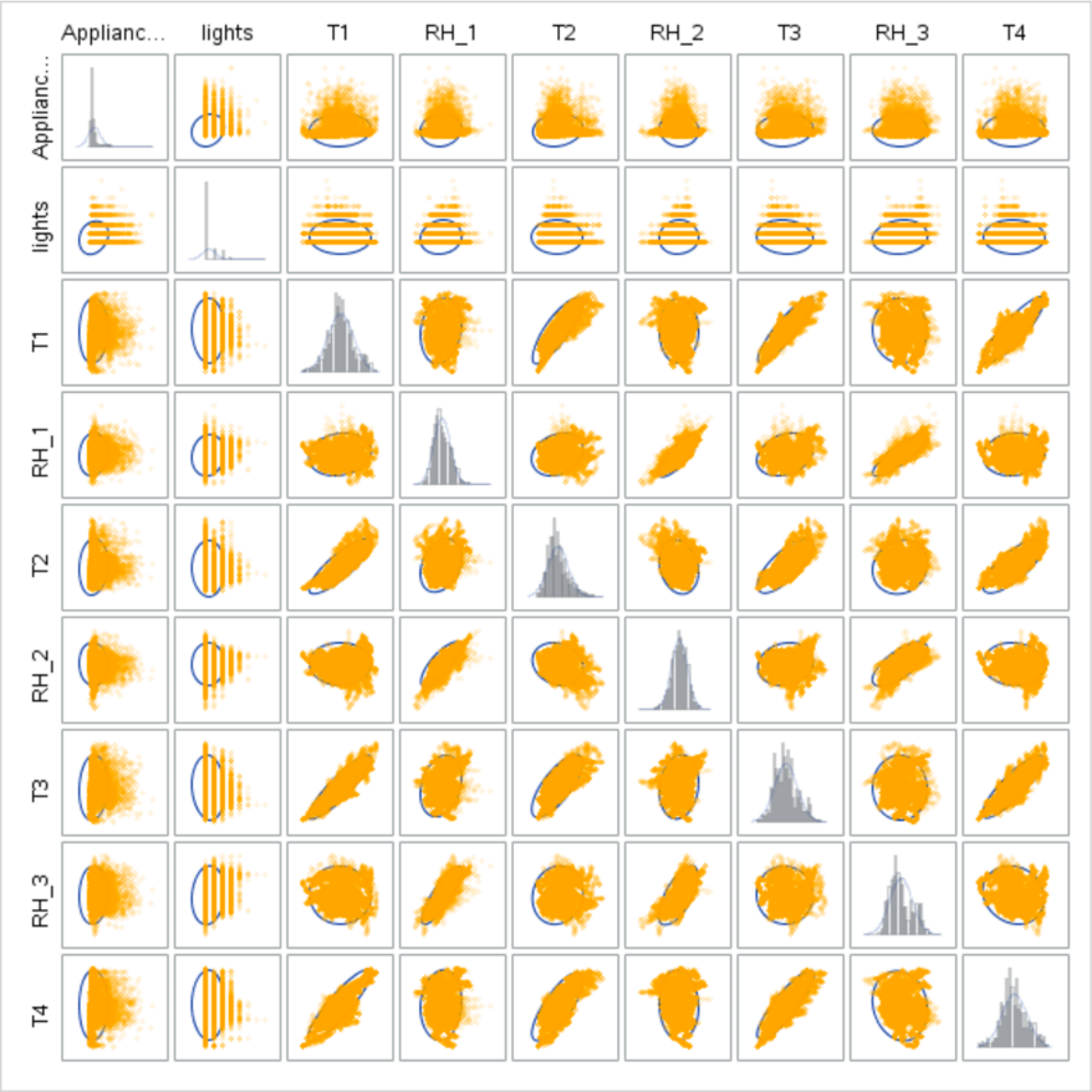
Although the R-Square statistic returned by this modeling process remains modest, applying transformations on the response resulted in an improvement on the regression statistics generated by the analyses of the original study, which also did not employ PCA to gain insight into explanatory variable relationships.

