

Constructing Baseline of Customer's Hourly Electric Usage in SAS®

Yuqing Xiao, Bob Bolen, Diane Cunningham, Jiaying Xu
Southern Company Services, Atlanta, GA

ABSTRACT

"PowerRewards" is a pilot program offered by the Georgia Power Company in 2008 in an effort to reduce residential customer usages during summer peak hours. As load analysts, we were challenged to construct models for estimating baseline electric usages for a group of residential customers on a series of events defined by the Georgia Power company. This paper briefly describes the background of the program, process flow during the pilot, analytical method used, some SAS® code in the final product and shadow models developed after the pilot.

INTRODUCTION

Residential and commercial electricity use often varies dramatically during different times of the year and different hours of the day. Electric systems and grids are typically scaled to meet projected peak demand. Significant capital investment might be required for new power plant construction in response to increasing demand at peak hours. Demand control technologies and programs are hot topics in today's utility industry. It provides a means of lowering peak hour demand and thus reduces capital cost to utility companies, cuts harmful emissions and lowers electricity prices to consumers.

In 2008, Georgia Power Company, one of the leading utility companies in the southeast, launched a new pilot program called "PowerRewards" to incent residential customers during hot summer peak hours. It gives customers credit when they reduce their electric usages during critical peak events called by the company. In order to determine the rewards, baselines of each customer's hourly electric usage needed to be constructed and rewards were calculated based on the differences between customers actual usage and the projected usage, called the baseline (Figure 1). Several algorithms for constructing the baselines were studied and a two-step estimation methodology was selected for the final model. Over 1,000 customers were enrolled in the 2008 pilot. A total of 6 events were called and each lasted 2 to 4 hours.

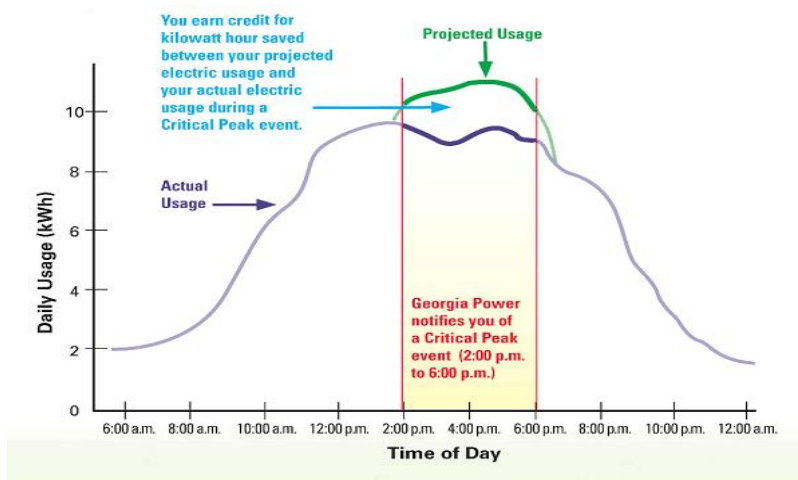


Figure 1. PowerReward concept

PROCESS FLOW

At the beginning of the program, a special team was dedicated to the collection of hourly usage data from enrollment group customers. Data was loaded and processed for gaps, reading errors and other data issues each day. Daily usage data was then appended to a master dataset. On the day the company decided to call an event (event day is generally the next day and when anticipated system load would be high), the master dataset was checked against minimum model building requirements and a potential call list was created. This call list was further checked for customer's eligibility to create the final call list. Customers in the final list were contacted by e-mail and/or phone the same day. During the next day's event hours, customers could choose to lower their electricity consumption by setting the air conditioner thermostat to a higher degree, turning off large, non-essential devices, such as pool pumps, or postponing use of large appliances like clothes dryers, dishwashers, etc. Within 3 business days after the event,

weather data was updated, baselines for each customer were constructed and final rewards were calculated and uploaded on the company's website. Customers could view their actual usage, baseline usage and calculated reward on each event day online. The entire program process flow diagram is shown in Figure 2.

