Regulated Price Plan Roadmap Pilot Program Interim Impact Evaluation: Summer 2018

Appendix H: Alternative Control Group Approach

2019-05-24

Prepared for:



*Prepared by:*

Peter Steele-Mosey, Associate Director

Navigant Consulting, Ltd.

First Canadian Place

100 King St. W. Suite 4950

P.O. Box 64

Toronto, ON, M5X 1B1

416.777.2440

navigant.com

# iNTRODUCTION

The London Hydro RPP Pilot was designed to be evaluated as a randomized control trial (RCT), using a “recruit-and-deny” enrollment strategy to ensure a robust control group for all treatments. During the enrollment period, London Hydro became concerned about its ability to obtain sufficient program applicants to support this strategy and asked Navigant to develop a contingency quasi-experimental evaluation approach, should insufficient RCT controls be available to support the evaluation.

To this end, Navigant developed, and described in the evaluation plan, a quasi-experimental approach for developing matched participant controls based on hourly, weather-dependent consumption patterns. Navigant had previously used such an approach for the evaluation of Alectra’s Advantage Power Pricing pilot.[[1]](#footnote-2)

The approach to developing this control group is described in Section 1, below.

At the end of the enrollment period, London Hydro had a sufficient number of applicants to the CPP (and CPP/RT) program streams to deliver a robust RCT control group for those treatments, but concerns arose regarding whether these customers would still be suitable as controls for the RT-only group, and the decision was made to proceed with the initial steps of matched control group development, and to use the pre-matching period as a testing ground to compare the two control groups.

The purpose of the testing conducted was to answer the question: *are the consumption patterns of the RT-only participants in the testing period more like those of the RCT controls’, or more like those of the matched controls’?*

Put another way, are applicants to the CPP treatment more like applicants to the RT treatment than non-applicants whose electricity consumption patterns are similar to RT treatment applicants in the period from February 1, 2017 through January 31, 2018?

What Navigant found was that the matches were poorer controls for RT-only participants than the RCT controls, and that the RCT controls, were, overall, extremely similar in consumption patterns to the RT-only participants in the test period. Based on this finding Navigant determined the most prudent course of action was to proceed with using the RCT controls as controls for all treatments, rather than to work to further refine the matching algorithm.

This appendix is included for completeness and transparency. It outlines the initial approach to matching deployed by Navigant and the findings of test-period comparison.

Navigant would emphasize that readers must bear in mind that the matched controls were abandoned as a strategy when it became apparent that the RCT control group would be a suitable control group for the RT-only participants. An RCT control group is always preferable to a quasi-experimental approach (e.g., matched controls) as it controls for selection bias – an unobservable trait that only experimental (i.e., RCT) approaches can control for.

Had no RCT control group been available, or had it been demonstrated that the RCT group in hand was meaningfully different from the RT-only participant group in the testing period, Navigant would have continued to refine its matching approach to address issues that its testing identified.

The remainder of this Appendix is divided into two sections:

* **Control Group Development.** This section provides a description of how the matched control group was developed.
* **Control Group Assessment.** This section summarizes Navigant’s analysis of the different control groups in the testing period, and provides the evidence for why the RCT controls were a suitable control group for the RT-only participants.

A note on terminology:

* **Program Period:** The period from May 1, 2018 through April 30, 2019
* **Matching Period:** Referred to in the body of this report as the “pre-program period”, this is the period from February 1, 2017 though January 31, 2018.
* **Testing Period:** The period from February 1, 2016 through January 31, 2017

# Control Group Development

Navigant’s approach to selecting for each participant an eligible non-participant with the most similar hourly electricity consumption patterns has previously been documented in its evaluation of Alectra’s Advantage Power Pricing pilot program. Navigant’s approach for this project is similar, but with some material differences. The entire approach is described in detail below.

To select the control group PowerStream provided Navigant with hourly AMI data for participants and for over 130,000 eligible non-participants extending from 2015-10-01 to 2018-02-05.

Matching took place in two distinct steps:

1. **Grouping.** A “pre-matching” step that limits the match search to those non-participants with a total consumption in the matching period (2017-02-01 through 2018-01-31) that is reasonably close to that of the given participant.
2. **Matching.** Participants are matched to controls that fall with the specified matching groups, according to hourly electricity consumption patterns.