Conclusion

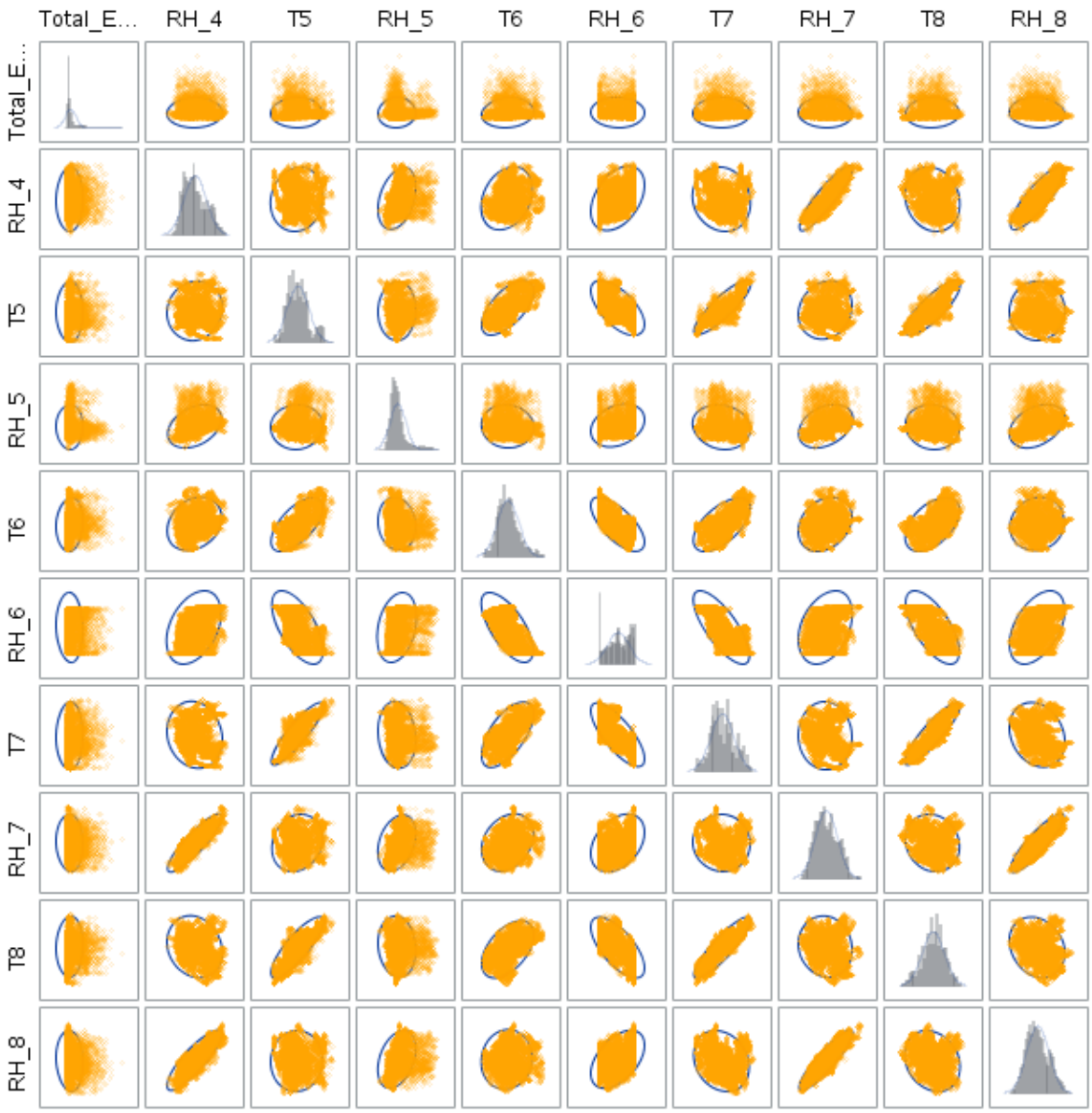
The authors of the original study considered four models in their analysis: linear regression, random forest, support vector machines and gradient boosting machines. In the original study, the authors found the linear regression model to be the least predictive in explaining appliance energy usage (R-Square: .18). In combining the appliance and light energy variables, our focus was in exploring the explanatory variable relations to the extent that their linear combinations as principle components explained the variance in total energy usage. Although PCR does not account for as much variance as the other algorithms, PCA does shed unique light on the multivariate complexity of the most important factors in considering the question of energy usage. The analysis confirms the relative importance of temperature and humidity throughout different parts of the household, particularly in primary living spaces (living rooms) and in frequently occupied areas with high-energy, heat- and moisture-generating appliances (kitchens, bathrooms). Moreover, though they are difficult to interpret, subordinate components in the hierarchy provide further insight into the relation between diurnal and seasonal outdoor weather patterns (outdoor temperature/humidity measurements as well as air conditions/quality) and corresponding indoor patterns of energy-consuming human activity.

Appendix

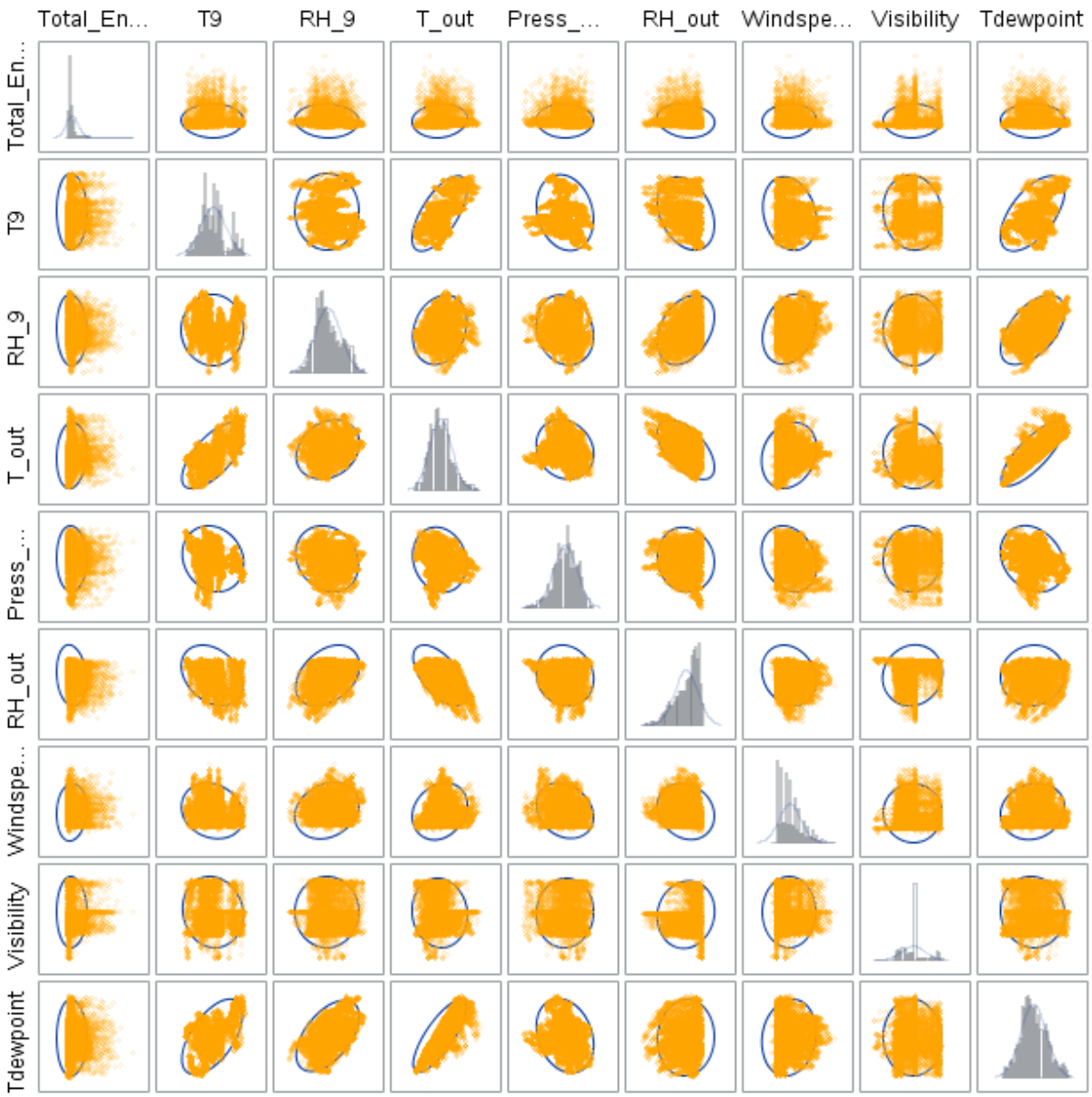
Scatter plots, comparing variable by pairs, follow.



Appendix Figure 1: Pair plots for variables Appliance, Lights, T1-T4, and R1-R3



Appendix Figure 2: Pair plots for variables Total Energy, RH4-RH8,



Appendix Figure 3: Pair plots for Energy, T9, RH9 and external weather conditions

SAS Data Code is split into two parts. The first part considers analysis on the entire data set, converting all variables to appropriate numeric equivalets. The second part parses one of the original variables, Date Time, into two separate attributes, Day of Week and Time of Day, and also splits the original data set into testing and training sets.

Part 1 of 2

```
FILENAME REFFILE '/home/...../energydata_complete.csv';

PROC IMPORT DATAFILE=REFFILE
        DBMS=CSV
        OUT=PCR_energy;
        GETNAMES=YES;
RUN;

PROC CONTENTS DATA=PCR_energy; RUN;

*Data Cleaning: Formatting variable types;

proc contents data=PCR_energy out=vars(keep=name type) noprint;

data vars; set vars;
if type=2 and name ne 'date';
newname=trim(left(name))||"_n";

options symbolgen;

proc sql noprint;
select trim(left(name)), trim(left(newname)),
     trim(left(newname))||'='||trim(left(name))
into :c_list separated by ' ', :n_list separated by ' ',
    :renam_list separated by ' '
from vars;
quit;

data PCR_energy2; set PCR_energy;
array ch(*) $ &c_list;
array nu(*) &n_list;
do i = 1 to dim(ch);
  nu(i)=input(ch(i),8.);
end;
drop i &c_list;
rename &renam_list;
run;
```

```
proc contents data = PCR_energy2; run;

data PCR_energy3; set PCR_energy2;
*Convert the character string to SAS datetime value;
datetime =input(date, anydtdtm19.);
*Apply format to the SAS date time value;
format datetime dateampm17.;        *to convert to seconds, use anydtdtm19. as format;
drop date;                          *to convert to numeric date, use dateampm17. format;
run;

proc contents data = PCR_energy3; run;

proc means data = PCR_energy3 n min q1 median mean q3 max range nmiss std var kurt skew;
run;

*EDA: Exploring variable distributions and relationships;

proc sgscatter data= PCR_energy3;
matrix Appliances lights T1 RH_1 T2 RH_2 T3 RH_3 T4 /
        ellipse diagonal=(histogram normal) markerattrs = (color = orange) transparency=0.8;
run;

proc sgscatter data= PCR_energy3;
matrix Appliances lights RH_4 T5 RH_5 T6 RH_6 T7 RH_7 T8 RH_8 /
        ellipse diagonal=(histogram normal) markerattrs = (color = orange) transparency=0.8;
run;

proc sgscatter data= PCR_energy3;
matrix Appliances lights T9 RH_9 T_out Press_mm_hg RH_out Windspeed Visibility Tdewpoint /
        ellipse diagonal=(histogram normal) markerattrs =(color = orange) transparency=0.8;
run;

proc corr data= PCR_energy3 outp = corr(where=(_type_='CORR')) noprint;
run;

proc transpose data=corr(rename=(_name_=RowLab))
        out=t(rename=(col1=Value _name_=ColLab));
by notsorted rowlab;
run;

data t2; length v $ 5; set t;
v = ifc(-1e-8 le value le 1e-8, ' 0', ifc(rowlab eq collab, ' ---', put(value, 5.2)));
AbsValue = ifn(rowlab eq collab, 0, abs(value));
run;
```

```
ods graphics on / height=9in width=9in;

proc sgplot noautolegend;
title h=7pt 'Correlation Heatmap';
heatmapparm y=rowlab x=collab colorresponse= absvalue / colormodel=(cxFAFBFE cx667FA2
cxD05B5B);
text y=rowlab x=collab text=v;
%let opts = display=(nolabel noticks) valueattrs=(size=7) offsetmin=0.05 offsetmax=0.05;
xaxis &opts;
yaxis &opts reverse;
run;

title;

PROC SQL;
create table T_data as
select * from
(SELECT t1 as TVal, 'T1' as Name FROM PCR_energy3
union
SELECT t2 as TVal, 'T2' as Name FROM PCR_energy3
union
SELECT t3 as TVal, 'T3' as Name FROM PCR_energy3
union
SELECT t4 as TVal, 'T4' as Name FROM PCR_energy3
union
SELECT t5 as TVal, 'T5' as Name FROM PCR_energy3
union
SELECT t6 as TVal, 'T6' as Name FROM PCR_energy3
union
SELECT t7 as TVal, 'T7' as Name FROM PCR_energy3
union
SELECT t8 as TVal, 'T8' as Name FROM PCR_energy3
union
SELECT t9 as TVal, 'T9' as Name FROM PCR_energy3) as T_data order by name;
QUIT;

proc sgplot data= T_data;
vbox TVal / category= Name;
run;

PROC SQL;
create table RH_data as
select * from
```

```
(SELECT rh_1 as RHVal, 'RH_1' as Name FROM PCR_energy3
union
SELECT rh_2 as RHVal, 'RH_2' as Name FROM PCR_energy3
union
SELECT rh_3 as RHVal, 'RH_3' as Name FROM PCR_energy3
union
SELECT rh_4 as RHVal, 'RH_4' as Name FROM PCR_energy3
union
SELECT rh_5 as RHVal, 'RH_5' as Name FROM PCR_energy3
union
SELECT rh_6 as RHVal, 'RH_6' as Name FROM PCR_energy3
union
SELECT rh_7 as RHVal, 'RH_7' as Name FROM PCR_energy3
union
SELECT rh_8 as RHVal, 'RH_8' as Name FROM PCR_energy3
union
SELECT rh_9 as RHVal, 'RH_9' as Name FROM PCR_energy3) as RH_data order by name;
QUIT;

proc sgplot data= RH_data;
vbox RHVal / category= Name;
run;

*Horizontal sum to create single energy parameter as response variable (log-transformed);

data PCR_energy4; set PCR_energy3;
Energy = Appliances + lights;
log_Energy = log10(Energy);
run;

*PCA with 15 PCs --> .1898 Adjusted R^2;

proc princomp plots=all  data=PCR_energy4 out= PCR n = 15;
var datetime Press_mm_hg RH_1 RH_2 RH_3 RH_4 RH_5 RH_6 RH_7 RH_8 RH_9 RH_out T1 T2
T3 T4 T5 T6 T7 T8 T9 T_out Tdewpoint Visibility Windspeed rv1 rv2
;
run;

proc reg data= PCR plots(maxplots=none) = all;
model Energy = Prin1-Prin15;
run;

*Improved Performance with log-transformation of right-skewed Energy variable;
```

```
proc reg data = PCR plots(maxplots=none) = all;
model log_Energy = Prin1-Prin15;
run;

*PCA with all PCs --> .2562 Adjusted R^2;

proc princomp plots=all  data=PCR_energy4 out= PCR_2 n = 27;
var datetime Press_mm_hg RH_1 RH_2 RH_3 RH_4 RH_5 RH_6 RH_7 RH_8 RH_9 RH_out T1 T2
T3 T4 T5 T6 T7 T8 T9 T_out Tdewpoint Visibility Windspeed rv1 rv2;
run;

proc reg data = PCR_2 plots(maxplots=none) = all;
model log_Energy = Prin1-Prin27;
run;

*Comparison with PROC REG output --> .2562 Adjusted R^2;

proc reg data = PCR_energy4 plots(maxplots=none) = all;
model log_Energy = datetime Press_mm_hg RH_1 RH_2 RH_3 RH_4 RH_5 RH_6 RH_7 RH_8
RH_9 RH_out T1 T2 T3 T4 T5 T6 T7 T8 T9 T_out Tdewpoint Visibility Windspeed rv1 rv2;
run;

*Blocked Cross-Validation to account for serial correlation-- default lag of n = 7;

proc pls data=PCR_energy4 method=pcr cv = block plots = all;
model log_Energy = datetime Press_mm_hg RH_1 RH_2 RH_3 RH_4 RH_5 RH_6 RH_7 RH_8
RH_9 RH_out T1 T2 T3 T4 T5 T6 T7 T8 T9 T_out Tdewpoint Visibility Windspeed rv1 rv2
;
run;

*Comparison with random forest output on most significant contributing individual variables;

proc hpforest data = PCR_energy4;
target log_Energy / level = interval;
input datetime Press_mm_hg RH_1 RH_2 RH_3 RH_4 RH_5 RH_6 RH_7 RH_8 RH_9 RH_out T1
T2 T3 T4 T5 T6 T7 T8 T9 T_out Tdewpoint Visibility Windspeed rv1 rv2 / level = interval;
run;
```

Part 2 of 2

```
/*************************************
Created 2 new variables in excel
->dayofweek - This will have week names Sunday to Saturday
-> HourofDay - THis has the truncated value of the hour from the timestamp
*********************************************/
/***********Changing Data types*************/
        data WORK.ENERGY    ;
         infile '/home/llharris0/Homework/Stats2/energydata_complete2.csv' delimiter = ','
MISSOVER DSD  firstobs=2 ;
            informat date $21. ;
            informat Appliances  ;
            informat lights  ;
            informat T1  ;
            informat RH_1  ;