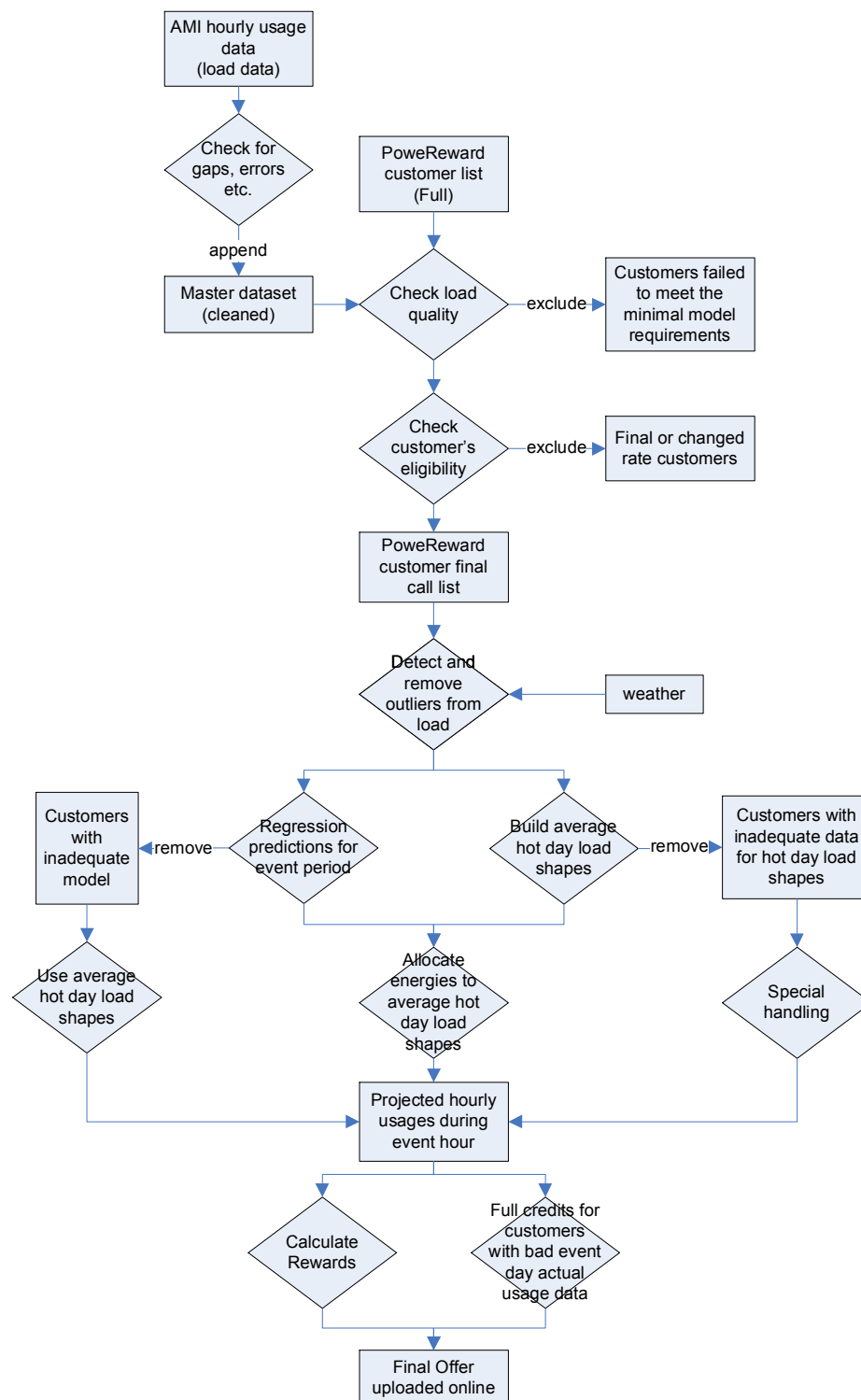


Figure 2. Program Process Flow

STATISTICAL METHOD

Four estimation methodologies for constructing the baselines were studied, including simple average, combined ratio, linear regression and time series. About 20 model specifications were developed and evaluated based on their data requirements, model performance, ease of implementation and ease of interpretation. A two-step estimation methodology was selected for the final production model.

From past experience, residential customer's summer electricity usage is known to change rapidly from hour to hour and is largely driven by air conditioning and customer behavior. Regression on temperature relates customer usage with air conditioning and is a good logical approach. However, **hourly** regression model ignores other effects, such as equipment cycling, customer behavior, etc., and tends to have large errors. Instead, we chose an energy model on the entire event as the first step, which predicts how many kilowatt hours a customer would have used during event hours had an event not been called. Weighted least squares regression model was used to give high temperature days more influence in the model. The full model definition is as follows,

$$NEU = \beta_0 + \beta_1 * AvgT + \beta_2 * AvgDP + \beta_3 * Lag1AvgT + \beta_4 * Lag2AvgT + \varepsilon$$

where,

NEU stands for customer's normal electric usage

AvgT is the average dry bulb temperature over the event period

AvgDP is the average dew point temperature over the event period

Lag1AvgT is the average dry bulb temperature across 26 hours prior to peak period

Lag2AvgT is the average dry bulb temperature across 28 hours prior to Lag1avgT.

