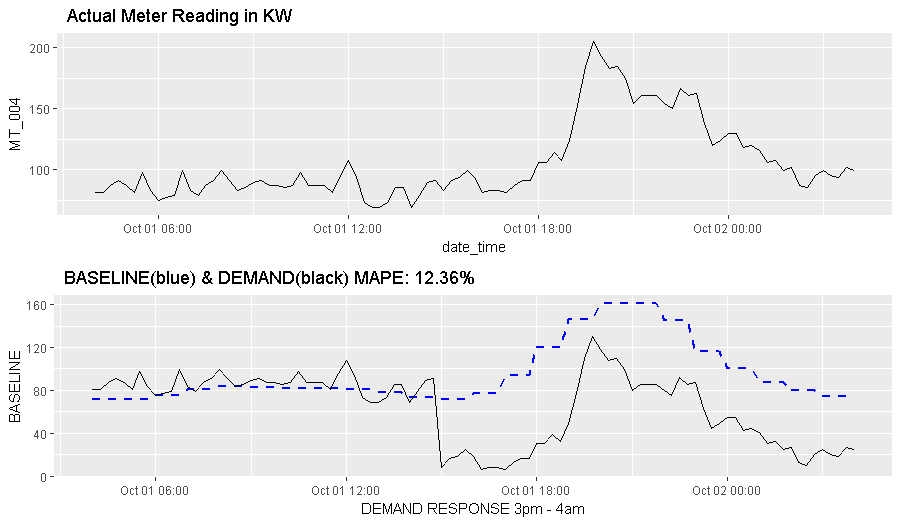
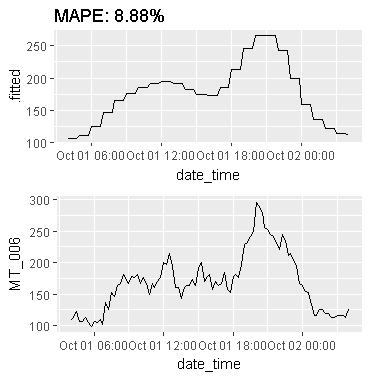
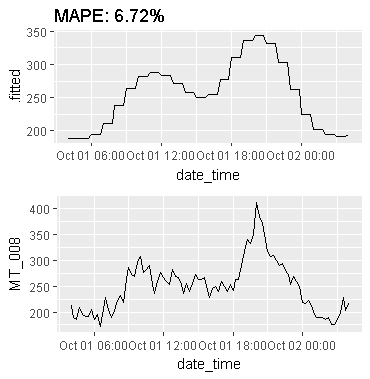
# *Local Weighted Regression for Baseline Calculation*

Local Regression is the most common method used to smoothen a volatile time series. It is a non-parametric methods where least squares regression is performed in localized subsets, which makes it a suitable candidate for smoothing any numerical vector.

* The model is considered a Baseline type 1 method. According to DOE, FERC, and AEIC this methodology is currently the most prominent in Demand Response programs.
* The model takes key aspects from a variety of other commonly applied statistical methods within the Baseline type 1 class including:
  + Rolling Average/ Average
  + Comparable Day/ Day Matching
  + Regression
* Local Regression is flexible enough to include additional variables as they become available
* Robust enough to give consistently accurate estimates with limited data
* The model has 4 basic components (hyper-parameters)
  + - Weight Function: More recent data has higher impact
    - Degree/Family: Allows for flexibility in the estimates
    - Span: % of Data considered when making each estimate
    - Distribution of Errors: Normal Distribution or Symmetric
* The model works by iteratively taking small subsets of data and then fitting a line to each of these subsets (these line need not be straight). The aggregation of these individual lines give us the output or expected value at each point in time for the series.
* **Baseline Specifications Particular to proposed Model**
  + Predictors: Hour, Day, Month, & Interaction between Month and Hour
  + Baseline model components specifications:
    - Weight Function: (1-(dist/maxdist)3)3
      * weights decrease steadily from the closest observation (only observations within the span are weighted)
    - Degree: 2
      * Allows many small local regression fits to bend accordingly and better fit to the observed capacity
    - Span: 3 months maximum (well within AEIC 1 year maximum)
      * At most, only consider observations from most recent season (3-months).
    - Distribution: Symmetric similar to t-distribution
      * Errors are approximately normally distributed but do have some what fatter tails than the Gaussian distribution.
  + **Description of Process:** The Model attempts to minimize the Error between the estimated load at time t and the actual load at time t. For any given day we choose to estimate baseline, by design the model considers the marginal effect of that particular day, month, hour, as well as the interaction between the day and month. In addition it weights the most recent data more heavily than older data, and indirectly reduces error caused by autocorrelation. All parameters have been tested on 4 (moderate to high load) customers and resulted in highly statistically significant results for each variable. The model design are based on pg. 61 *Measurement and Verification for Demand Response,* commissioned by the DOE and FERC.
* Preliminary Results:





* Average mean absolute percentage error using 2 day rolling average for the 3 meters above was 20.88%.
* Average mean absolute percentage error using local regression for the 3 meters above was 9.32%

NOTE: GAM models may prove to be more accurate particularly for high load customers but they required a vast number of parameters and therefore may not be suitable method for calculating baseline currently. However, this method may yield superior results as we collect more data.

**References:**

<http://www.stat.ucla.edu/~cocteau/stat204/readings/cleveland.pdf>

<https://www.ferc.gov/industries/electric/indus-act/demand-response/dr-potential/napdr-mv.pdf> (pg. 60 – 62)

<https://library.cee1.org/sites/default/files/library/10774/CEE_EvalDRBaseline_2011.pdf>

<http://r-statistics.co/Loess-Regression-With-R.html>