            informat T2  ;
            informat RH_2  ;
            informat T3  ;
            informat RH_3  ;
            informat T4 ;
            informat RH_4  ;
            informat T5  ;
            informat RH_5  ;
            informat T6 ;
           informat RH_6 ;
           informat T7 ;
           informat RH_7 ;
           informat T8 ;
           informat RH_8  ;
           informat T9  ;
           informat RH_9  ;
           informat T_out  ;
           informat Press_mm_hg  ;
           informat RH_out  ;
           informat Windspeed  ;
           informat Visibility  ;
           informat Tdewpoint  ;
           informat rv1  ;
           informat rv2  ;
           informat dayofweek $21.;
           informat HourofDay bestd20.19 ;

           format date $21. ;
           format Appliances  ;
```

```
  format lights bestd20.19  ;
  format T1 bestd20.19  ;
  format RH_1 bestd20.19  ;
  format T2 bestd20.19  ;
  format RH_2 bestd20.19  ;
  format T3 bestd20.19  ;
  format RH_3 bestd20.19  ;
  format T4 bestd20.19  ;
  format RH_4 bestd20.19  ;
  format T5 bestd20.19 ;
  format RH_5 bestd20.19  ;
  format T6 bestd20.19  ;
  format RH_6 bestd20.19  ;
  format T7 bestd20.19  ;
  format RH_7 bestd20.19  ;
  format T8 bestd20.19  ;
  format RH_8 bestd20.19  ;
  format T9 bestd20.19  ;
  format RH_9 bestd20.19  ;
  format T_out bestd20.19  ;
  format Press_mm_hg bestd20.19  ;
  format RH_out bestd20.19  ;
  format Windspeed bestd20.19  ;
  format Visibility bestd20.19  ;
  format Tdewpoint bestd20.19 ;
  format rv1 bestd20.19 ;
  format rv2 bestd15.14 ;
  format dayofweek $21.;
  format HourofDay bestd20.19 ;
input
        date $
        Appliances
        lights
        T1
        RH_1
        T2
        RH_2
        T3
        RH_3
        T4
        RH_4
        T5
        RH_5
        T6
```

```
                RH_6
                T7
                RH_7
                T8
                RH_8
                T9
                RH_9
                T_out
                Press_mm_hg
                RH_out
                Windspeed
                Visibility
                Tdewpoint
                rv1
                rv2
                dayofweek $
                HourofDay
        ;
        /***Randonmly partitioning data into training and test data****/
        if ranuni(12345) < 0.7 then set="TRAINING";
        else set = "TESTING";
       run;

 data energy;
 set energy;
 Total_Energy_Consumption=Appliances+lights;
 run;
 /**************Splitting data into Train***************/
 data Energy_train;
 set energy;
if set = "TESTING" then delete;
run;

/************Splitting data into Test*****************/
data Energy_test;
 set energy;
if set = "TRAINING" then delete;
run;

data corr_energy_train ;/*Creating a new Train dataset to pass to IML*/
  set energy_train(keep=Total_Energy_Consumption T1 RH_1 T2 RH_2 T3 RH_3 T4 RH_4 T5
RH_5 T6 RH_6 T7 RH_7 T8 RH_8 T9 RH_9 T_out Press_mm_hg RH_out Windspeed
Visibility Tdewpoint dayofweek HourofDay);
  run;
```

```
proc corr data=work.corr_energy_train outp=corr(where=(_type_='CORR')) noprint;
run;

/************************ Understand relationship between the Weekday/ Time of Day
and Energy Consumption*******
I wasn't able to figureout how to create the below multivariable heap map in SAS. I did this in
Tableau instead.