The weight was defined as a squared function of AvgT to achieve higher precision on prediction on high temperature days when an event is more likely to be called. Also, stepwise model selection was used at run time with fixed effect on AvgT and others to be selected so that each individual customer can have slightly different reduced model corresponding to his particular usage pattern.

In the next step, a typical hot day load shape was created by averaging the 3 hottest days within the 15 weekdays prior to the event day. The '15 weekdays' restriction attempted to find the best representation of customer's most recent behavior. The predicted event period energy was then allocated to each hour based on the hot day load shape.

$$kWh_i = NEU * AvgkWh_i / \sum AvgkWh$$

where,

kWh_i is the estimated customer usage at event hour i

AvgkWh_i is customer's average hot day usage at hour i.

The final reward was calculated as the sum of positive differences between the estimated baseline and customer's actual usage across event hours.

DATA CLEANING AND MODEL VALIDATION

Despite the strong influence weather has on electricity use, we saw many abnormal usages in our database due to customer's unusual behavior and/or data collection error. To reduce large model errors caused by them, R-student residual was chose to detect and delete outliers. R-student residual is basically the residual if we fit the model with the current observation deleted. Since it removes the impact of the current observation, large R-student residual generally indicates a high influence point and is commonly used for outlier diagnosis. Under the usual regression assumptions, it follows t_{n-p-1} distribution. With SAS, R-student residuals can be calculated and output through 'RSTUDENT' option with PROC REG's 'OUTPUT' statement. To find the cut-off point for R-student residuals, model number of parameters and error degrees of freedom were output through 'OUTEST=' option and 95% confidence bound was calculated using 'TINV' function.

SAS code to get R-student residuals and confidence bounds:

```
options nonotes;
/* stepwise regression and output model statistics */
proc reg data=&cust(where=(dt<&cpp and reg='Y')) noprint outest=est edf;
  by ky_spt;
  model cpp_kwh = avg_t avg_dp tlg1 tlg2 / selection=stepwise include=1;
  weight wt;
  output out=rstudent rstudent=rstud;
run;
options notes;
/* calculate confidence bounds */
data quantile(keep=fmtname ky_spt label rename=(ky_spt=start));
```

```

set est(keep=ky_spt _p_ _edf_);
retain fmtname 'quantile';
n = _p_+_edf_;
label = tinv(1-.025/n,_edf_);
run;
proc format library=work cntlin=quantile; run;
/* set flags for outliers */
data rstudent(keep=ky_spt dt cpp_kwh rstud outlier);
set rstudent;
if abs(rstud) > put(ky_spt,quantile.) then do; cpp_kwh = .; outlier = 1; end;
run;
/* update master dataset with outlier info. */
data &cust;
update &cust rstudent updatemode=nomissingcheck;
by ky_spt dt;
format rstud 10.2;
run;

```

Several statistics were used to address model inadequacy. One is model error degrees of freedom. We drop regression models with small degrees of freedom, which have too few data points for a meaningful model. Another one is PRESS statistic, or Prediction Error Sum of Squares, defined as the sum of the squared PRESS residuals. It is generally considered as a measure of how well a regression model will perform in predicting new data. Large PRESS value indicates high prediction error and was flagged for model removal. Also, negative predictions or predictions lower than the customer's historical minimum can be problematic and were excluded. More model statistics were output for reference, like R^2 , adjusted R^2 , regression model coefficients, etc.

SAS code to get statistics for regression model validation:

```

/* regression model and model statistics */
proc reg data=master(where=(reg='Y')) outest=est rsquare adjrsq press noprint;
by ky_spt;
model cpp_kwh=avg_t avg_dp tlg1 tlg2 / selection=stepwise include=1 alpha=.05;
weight wt;
output out=pred(keep=ky_spt dt est_kwh resd press lb)
p=est_kwh r=resd press=press lcl=lb;
run;
/* compute average residual, press residual and usage, minimum usage and total
weights by temperature group(high, medium or low) */
proc means data=master(where=(dt<&cpp and cpp_kwh ^= .)) noprint;
by ky_spt;
class avg_t;
var resd press cpp_kwh;
format avg_t temp_gp.;
output out=avg_resd(where=( _type_=1) rename=(avg_t=temp_gp))
mean(resd)=avg_resd mean(press)=avg_press mean(cpp_kwh)=avg_kwh2
min(cpp_kwh)=min_kwh2;
output out=weights(where=( _type_=0) drop=avg_t) sum(wt)=tot_wt
mean(cpp_kwh)=avg_kwh min(cpp_kwh)=min_kwh;
run;
/* create event date dataset with model STAT */
data prediction;
merge master(where=(dt=&cpp))
est(in=a rename=(avg_t=beta1 avg_dp=beta2 tlg1=beta3 tlg2=beta4))
weights(drop=_type_ _freq_);
by ky_spt;
if a then do;
msr_press = sqrt(_press_/tot_wt);
ratio = msr_press/avg_kwh;
tot_beta = sum(of beta1-beta4);
end;
run;