## Grouping

Matching with high frequency data is computationally intensive and can be quite time-consuming. The “Grouping” step reduces the resource demands of matching by limiting a participant’s potential matches only to those non-participants with a reasonably similar total energy consumption within the matching period (2017-02-01 through 2018-01-31).

This initial step proceeds in the following manner:

1. **Eliminate Outliers and Participants with Incomplete Data Series.** Non-participants were eliminated from the pool to be used to generate matches when:
   1. More than 20% of expected observations missing in either the matching period or the testing period.
   2. They were missing more than one complete month of data in either the matching period or the testing period.
   3. Annual consumption exceeded 60 MWh in the matching period.
2. **Non-Participants Grouped.** After excluding non-participants per the criteria above, non-participants are assigned to one of 20 groups of equal size based on their estimated total consumption in the period.[[2]](#footnote-3)

As part of this grouping, the maximum and minimum annual kWh of all non-participants within each group is preserved

1. **Assign Groups to Participants for Matching.**
   1. Each participant is assigned a “range” of annual kWh consumption. This range extends from 25% below their annual consumption to 25% above their annual consumption.
   2. This range (for each participant) is then compared to the threshold (minimum and maximum) values of each of the non-participant groups.
   3. This then delivers a set of non-participant groups for each participant – for example:
      1. If the upper threshold (maximum annual kWh) of a given non-participant group is more than the bottom kWh value of the participant’s range, that group is included in the set of groups from which a control customer may be selected for the given participant.
      2. Likewise, if the lower threshold (minimum annual kWh) of a given non-participant group is less than the top kWh value of the participant’s range, that group is also included in the set of groups from which a control customer may be selected for the given participant.

The thresholds used and number of potential control customers in each group is summarized in below.

Figure H - 1: Distribution of Consumption Values by Non-Participant Group

|  |  |  |  |
| --- | --- | --- | --- |
| **Group Number** | **Max Annual Consumption** | **Min Annual Consumption** | **Number of Customers in Group** |
| 1 | 2,486 | 0 | 4,838 |
| 2 | 3,316 | 2,486 | 4,839 |
| 3 | 3,929 | 3,316 | 4,839 |
| 4 | 4,437 | 3,929 | 4,839 |
| 5 | 4,887 | 4,437 | 4,839 |
| 6 | 5,307 | 4,887 | 4,839 |
| 7 | 5,720 | 5,307 | 4,839 |
| 8 | 6,128 | 5,720 | 4,839 |
| 9 | 6,541 | 6,129 | 4,839 |
| 10 | 6,968 | 6,541 | 4,839 |
| 11 | 7,417 | 6,969 | 4,839 |
| 12 | 7,902 | 7,418 | 4,839 |
| 13 | 8,445 | 7,902 | 4,839 |
| 14 | 9,028 | 8,445 | 4,839 |
| 15 | 9,725 | 9,028 | 4,839 |
| 16 | 10,586 | 9,725 | 4,839 |
| 17 | 11,681 | 10,586 | 4,839 |
| 18 | 13,201 | 11,682 | 4,839 |
| 19 | 15,863 | 13,202 | 4,839 |
| 20 | 58,314 | 15,864 | 4,839 |

On average, each participant was compared to non-participants in approximately six of the non-participant groups. Put another way, each participant was, on average, compared to nearly 30,000 different non-participants.

## Matching

For matching controls to participants, Navigant established a set of 30 day-types, and compared participant and non-participant (within the groups specified above) average hourly loads on those 30 days from the period beginning 2017-02-01 through to 2018-01-31. Each participant’s consumption pattern in this period was compared with the consumption patterns of all available non-participants from the relevant groups (see above) in the same period. So, for example if a given participant’s set of non-participant groups included groups 3, 4, and 5, then only the non-participants in those groups would be considered as matches for the given participant.

The non-participant whose historical consumption patterns by hour of day and day-type deviated the least from those of the given participant became that participant’s match. Historical consumption patterns for each individual and each season were summarized by a 720 element vector of average levels of consumption (30 day-types times 24 hours – see below).

The matching period was divided into three two-month periods. In the winter: January/February, March/December, and April/November. In the summer, May/October, June/September, and July/August. Five day-types were applied to each period. Four of these day-types were assigned to non-holiday weekdays where the average daily temperature met a certain temperature threshold (see for the criteria), and the fifth was applied to all holidays and weekends.

Figure H - 2: Day-Type Average Daily Temperature (Celsius) Thresholds

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Day Type | | | | |
|  |  | Weekdays | | | | Weekends/ Holidays |
|  |  | 1 | 2 | 3 | 4 | 5 |
| Winter | **Period A:** Jan & Feb | <= -11 | -7 to  -11 | -3 to  -7 | > -3 |  |
| **Period B:** Mar & Dec | <= -10 | -6 to  -10 | 1 to  -6 | > 1 | All Days |
| **Period C:** Apr & Nov | <= 1 | 2 to  1 | 6 to  2 | > 6 |  |
| Summer | **Period D:** May & Oct | >= 18 | 15 to  18 | 11 to  15 | < 11 |  |
| **Period E:** Jun & Sept | >= 23 | 21 to  23 | 16 to  21 | < 16 | All Days |
| **Period F:** Jul & Aug | >= 23 | 21 to  23 | 20 to  21 | < 20 |  |

*Source: Environment Canada weather data and Navigant analysis*

So, for example, if the average temperature on a non-holiday February weekday was -11 C or below, that day was assigned day-type A\_1.

The weather data thresholds were selected based on the observed weather in matching period (2017-02-01 through 2018-01-31) such that approximately 10% of the non-holiday weekdays were day-type 1 (the most extreme temperature), 20% were day type 2, 30% were day type 3 and the remaining 40% were day type 4. Effectively, the day-type thresholds were selected in such a way as to increase the weighting of more extreme temperature days, improving the likelihood that matched controls temperature-sensitive loads will be very close to those of the participants.

This approach was taken based on previous experience with this technique in which it was found that a more balanced weighting delivers slightly better matches on “average” temperature days but considerably worse matches on more extreme temperature days. The intention of the trade-off applied by Navigant (i.e., to over-weight extreme days) is to deliver a data set of considerably greater flexibility, with matches on “average” or unseasonably mild days that are still quite good, but with much better-quality matches on extreme temperature days. Based on Navigant’s observations in the testing period (see below), Navigant has since concluded that, if anything, there was insufficient weight placed on the most extreme temperature days in each season.

These day types thresholds were assigned to the matching period customer AMI data, and these interval data were then averaged by day-type and hour, delivering the aforementioned 720 element vector of values. The non-participant whose vector of values had the shortest Euclidean distance from the participant’s was selected as that participant’s match.

# Control Group Assessment

The best way to test a control group for bias is to compare it to the participant group in some pre-program period using the same approach that will be used to evaluate the program effects. In this case, Navigant estimated a slightly simplified[[3]](#footnote-4) (but identical in terms of delivering ex-post impacts) version of the regression equation used to estimate program impacts using the testing period data (i.e., consumption from May 1 through October 31 of 2016).

The parameter estimates associated with the treatment variables provide an indication of how similar the control and participant groups are. These parameters should be as close as possible to zero, and definitely not be statistically significant for the control group to be considered as a reasonable control group for the analysis. An estimate of a parameter of interest (i.e., that captures a program effect) that is statistically significant is an indication that any program period estimates could be biased. This finding indicates that there exists a consistent difference between the consumption patterns of the participants and controls which is not being adequately controlled for and is therefore being attributed to (in the testing period) a non-existent “program”.

Navigant performed this test using both control groups (RCT controls and matched controls) and compared the p-values and parameter estimates of the parameters of interest, see Figure H - 1, below. Recall that a p-value is a measure of estimate uncertainty – the higher the p-value, the more uncertain the result. A p-value of 0.1 or less indicates that a parameter estimate is statistically significant at the 90% confidence level. A p-value of 0.5, on the other hand, would indicate that a parameter is only statistically significant at the 50% confidence level.

Figure H - 3: Parameter Estimates and P-Values

|  |  |  |  |
| --- | --- | --- | --- |
| **TOU Period** | **Value Type** | **RCT Counterfactual** | **Matched Control Counter Factual** |
| On-Peak | Parameter | 0.001 | -0.076 |
| P-value | 0.992 | 0.227 |
| Mid-Peak | Parameter | -0.020 | -0.049 |
| P-value | 0.809 | 0.364 |
| Off-Peak | Parameter | 0.089 | -0.069 |
| P-value | 0.575 | 0.517 |
| Weekend Off-Peak | Parameter | 0.002 | -0.103 |
| P-value | 0.995 | 0.630 |

There are two note-worthy facts about these outputs:

* **Parameters estimated using RCT Controls are smaller in absolute value.** For every period except for the weekday Off-Peak, the parameters delivered by the data-set that used the RCT controls (not the matched controls) are much smaller in absolute value. In the one instance in which the absolute value of the parameter associated with the RCT control data set is larger, the difference between the absolute value of the two parameters is smallest of all four parameters.
* **P-values associated with parameters using RCT controls are much bigger.** P-values are substantially higher for the parameters drawn from the RCT control data set estimation, indicating much weaker evidence against the null hypothesis (that the effect is non-significant) than for the matched controls-based data set.

The finding above was the main driver for Navigant concluding that it was more appropriate to proceed with the RCT controls than the matched controls. Navigant also noted that visual comparisons of load profiles in the test period suggested that

As part of Navigant’s due diligence, Navigant undertook some additional investigation to determine whether these effects were indicative of any consistent underlying issue with the matched controls, or simply a chance variation in consumption.

A visual comparison of the difference between RCT control and participant loads, and the difference between matched control and participant loads on the hottest days of the year in the matching period and the testing period were quite revealing. In both cases, the control customers had lower demands on average than the participants in the peak parts of the day.

However, when the two time periods were compared, it became apparent that the difference between RCT and participant was much more consistent across time. Given this consistency, these differences are controlled for via the pre-period consumption value included on the right-hand side of the regression equation.

In contrast, the difference between the matched controls and the participants *grew* in the testing period (hence the positive coefficients shown above in Figure H - 3). This was suggestive of the fact that although the matched controls had very similar patterns in the matching period, these patterns deviated in a meaningful way, correlated with weather, outside of that period. Put another way: the matching algorithm did not sufficiently weight consumption at high temperatures.

A major reason for this was simply that the matching period was a very mild summer, compared to the test summer: in 2017 (the test summer) there were only three F\_1 day-types (non-holiday weekdays in July or August where average daily temperature exceeded 23 degrees Celsius). In contrast in the test period (summer 2016) there were 17 instances of this day-type, and in the program period (summer 2018) there were 11 instances of this day-type.

This difference in seasonal temperatures is also evident when examining the trend of daily temperature values across the three summers as shown in below. Note how much lower the fitted curve is for the matching period summer (blue dotted line) than for the testing period (green dotted line) or program period (purple dotted line) summers.

Figure H - 4: Average Daily Weekday Temperatures, 2016, 2017, and 2018

Navigant’s conclusion from this analysis is that had the RCT controls in the end been found to be unsuitable controls for the RT-only group, Navigant would have needed to make further adjustments to its matching algorithm to put a higher weight on the highest temperature period, to minimize match “drift” in hotter summers.

For completeness, and at the request of the OEB, Navigant did estimate the RT-only impacts when the match controls were used. Navigant found that using the uncorrected matched controls the model delivered a small, but statistically significant increase in consumption in the Mid-Peak, Off-Peak and weekend Off-Peak periods. Impacts in the On-Peak period were statistically insignificant. Navigant would emphasize that these estimates are the result of using a control group under development as a contingency, a control group that was abandoned before it was finalized as soon as it became apparent that this contingency was no longer required.

1. Navigant, prepared for Alectra Utilities, *Advantage Power Pricing Pilot: Impact and Process Evaluation: Winter 2015/2016 and Summer 2016*, July 2017

   <https://www.powerstream.ca/attachments/Navigant_Evaluation_Advantage_Power_Pricing_2015-2016.pdf> [↑](#footnote-ref-2)
2. Since non-participants missing some observations may be included, estimation of some values is required in order to obtain a consistent measure of annual consumption in the matching period. This delivered by multiplying each non-participant’s average hourly consumption in the period by the number of hours in the matching period. [↑](#footnote-ref-3)
3. Instead of including temperature interactions, only dummy variables are used to capture program effects. [↑](#footnote-ref-4)