Questions and Comments Following PJM Response to Model Comments

PJM's comments from their response, "20190806-pjm-comments-for-pseg-feedback-post-meeting.pdf ", are contained in the text boxes with my comments and questions below.

Change #1 – End-Use Characteristics

> In the interest of being concise, we may have used the term "equipment index" too loosely. In the proposed model, Residential and Commercial models are built as a function of Heating, Cooling, and Other. Heating, Cooling, and Other are in turn a function of weather (in the case of Heating and Cooling), the respective equipment index, and economic factors. Each sector takes into consideration the identity:
>
> Total Use = Heating Use + Cooling Use + Other Use
>
> Since Heating Use, Cooling Use, and Other Use are not observed but estimated, these are fit into a statistical model (separately for Residential and Commercial):
>
> Total Use = b1 x (Heating Use) + b2 x (Cooling Use) + b3 x (Other Use)

I see no documentation of these model estimations in Manual 19. If this is an entirely new method that PJM is using than the methodology can't be evaluated without knowing what data is used to estimate the coefficients and what the regression results are. If these results were provided and I just missed them (unfortunately, something not out of the realm of possibility), please direct me to them. If they weren't provided, please provide.

> In the End-Use data, Commercial intensity figures are supplied in terms Use per Square Foot. So Working Age Population is a proxy for Square Footage to help the model translate Intensity to Total Energy. Though not on our slide, we also couple Working Age Population with a Square Foot to Population conversion based on National data to take into account that Square Footage is increasing relative to population.

I understand that commercial square footage data is difficult to obtain but my question remains – Why are you using Working Age Population? Granted, employment is not a good proxy for commercial floor space given automation. However, if economic theory would suggest that the demand for retail and other commercial services is a function of income, household, and population, I don't understand why Working Age Population was chosen. With working age population, defined as ages 15-64, this metric omits age cohorts that have a demand for Education & Health Services which is one of the fastest growing sectors in the economy. Again, there appears to be additional modeling (the working age to square footage translation) for which the documentation has somehow eluded me. If these results were provided and I just missed them, please direct me to them. If they weren't provided, please provide.

> History is ***Electricity Retail Sales to the Industrial Sector*** from EIA divided by ***Real Goods Output*** from Moody's Analytics. Forecast is then driven off of ***Electricity Use per dollar of Real Shipments*** from the EIA's Annual Energy Outlook.
>
> As for the abrupt change, we don't know specifics, it would be a combination of their thoughts on industrial mix, disposition to using electricity, and industrial process efficiency.

The industrial energy use in the AEO "includes energy for combined heat and power plants that have a non-regulatory status and small on-site generation systems." Is it possible that your history is based on Retail Sales of electricity and the forecast is based on Total Electricity Use (purchased and other) so that the series aren't consistent?

Change #2 – Modeling Non-weather sensitive Load

> We agree that we are not doing two-stage least squares.
>
> What we are trying to accomplish is to rectify the growing disconnect between non-weather sensitive load trends and the variable set that the current model uses to represent that trend (i.e. the current model's Other Equipment Index combined with the Economic Index). The solution we propose is to estimate non-weather sensitive load, forecast it using a calibrated non-weather sensitive variable (a combination of Residential, Commercial, and Industrial trends), and then directly use this as an input to the model of total load. We believe that employing this method will allow the model to better distinguish historical trends and ultimately provide a more accurate forecast.

Well, actually the proposed PJM model is utilizing two-stage least squares – Stage 1 estimating the non-weather sensitive load and Stage 2 using that result in the forecast model. My point was that 2SLS is not being used for the standard econometric reasons that it was developed to address. If the current model is not capturing non-weather sensitive load trends, then it implies that there is a specification problem or problems in the model or the estimation data set that should be examined and rectified. It seems very extreme to use an untested, and to me questionable, methodology that uses data that is not in the season of the peak load - the main objective to forecast. In addition, if the peak hours in the shoulder months are used to estimate the base load, then the hour used is also not appropriate.

> We are working on changes that consider your concern regarding the absence of Summer/Winter months in the non-weather sensitive load calculation.
>
> In establishing the history of non-weather sensitive load, we don't necessarily care about the why, only about the trend. The why is then considered when the non-weather sensitive variable is estimated and forecasted in a regression model which uses the calibrated non-weather sensitive variable (a combination of Residential, Commercial, and Industrial trends).

I don't think that the history of non-weather sensitive load can be accurately established without considering why it changes over time. Annual binaries, as I stated before, capture all the omitted variables, including some imprecision in the weather variables. If you have a driver that you're confident using to forecast non-weather sensitive load, it should be utilized in the estimation of non-weather sensitive load. If it can't be used in the estimation, it shouldn't be used in the forecasting.

Again, there appears to be additional estimations for which the documentation has somehow eluded me. If these results were provided and I just missed them, please direct me to them. If they weren't provided, please provide.

Change #3 – Building weather Variables

> We are using both CDD and THI because they are both found to add value to the model.

I assume, from the statement above, that you're using CDD because the models "performance", e.g. R^2, MSE, etc., is reduced if you include CDD. This, however, is not a justification for using a variable it is an introduction of bias in your estimates. The rationale for this bias is that, with the regression model, you're trying to quantify the conditional expectation, i.e:

$$E(y|X\beta) \qquad [1]$$

of the theoretically valid model that you've constructed. If variables are selected to, for example, minimize MSE then a different conditional expectation is being quantified, i.e:

$$E(y|X\beta,\text{minimize}(MSE)) \qquad [2]$$

The bias arises because you get a different answer with [2] than from [1] because, in effect, they are two different questions.