```

For each event, 1 summary report and 3 Excel workbooks were created with SAS using the ODS tagset and were reviewed and examined for large average residuals, large differences between the regression prediction and the 3 hottest day average prediction, negative slopes for the regression model, etc. Scatter plots and Load shapes on event days were used to identify issues for the individual customer (Figure 3).

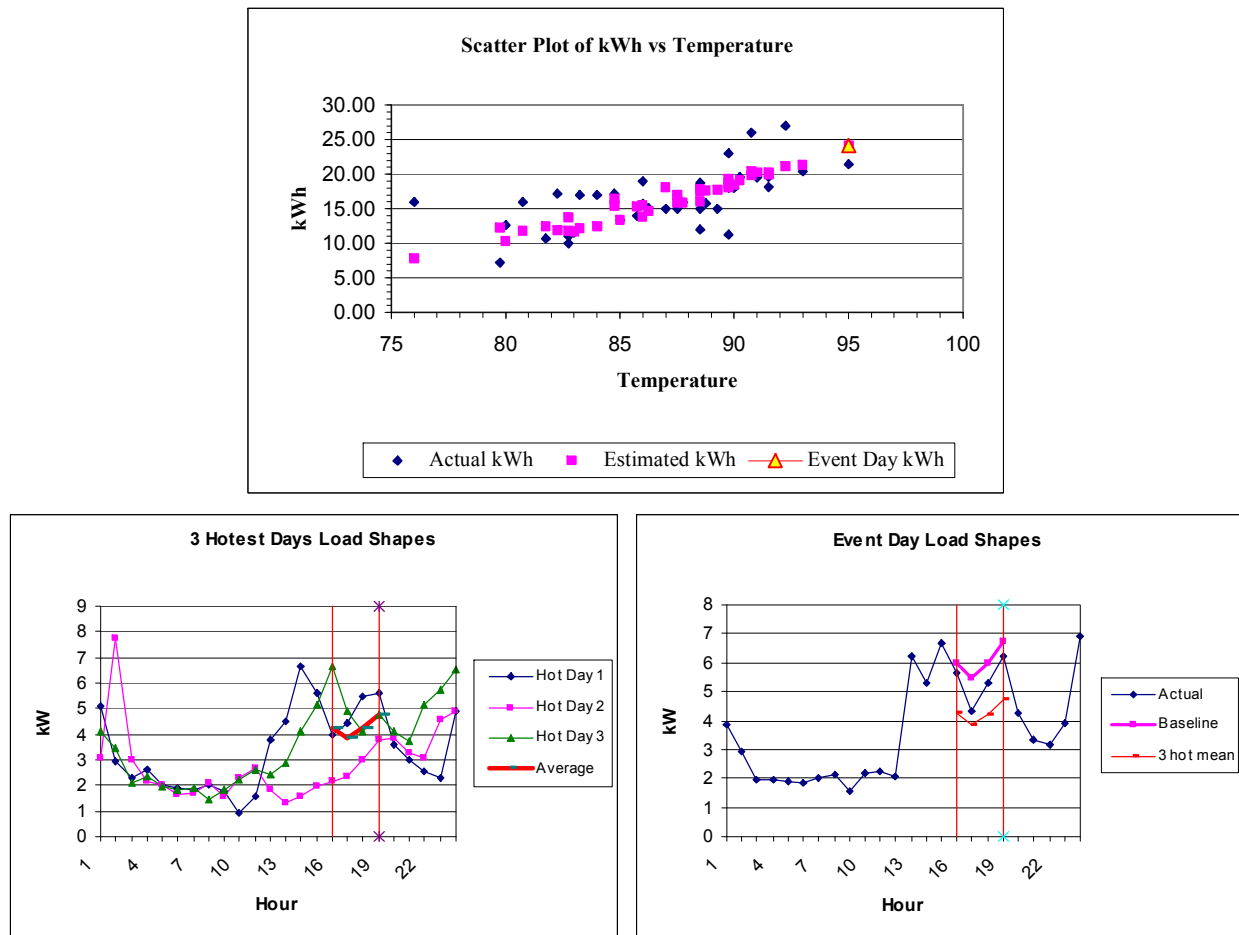


Figure 3. Scatter Plot and Event Load Shapes

CODE FOR CALCULATING LAG VARIABLES

In the full regression model, there were 2 lag variables. They crossed 3 days and depended on the called event hours. A macro function with beginning and ending hour parameters was written to create the lag variables.