Below is the link for this:
https://public.tableau.com/profile/ashwin.thota#!/vizhome/Applianceenergyconsumption-
Heapmap/Sheet1?publish=yes
************************************/

/********* regression analysis*************/

data energy_train;
set energy_train;
log_Total_Energy_Consumption=log10(Total_Energy_Consumption);
run;

proc reg data = energy_train PLOTS(MAXPOINTS=NONE)= all;

model log_Total_Energy_Consumption = lights T1 RH_1 T2 RH_2 T3 RH_3 T4 RH_4 T5 RH_5 T6
RH_6 T7 RH_7 T8 RH_8 T9 RH_9 T_out Press_mm_hg RH_out Windspeed
Visibility Tdewpoint / vif;
run;


proc glmselect data=energy_train PLOTS = all;
class dayofweek HourofDay;
model log_Total_Energy_Consumption =  T1 RH_1 T2 RH_2 T3 RH_3 T4 RH_4 T5 RH_5 T6 RH_6
T7 RH_7 T8 RH_8 T9 RH_9 T_out Press_mm_hg RH_out Windspeed
Visibility Tdewpoint dayofweek HourofDay/selection=lasso(stop=cv) cvmethod=random(5)
stats=adjrsq;
run;

proc glm data=energy_train plots(maxpoints=none)=all;
class HourofDay;
model log_Total_Energy_Consumption=

T2
T3
T6
RH_8
RH_out
```

```
HourofDay;
run;

proc princomp plots=all data=energy_train out=pca;
    var T1 RH_1 T2 RH_2 T3 RH_3 T4 RH_4 T5 RH_5 T6 RH_6 T7 RH_7 T8 RH_8 T9 RH_9 T_out
Press_mm_hg RH_out Windspeed Visibility Tdewpoint;
    run;
proc glm data=pca plots(maxpoints=none)=all;
class dayofweek HourofDay;
          model log_Total_Energy_Consumption= prin1-prin10 dayofweek HourofDay;
          run;

/**********************PCA*************************************/

ods output Eigenvectors=Output ;       /* the data set name is 'Output' this should hold the
Eigen vectors from Train*/
proc princomp data=energy_train out=pca1;
var T1 RH_1 T2 RH_2 T3 RH_3 T4 RH_4 T5 RH_5 T6 RH_6 T7 RH_7 T8 RH_8 T9 RH_9 T_out
Press_mm_hg RH_out Windspeed
Visibility Tdewpoint;
run;

proc print data=Output noobs/*noobs suppresses Obs numbers in the output*/;
run;

data energy_train_iml;/*Creating a new Train dataset to pass to IML*/
  set energy_train(keep=T1 RH_1 T2 RH_2 T3 RH_3 T4 RH_4 T5 RH_5 T6 RH_6 T7 RH_7 T8 RH_8
T9 RH_9 T_out Press_mm_hg RH_out Windspeed
Visibility Tdewpoint);
  run;

/*********** Extracting PC's from Test Data************/
proc iml;
use Output;
read all var {Prin1       Prin2    Prin3    Prin4    Prin5 Prin6 Prin7 Prin8 Prin9 Prin10} into
Test_Vector;/*Reasing first 5 significant PC's in to a matrix of 25x5*/
use energy_train_iml;
read all var {T1 RH_1 T2 RH_2 T3 RH_3 T4 RH_4 T5 RH_5 T6 RH_6 T7 RH_7 T8 RH_8 T9 RH_9
T_out Press_mm_hg RH_out Windspeed
Visibility Tdewpoint} into energy_train_iml;/*NOTE: Should include the same # of variables as
you have from the PCA analysis on TRAIN data*/
train_pca=(test_vector)`*(energy_train_iml)`; /*Multiply (5x25) with (25xN) to get 5xN*/
train_pca_T=(train_pca)`; /*Transpose 5xN to Nx5 N= number of obs in Train data and 5= # of
significant PC's from Test */
```

```
/*print train_pca_T;*/
create Energy_Train_PCA from train_pca_T[colname={"PC1" "PC2" "PC3" "PC4" "PC5" "PC6"
"PC7" "PC8" "PC9" "PC10"}];/* This will create a SAS data set from IML*/
append from train_pca_T;
close Energy_Train_PCA;
RUN;
proc print data=Energy_Train_PCA;run;

/********Running the Model on TEST Data*********/
data energy_test;
set energy_test;
log_Total_Energy_Consumption =log10(Total_Energy_Consumption);run;

data energy_test_subset;/*Creating a new Train dataset to pass to IML*/
  set energy_test(keep= log_Total_Energy_Consumption dayofweek HourofDay);
  run;

data combined;
   merge energy_test_subset Energy_Train_PCA;
run;

proc glm data=combined plots(maxpoints=none)=all;
class dayofweek HourofDay;
        model log_Total_Energy_Consumption= pc1-pc10 dayofweek HourofDay;
        run;
```