Now, you know as well as I do that we're all guilty of doing this. As a matter of fact software exists to do this on a large scale (see: Stepwise Regression). My problem with it arises when it is used to include variables that should not be in the regression. As far as I know, there is no engineering or behavioral reason for cooling load to be a function of temperature without taking humidity into account. I think that using CDD in the models is fundamentally wrong.

> It has not been verified that the max daily THI never occurs after the peak, nor minimum daily WWP. Weather conditions that occur after the peak are highly collinear with conditions that occur before the peak. Weather variables are meant to be indicative of the weather conditions that customers face, and the model establishes the relationship between load and weather. If the weather conditions are not informative to determining the relationship, then the model will give less weight to them.

In the analysis of peak load, the independent variable that is most accurately measured is the weather. As a result, the weather data that is used should reflect the weather that contributes to the peak load whether it be at time of peak or prior to the peak. There is no reason to use weather that is "highly collinear with conditions that occur before the peak". The weather variables are meant to be the measure of weather that impacts load. There's no reason to settle for anything less – especially if alternative measure occurs after the peak. Again, using estimation results to justify using an incorrect specification is not appropriate.

> The idea behind running the univariate regression models was to get each individual variable's explanatory power. Since the Summer variables and Winter variables being examined are highly collinear, using a multivariate regression model would defeat this purpose. While there is undoubtedly omitted variable bias, this is predominately an issue for the parameter estimates in each of the univariate models (which never get used).

The omitted variable bias not only affects the parameter estimates but, since the parameter estimates are used to calculate the fitted values, the omitted variable bias also impacts the Sum of Squared Errors which is used. Multicollinearity does not preclude using multivariate regression. Multivariate regression in the presence of multicollinearity will result in unbiased estimated coefficients – the desired result. The estimates, however, are not as precise as one would hope. Using less precise unbiased estimators is preferable to using biased estimates that are just wrong.

> The goal here is not to save on degrees of freedom. The origin of pursuing this was that the current weather specification is a mix of linear and spline terms. Summer weather is specified by CDD (linear term), lag CDD (linear term), and THI (spline terms). Much of the weight in the final model solution is being given to the linear term, and thus the current model is perhaps not fully capturing the saturation effect at extreme weather conditions and thus potentially contributing to overforecasting. We believe having a single weather variable will allow us to have a more cohesive and consistent treatment of weather.

Perhaps using more appropriate weather, THI rather than CDD, would be a better first step in attempting to fully capturing the weather effect.

> You are correct that the variable with the greatest weight is CDD. As stated on the prior page, this is true in the current model as well. By itself CDD tends to be more predictive of daily peak load than does THI. We tried iterations with daily average THI in place of CDD and those tended to not do as well. We agree that THI is more intuitive, and if building an hourly model rather than a daily model, THI may very well work out to be the superior variable. We have hypothesized that it may have to do with morning and early afternoon humidity not being all that influential in final daily peak determination. This type of humidity (if it burns off by peak time) would boost a daily average THI out of proportion with its load impact, whereas CDD by contrast would not. Moreover, the culmination of a daily peak has to do with the aggregate thermal buildup throughout the day rather than a single point in time like max daily THI.

The statement that the THI doesn't work because the average daily THI or the daily maximum THI doesn't work is the result of using a daily average and a daily maximum that, as I noted before, contains weather data that occurs after the peak and, as a result, is irrelevant. It avoids the question as to why the THI at the time of peak and during the appropriate periods prior to peak is not used.

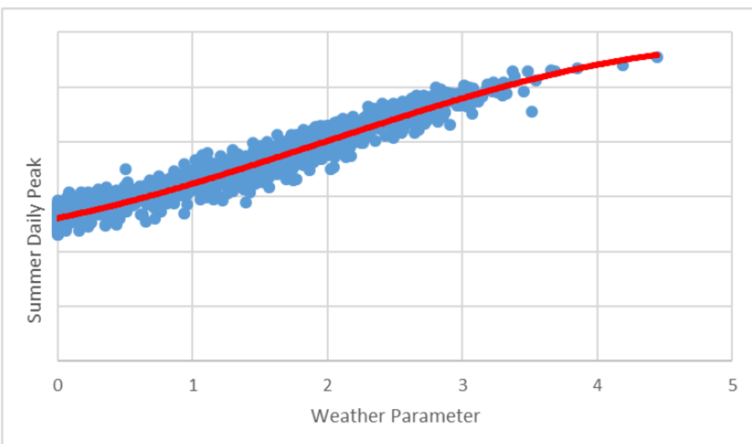
It is also noted in the PJM response that "the culmination of a daily peak has to do with the aggregate thermal buildup throughout the day". If this is the case, why are variables quantifying this buildup, e.g. one-hour lag, three-hour lag, etc., not included in the model specifications.

> A one day lag for CDD has been found to add value to the model. This would be highly correlated with overnight weather.

Again, “add value to the model” is not a sufficient reason to use a variable. My question was “What is the evidence supporting using a one day lag for CDD”. I would like to know the thought process behind the choice. The fact that it would be correlated with overnight weather is not a reason to use the one day lag rather than the overnight weather, if the overnight weather is the appropriate variable, since it is available.

> We don’t think it’s appropriate to read too much into over- or under-stating based on that graph. The graph is purely a presentation of multiple years of data points versus the weather parameter, thus obscures some of the other factors the model would be taking into account (i.e. year-to-year trends and calendar effects). As for the selection of the cubic polynomial, it was chosen based on visual inspection of the data. Here are two examples:

I was not reading too much into the shape of the graph. It was explicitly presented by PJM as the justification for the use of the cubic polynomial.



Referencing the graph above, how is this taking into account year-to-year trends and calendar effects?