SAS code to calculate lag variables:

```
%macro avg_lag(beg, end, var);
  %let n=%eval(&beg-1);
  %let m=%eval(&end-2);
  %let k=%eval(&end-&beg+1);
  (
    %do i=&n %to &m; lag&i(&var)+ %end;
    lag&i(&var))/&k
  %mend;
%let pos1 = %eval(&hr_end-&hr_beg+1);
%let pos2 = %eval(&hr_end-14+1);
%let pos3 = %eval(&pos2+26);
%let pos4 = %eval(&pos3+1);
%let pos5 = %eval(&pos3+28);
data weath1 (keep=avg_t avg_dp tlg1 tlg2 index wt dt stn);
  set weath0;
  avg_t = %avg_lag(1,&pos1,db);
  avg_dp= %avg_lag(1,&pos1,dp);
```

```

tlg1 = %avg_lag(&pos2,&pos3,db);
tlg2 = %avg_lag(&pos4,&pos5,db);
if &start <= dt <= &cpp and hr = &hr_end then do;
    index = .5*avg_t+.2*avg_dp+.2*tlg1+.1*tlg2;
    avg_t = max(avg_t,76);
    avg_dp= max(avg_dp,58);
    tlg1 = max(tlg1,73);
    tlg2 = max(tlg2,73);
    wt = (avg_t-75)**2;
    if weekday(dt) ^in (1,7) and dt ^in (&exclude) then output;
end;
run;

```

SPECIAL CASE HANDLING

Although there were procedures to validate the customer's hourly usage data before the customer was informed of an event, there were still several cases where special handling was required. The most common one was an inadequate regression model. In this case, the average of the 3 hottest days was used as the final baseline. A few times, customer's actual load on event day was missing. In this instance, a fixed rebate amount was issued. Occasionally, there were missing hot day load shapes. Instead of borrowing load shapes from another customer, a manual review was made and each situation was accessed.

SHADOW MODEL

At the start of the pilot, it was felt that a long term strategy for the baseline model needed to be easily understood and implemented by the company's billing group. Because of the time constraint, we were unable to spend a lot of time studying and comparing different models during the pilot. Additional models (shadow models) were developed and tested after the pilot. The shadow models focused on various factors related to customer's energy use. New variables were introduced into the regression model, such as different lag variables, power functions of temperature, heat index (a function of dry bulb temperature and relative humidity), shoulder time energy, temperature with time distance adjustment, variation on weights, etc. An average most alike day method was tested against the average hottest day method. Model performance was evaluated and compared using Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE), where small MAPE and MAE were desired. The final statistical results showed that applying shoulder time energy and temperature with time distance adjustment helps reduce model errors.

CONCLUSION

The PoweRewards pilot program has been an excellent learning experience. Many different methodologies were researched and various models were developed and tested. The final shadow regression model out performed any other models and showed better statistics in errors. But implementation of the model requires comprehensive coding and may not be feasible in the company's billing system. Average most alike day takes into account temperature impact and customer behavior change and performed very closely to the regression model. It's ease of interpretation and implementation in the company's billing system might make it a better candidate for the future.

ACKNOWLEDGMENTS

We especially like to acknowledge David Glyer and Bruce Chapman of Christensen Associates, Dr. Hsu of Georgia State University and members of the Southern Company load research team for their time, expertise and contribution to the program.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Yuqing Xiao
 Southern Company Services
 241 Ralph McGill Blvd. NE., Bin 10206
 Suwanee, GA 30308
 Work Phone: 404-506-4619
 E-mail: yxiao@southernco.com

Bob Bolen
 Southern Company Services
 241 Ralph McGill Blvd. NE., Bin 10206
 Suwanee, GA 30308
 Work Phone: 404-506-3730
 E-mail: bebolen@southernco.com

Diane Cunningham
Southern Company Services
241 Ralph McGill Blvd. NE., Bin 10206
Suwanee, GA 30308
Work Phone: 404-506-3220
E-mail: dmcunnin@southernco.com

Jiaying Xu
Southern Company Services
241 Ralph McGill Blvd. NE., Bin 10206
Suwanee, GA 30308
Work Phone: 404-506-4887
E-mail: jxu@southernco.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies.