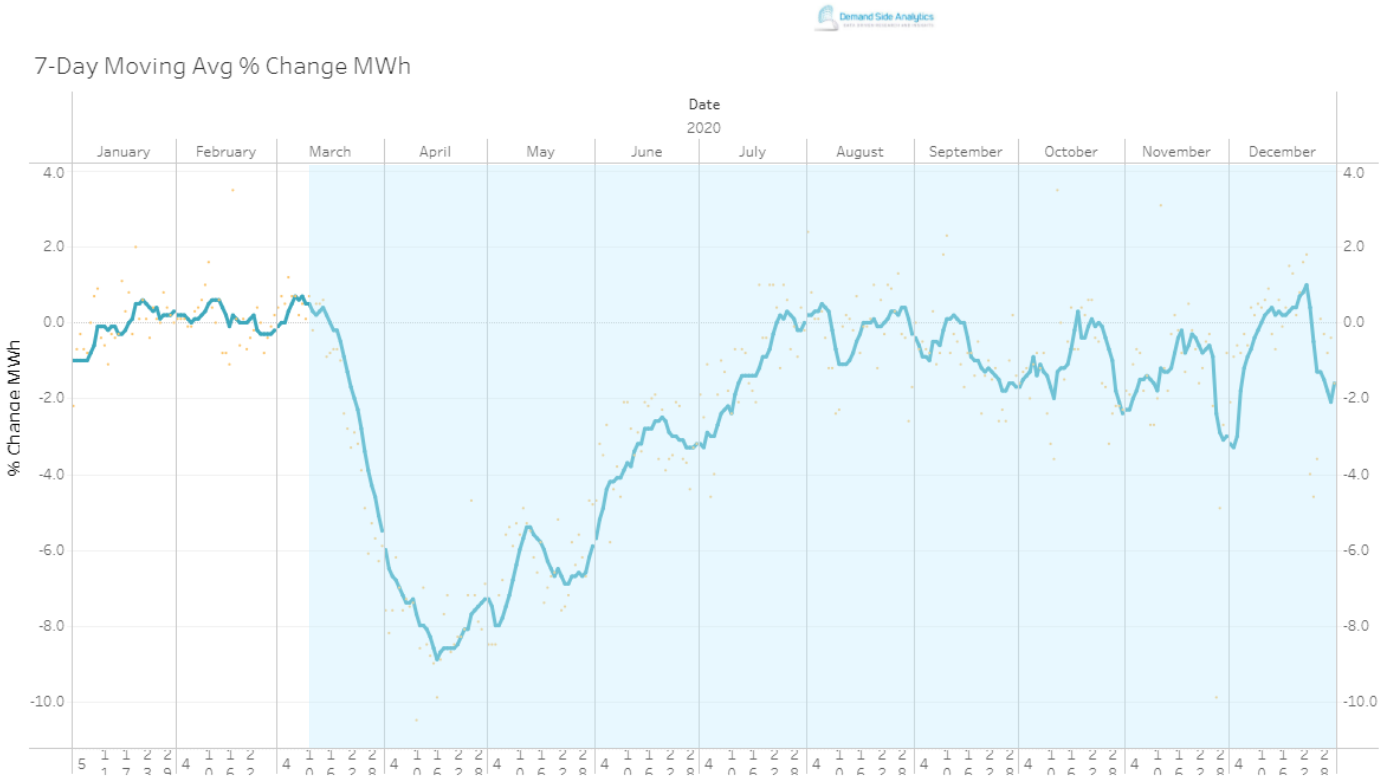
ASSESSING THE ACCURACY OF ENERGY EFFICIENCY EVALUATION METHODS IN A PANDEMIC

Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

APRIL 14, 2021

NON-ROUTINE EVENTS CAN INFLUENCE ENERGY USE

- COVID change occupancy and commute patterns
- The impact varies by industry, customer segment, and county
- The effect of COVID on energy use is dynamic, it evolves based on re-opening guideline and risk
- Other non-routine events also matter
  - Increased solar and battery storage penetration
  - Electric vehicle adoption
  - Default TOU rates

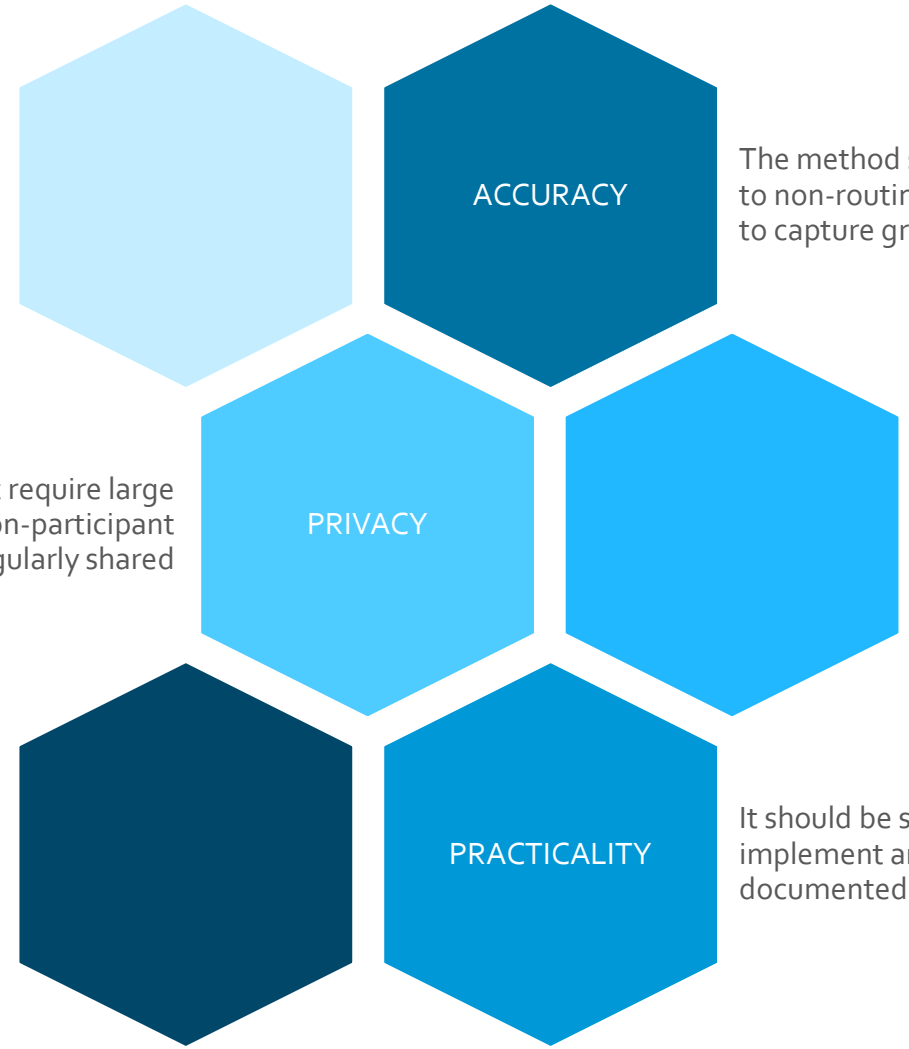


What would the savings estimate be if EE measure were installed February 28th, 2020?

Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

CONTROL GROUPS CAN HELP, BUT HOW TO IMPLEMENT?

- Study goals:
  - Conduct an accuracy assessment
  - Provide PG&E with estimates of program performance
  - Write & document open source code to generate the winning recommendation method's impacts



The method should be robust to non-routine events and able to capture granular impacts

It should not require large amounts of non-participant data to be regularly shared

It should be straightforward to implement and well documented

Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

ALL METHODS SHOULD BE EVALUATED ON THE SAME SET OF CRITERIA

Tournament framework allows us to empirically assess performance on a wide variety of methods	The final recommendation requires balancing multiple outcome metrics
▪ Define the models to be assessed ▪ Iteratively sample **pseudo** participants from the target population ▪ Aggregate across sample sizes and iterations ▪ Calculate impacts == Calculate error	▪ Quantitative ➢ How accurate is the model? ➢ Does the framework work for all customer classes and fuel types? ▪ Qualitative ➢ Straightforward to generate? ➢ Privacy concerns? ➢ Transparent?

STUDY APPROACH

ACCURACY TESTING FRAMEWORK

1. Identify placebo treatment and controls
- Historical participants (2017-2018) without an EE intervention in 2019-2020 (placebo treatment)
- Control pool (random sample)

2. Define methods for accuracy assessment
- CalTRACK v2 – pre-post
- CalTrack v2 + stratified matching control group
- Aggregated matched control group
- Synthetic controls - Use of non-participant aggregated profiles by segment as right-hand side variables
- Alternative pre-post models with matched control

3. For each method, assess performance
- Produce metrics for bias and fit

4. Pick the method based on accuracy and precision metrics and practical considerations

Repeat Process 200X for each sample size and method to produce distribution of errors

A. Sample participants

B. Estimate baseline for treatment period using method

Impacts are zero! No energy efficiency intervention took place. Can we accurately estimate the counterfactual?

C. Compare baseline to actual energy use for placebo treatment period

Any difference from zero is baseline error since, in fact, there was no energy efficiency intervention

D. Store the results
Used to estimate bias and fit metrics

Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

OPTIONS TO TEST EXPAND THE DECISION SPACE EXPONENTIALLY

Method

- CALTRACK v2 method (pre-post regression)
- CalTrack v2 + matched control group
- Aggregated matched control group
- Synthetic controls - Use of non-participant aggregated profiles by segment as right-hand side variables
- Alternative pre-post models + matched control group

Segmentation Granularity

- Residential
  - Climate zone
  - Climate zone, usage strata
  - Climate zone, usage strata, peak kW
  - Climate zone, usage strata, peak kW, EV/Solar status
  - Climate zone, usage strata, peak kW EV/Solar status, Weather sensitivity
- Small and Medium Business
  - Climate zone
  - Climate zone, rate class/size,
  - Climate zone, rate class/size, usage
  - Climate zone, rate class/size, usage, business type
  - Climate zone, rate class/size, usage, business type, and weather sensitivity

Matching Methods

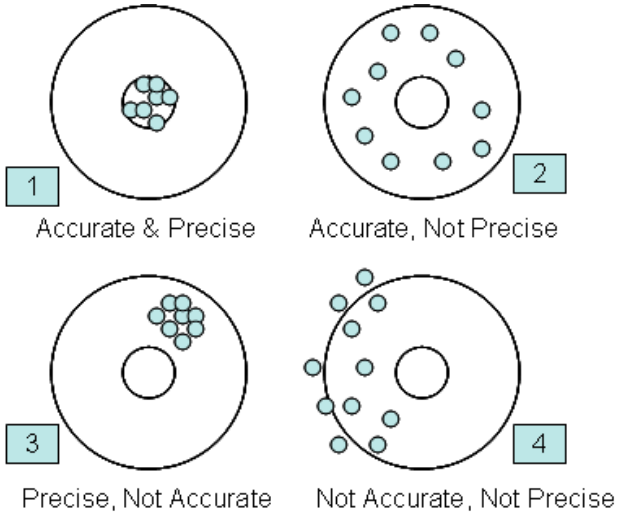
- Stratified Matching
- Stratified plus propensity score matching
- Stratified plus Euclidian distance matching

Matching On

- Customer size bins
- Load shape clusters
- Monthly consumption profiles
- DER Size

ASSESSING ACCURACY AND PRECISION

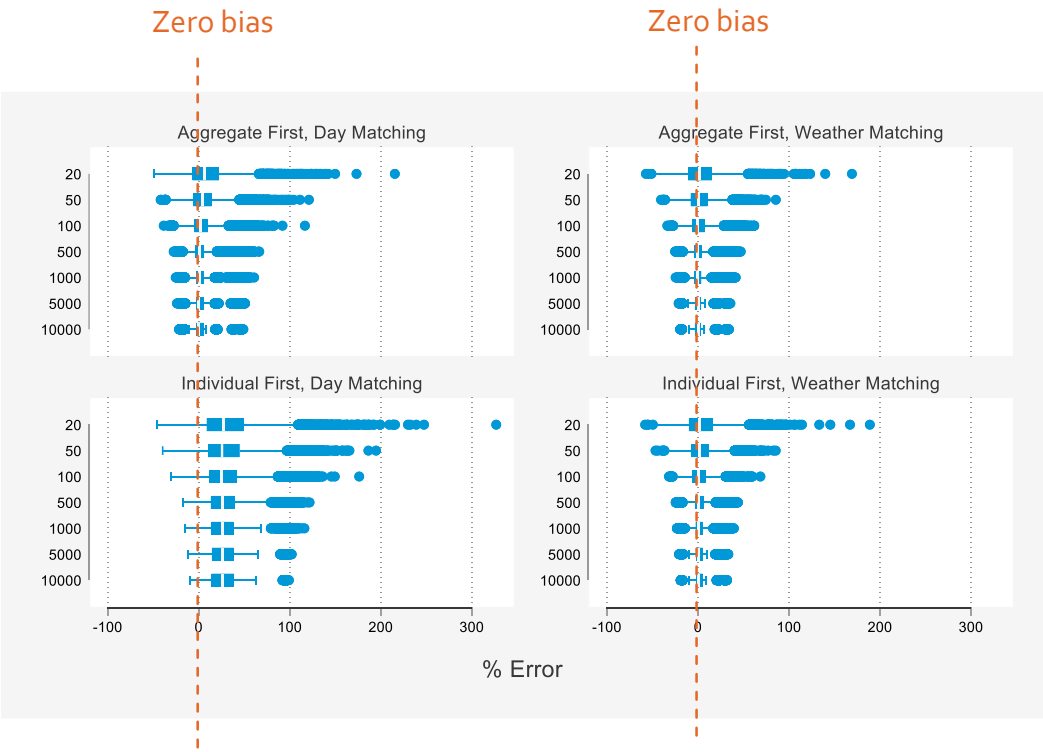
Type of Metric	Metric	Description	Mathematical Expression
Bias	% Bias	Indicates the percentage by which the measurement, on average, over or underestimates the true energy savings.	$\% \, Bias = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)}{\bar{y}}$
Precision	RRMSE	It measures the relative magnitude of errors, weighting more extreme errors more heavily.	$RRMSE = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\bar{y}}$



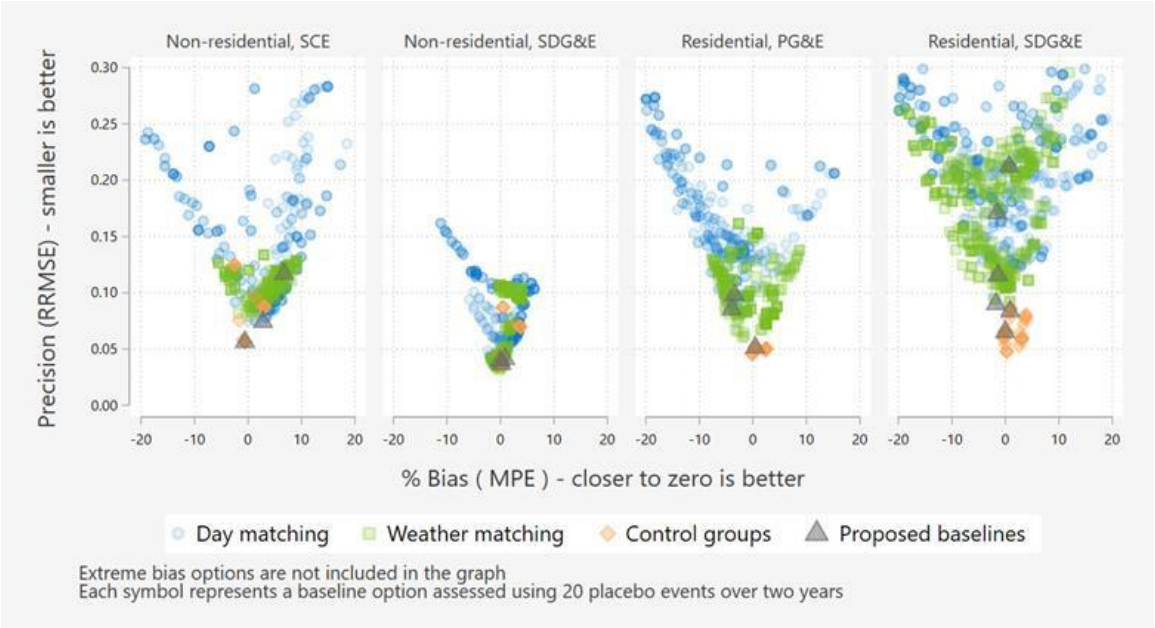
The best methods will be both accurate and precise

VISUALIZING BIAS AND PRECISION

Effect of Sample Size on Error



Balancing Accuracy and Precision

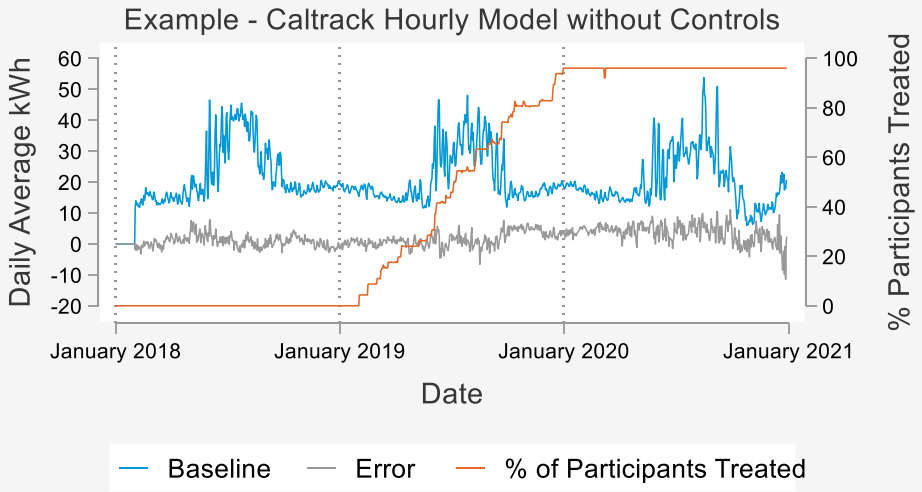


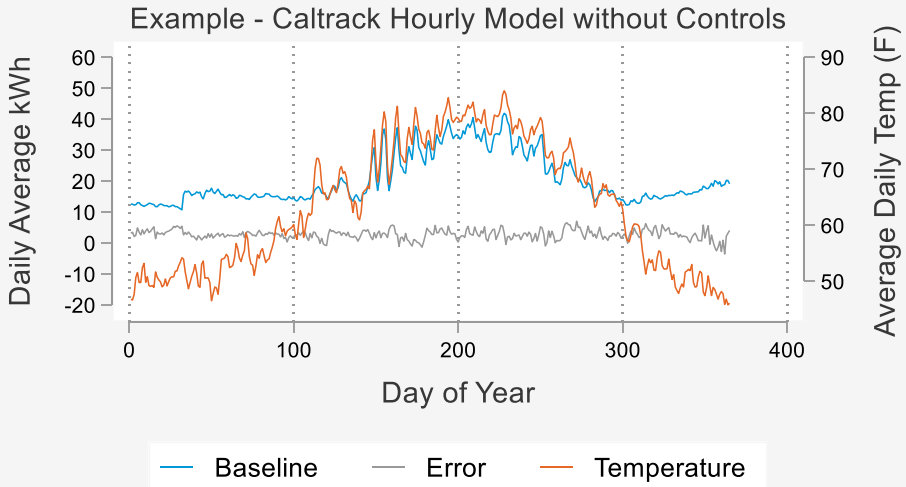
Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

METHOD 1: CALTRACK V2.0

- Building is modeled as base load, heating load, and cooling load.
- Heating load and cooling load are assumed to have a linear relationship
- Find the best balance points
- Run individual model for each location
- Aggregate individual results

Method implicitly assumes that the only difference between the pre and post period is weather and the intervention





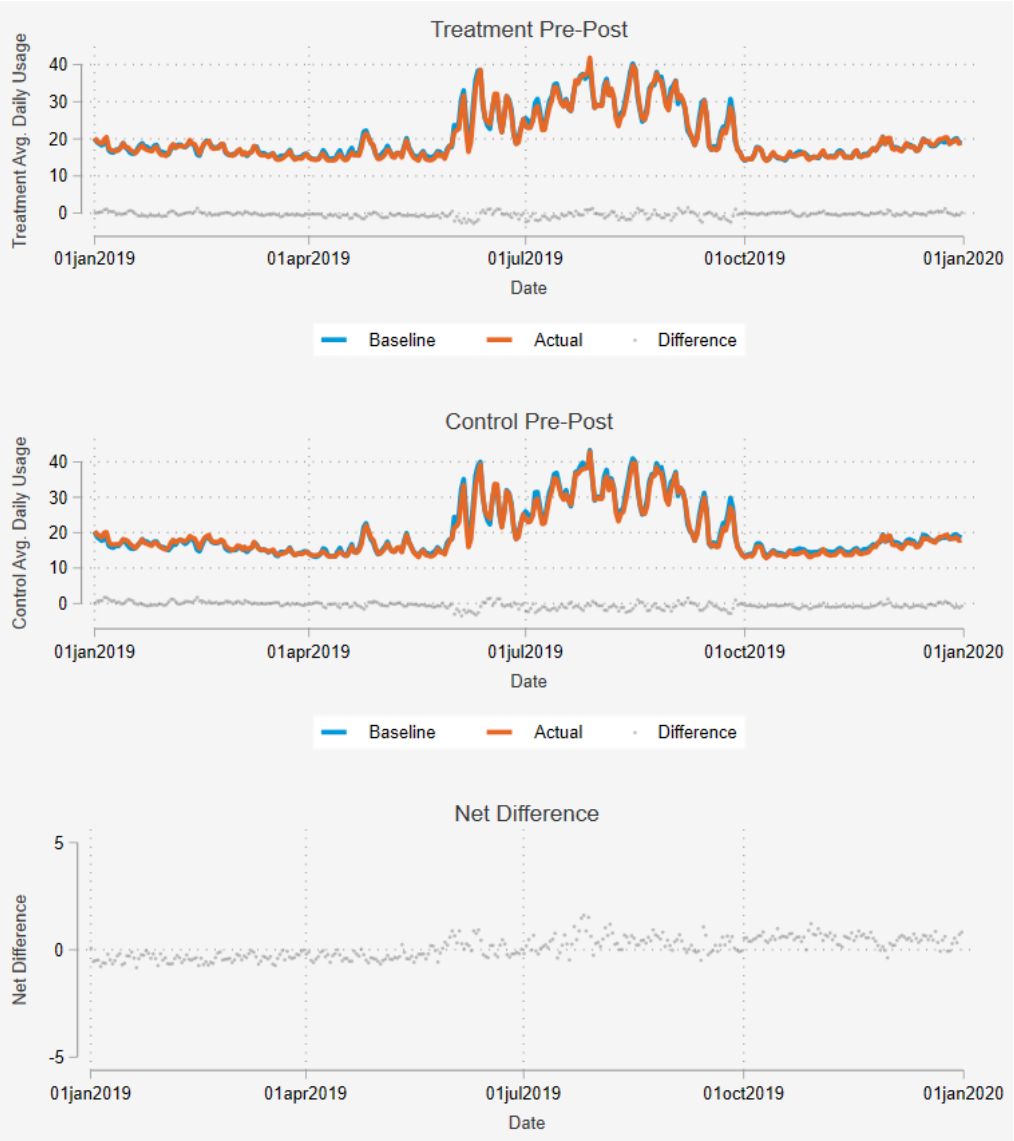
Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

METHOD 2: CALTRACK V2.0 + MATCHED CONTROLS

- Select stratified matched control group
- Run individual regressions on participant data, create counterfactual
- Run individual regressions on comparison group, create counterfactual
- Aggregate counterfactual and observed loads for treatment and control results
- Calculate the difference between counterfactual and observed load for treatment and control
- Take difference-in-differences

If we have a well matched control group:

- Before the intervention, the data should be nearly identical
- After the intervention, a noticeable shift occurs in the group that received the intervention
- The timing of shift coincides with the introduction of the intervention

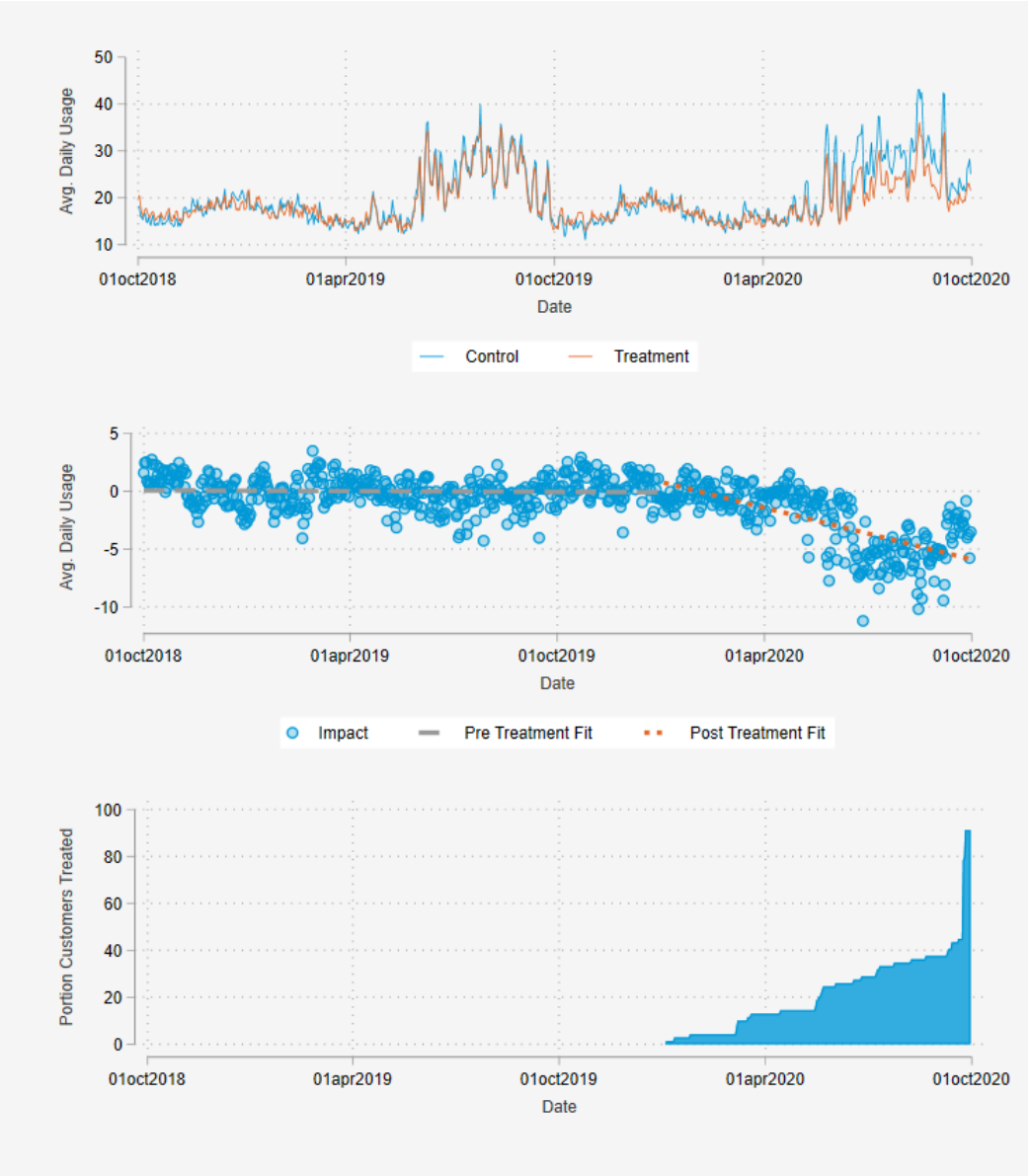


Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

METHOD 3: AGGREGATED MATCHED CONTROL GROUP

- Select stratified matched comparison group
- Aggregate loads for participant and control group
- Take the difference
- Run a regression that quantifies the difference as function of the intervention, weather, and other variables

- Faster computation time and intuitive
- Harder to disaggregate
- May need different cohorts based on intervention date
- Does not resolve concerns about privacy for control groups

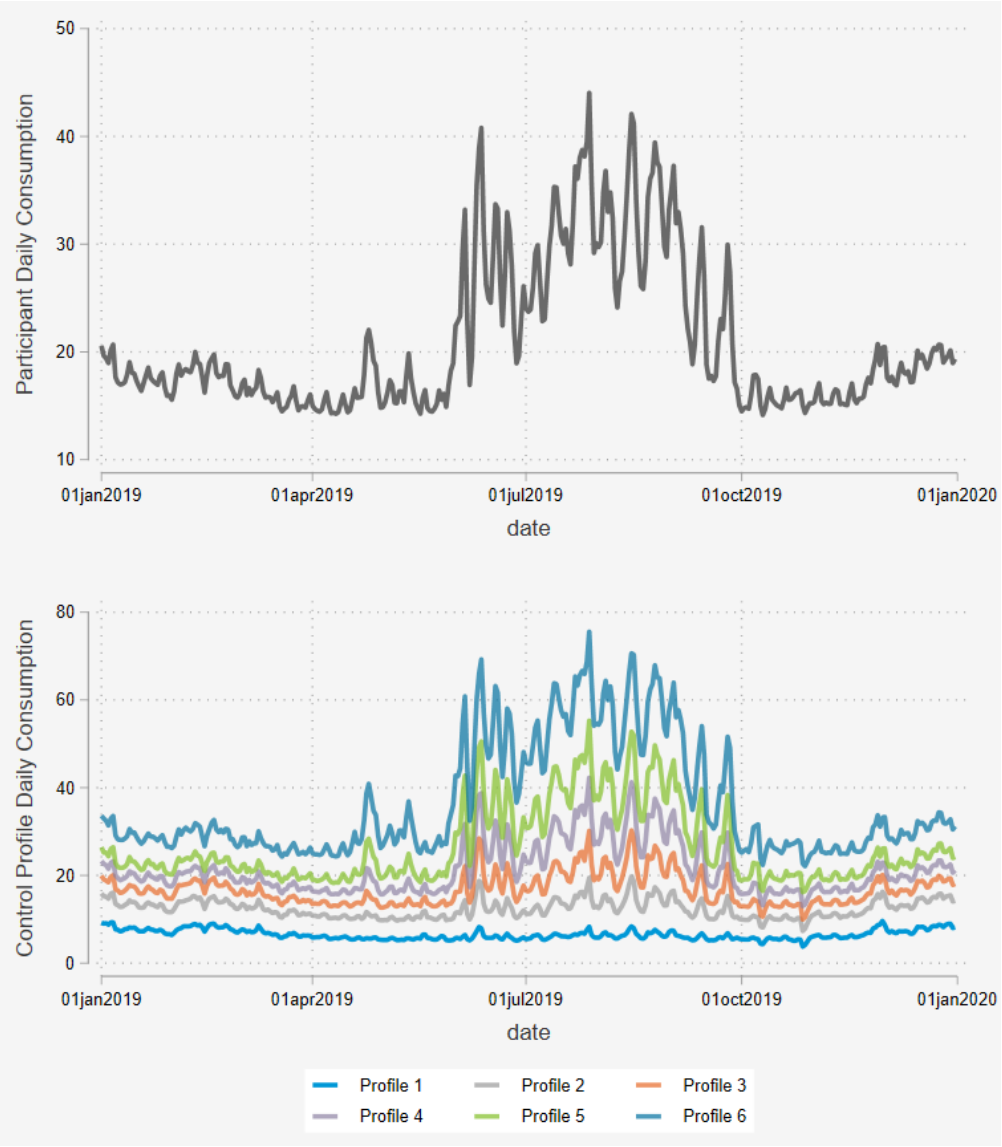


Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

METHOD 4: SYNTHETIC CONTROLS - USE NON-PARTICIPANT AGGREGATED PROFILES BY SEGMENT AS RIGHT-HAND SIDE VARIABLES

- Apply program criteria to comparison (optional)
- Create number of aggregated load profiles for different segments
- Model energy use as a function of weather and behavior of comparison profiles
- Counterfactual takes into account observed behavior by comparison group

- Resolves concerns about customer privacy
- Can be modeled at the individual level or in aggregate

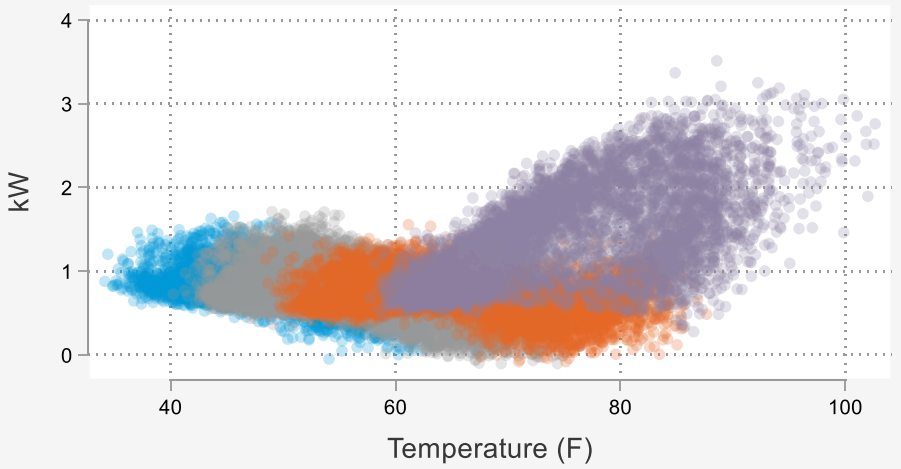


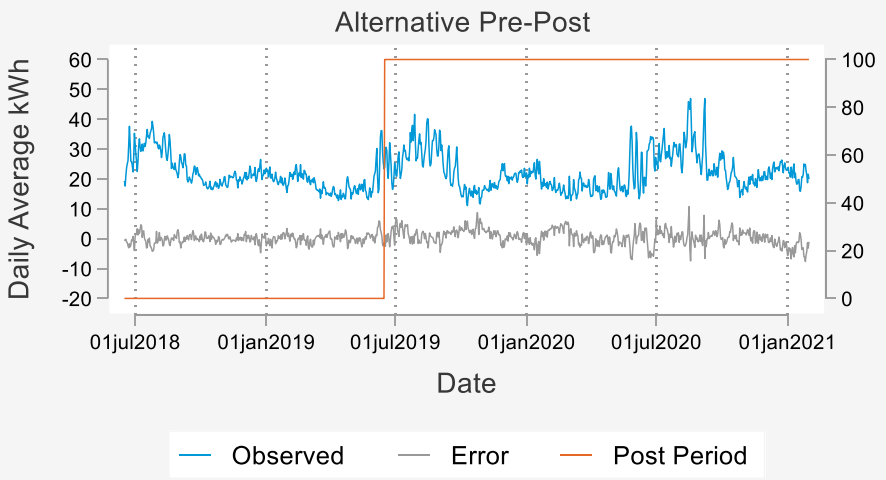
Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

METHOD 5: ALTERNATIVE PRE/POST MODELS WITH CONTROLS

- Select stratified matched control group
- Run individual regressions (non-CalTRACK models) on participant data, create counterfactual
- Run individual regressions (non-CalTRACK models) on comparison group, create counterfactual
- Aggregate counterfactual and observed loads for treatment and control results
- Calculate the difference between counterfactual and observed load for treatment and control
- Take difference-in-differences

- Same benefits/concerns as CalTRACK model with controls
- Change in specification may make the impact estimates more robust



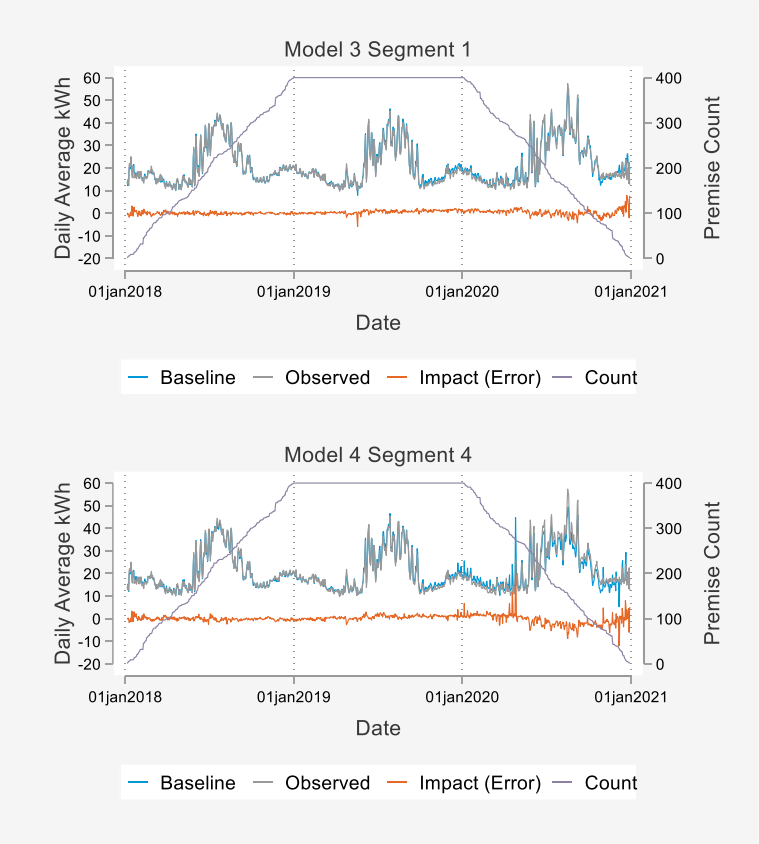


Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

A SUMMARY OF THE CAPABILITIES OF EACH METHOD

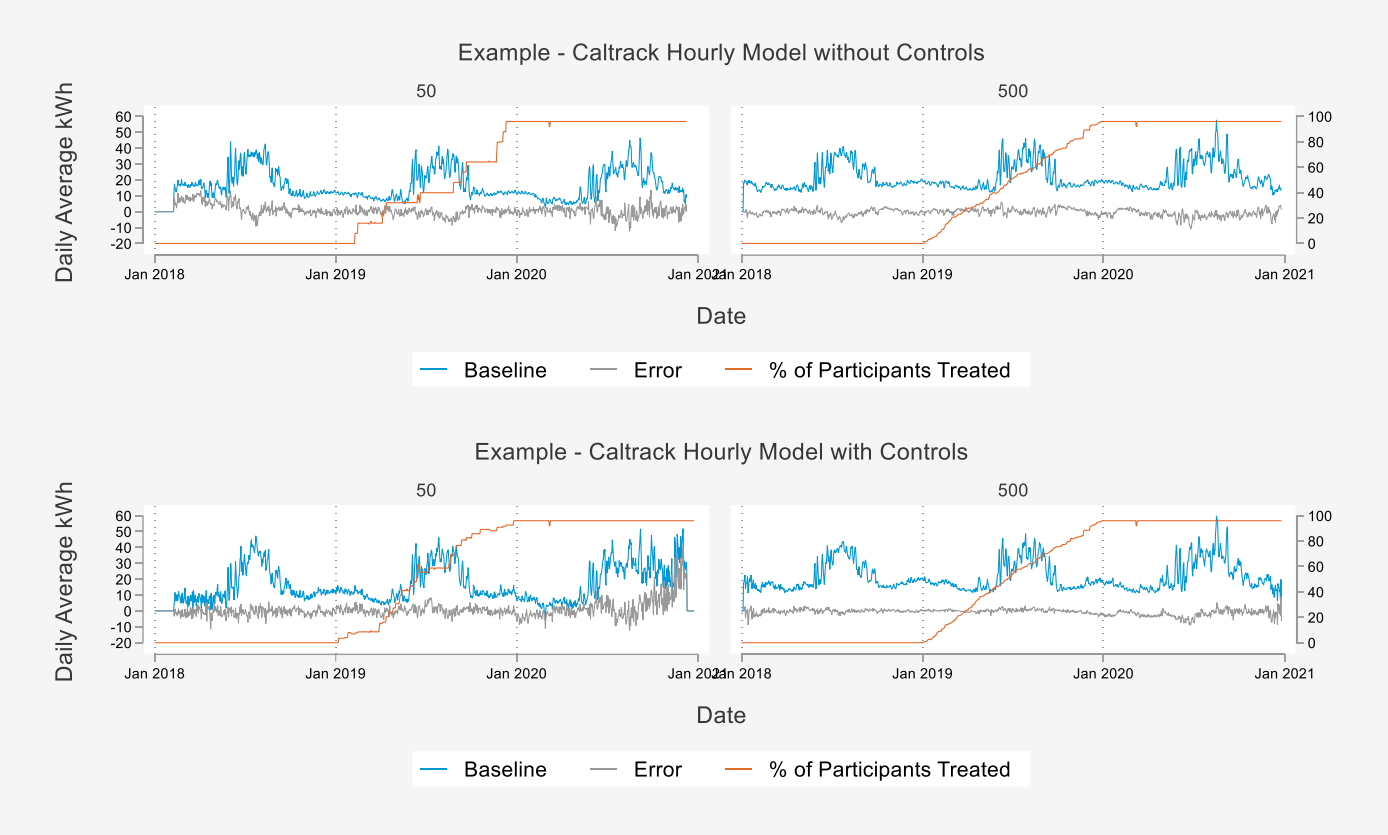
	Does Not Require Individual Non-Participant Data to be Shared	Does Not Require a Specific Matched Control Group to be Computed	Can Produce Granular Per-Customer or Segment Impacts	Ease of Implementation
CalTRACK w/o Controls	✓	✓	✓	✓✓
CalTRACK w/ Controls	✗	✗	✓	✓
Aggregated Matched Control Group	✓	✗	✗	✓
Synthetic Control	✓	✓	✓	✓✓✓
Alternative Pre-Post Models	✗	✗	✓	✓✓

EXAMPLE RESULT VISUALS – EACH GRAPH REPRESENTS 1 ITERATION



Synthetic Controls

Generally more precise impacts, but highly dependent on segmentation strategy



CalTRACK models with and without controls at different levels of aggregation

As expected, aggregation reduces estimation noise. Full analysis will quantify actual precision and accuracy gains

Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

REMAINING ANALYSIS

Accuracy Simulation

- 5 frameworks
- ~150 segmentation strategies
- 3 segmentation methods
- 4 regression specifications

Accuracy Analysis

- Aggregation bootstrapped across 200 iterations and 7 different sample sizes
- Compare results for:
  - Residential and Commercial populations
  - Gas & Electric fuel types

Documentation

- Determine final recommendation
- Write and document open source code
  - Stata
  - R
  - Python

EXAMPLES OF OPEN SOURCE CODE BEING DEVELOPED

Stata

```
version 16.0
program flex_match, rclass

/*
    Purpose: Matches customers based on hard and soft match criteria. Uses
    a variety of matching algorithms

    Assumed structure:
        data long by id(s)
        id(s) | treat | hard match vars | soft match vars

    Required inputs:

        [if]:           optionally specified
        id:             any combination of string and numeric unique identifiers
        treat:          binary treatment variable 0/1.
        method:         Matching algorithm. Options are:
                            random: randomly select x neighbors, no stratification
                            strata: stratified random matching
                            propensity: propensity score matching
                            euclidian: euclidian distance matching
        neighbors:      number of nearest neighbors. Must be positive integer
        seglist:        list of hard match segments (for all but random match)
        charlist:       list of soft match variables (for propensity &
                        euclidian distance matching)

    Outputs:
        The dataset returned has the following outputs:
        id:             the list of ids that were provided originally by user
        treat:          the treatment indicator provided by user
        _mcgroup:       the hard match group that the customer belongs to
        _match:         indicator for whether the customer was matched. A value
                        of 1 indicates the customer was matched
        _matchid:       an indicator for unique mappings of treatment customer
                        to their specific matched control(s). There will be one
                        treatment customer for each _matchid, and between 0 and
                        `neighbor' number of controls (depending on how many are
                        available)
        _method:        the method used for matching

*/
    syntax [if], id(string) treat(string) method(string) neighbors(real) ///
        seglist(string) charlist(string)
```

Python

```
# ==================================
# 2. Run Regression Loop for Each Prem ID by hour in Cohort
# ==================================
output = df_reg_data.copy()
df_loop = pd.DataFrame(columns=['t_prem_id', 'hour', 'baseline', 'date',
                                'weekday'])
# Loop by prem ID:
for group in df_reg_data['t_prem_id'].unique():
    ct = datetime.datetime.now()
    print(f"starting Prem ID: {group} at {ct}")
    # define subgroup
    df_sub = df_reg_data.loc[df_reg_data['t_prem_id'] == group]
    # Regress by Hour:
    for h in range(1,25):
        # Weekday dataset
        wd_reg = df_sub[(df_sub['post'] == 0) & (df_sub['past_year'] == 1) & (
                df_sub['weekday'] == 1) & (df_sub['hour'] == h)]
        # create dummies
        wd_reg = pd.get_dummies(wd_reg, columns = ['day_of_week', 'month', 'dst'])
        wd_reg.dropna(inplace=True)
        # Define X and Y for Weekday Regression
        # Y variable: consumption
        y = wd_reg[['t_kwh']]
        # explanatory variables: # update with full list of columns
        X = wd_reg.drop(columns = ['t_prem_id', 'date', 'hour', 'treatment',
                                   't_kwh', 't_avgtemp', 'cohort', 'ym',
                                   'climate_zone', 'source_int', 'past_year',
                                   'holiday'])
        x_list = X.columns.tolist()
        # Fit OLS model with Robust standard errors
        reg = sm.OLS(y,X).fit(cov_type='HC1')
        # Predict
        pred = df_sub[(df_sub['weekday'] == 1) & (df_sub['hour'] == h)]
        pred = pd.get_dummies(pred, columns = ['day_of_week', 'month', 'dst'])
        pred.dropna(inplace = True)
        pred_X = pred.drop(columns = ['t_prem_id', 'date', 'hour',
                                      'treatment', 't_kwh', 't_avgtemp',
                                      'cohort', 'ym', 'climate_zone',
                                      'source_int', 'past_year', 'holiday'])
        pred_X = pred_X[x_list]
        pred['baseline'] = reg.predict(pred_X)
        pred = pred[['t_prem_id', 'hour', 'baseline', 'weekday', 'date']]
        df_loop = df_loop.append(pred)
```

Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

REAL WORLD APPLICATION: PG&E EE PROGRAM EVALUATIONS

PG&E EE PROGRAM OVERVIEW

P4P Program

- Residential program implemented by multiple vendors
- Interventions include:
  - Measures with "deep savings" e.g insulation replacement, AC, pool pump replacement, etc.
  - "Light touch" measures – e.g. self-installed EE measures
  - Behavioral measures
- Implementation started in 2017 and total program enrollment has continued to increase over time

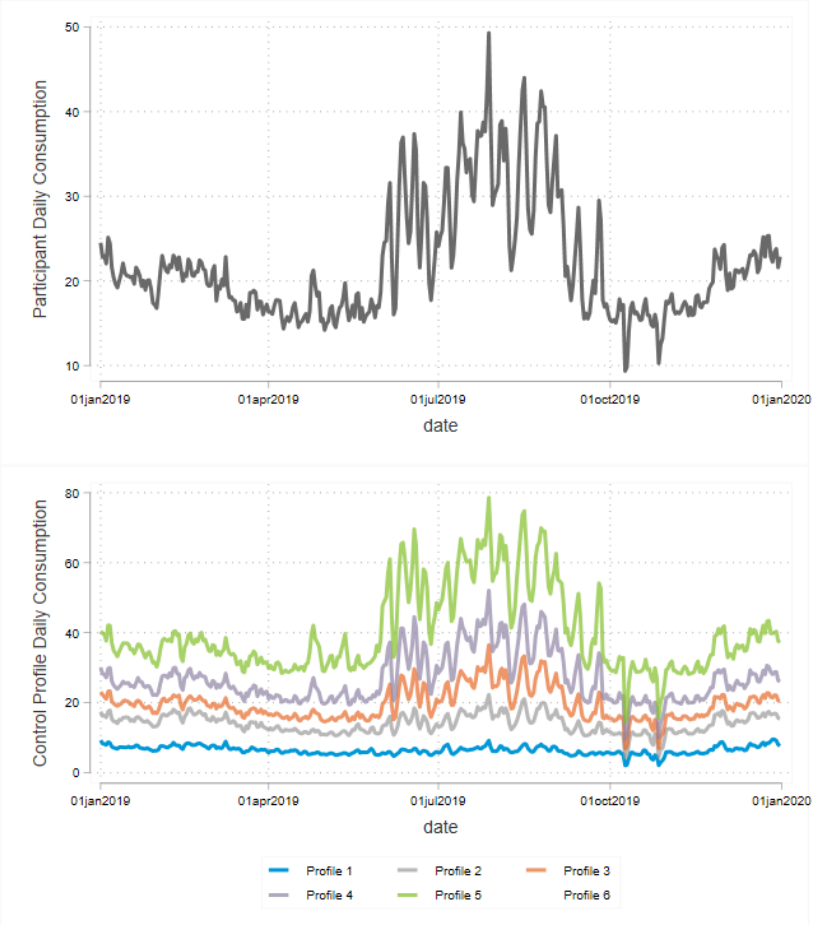
OBF Program

- Non-residential program implemented by multiple vendors
- Interventions include:
  - Lighting measures
  - HVAC measures
  - Custom measures
- Implementation started in 2017 and total program enrollment has continued to increase over time

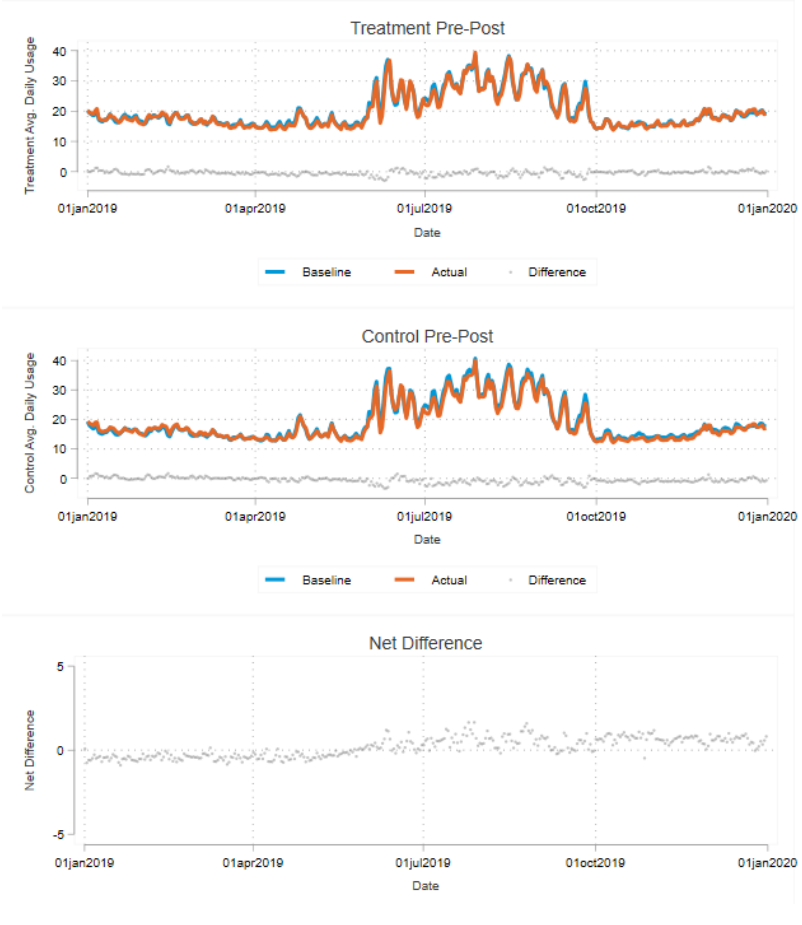
METHODS TESTED SO FAR

- Method
  - Individual regressions
  - Synthetic
- Matching
  - Stratified matching
  - Stratified + propensity score matching
- Eligible control candidates
  - Past EE participants
  - Non-Participants

Synthetic Control Profiles



Matched Control Group



Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

P4P MODEL ACCURACY IN OUT OF SAMPLE TESTING

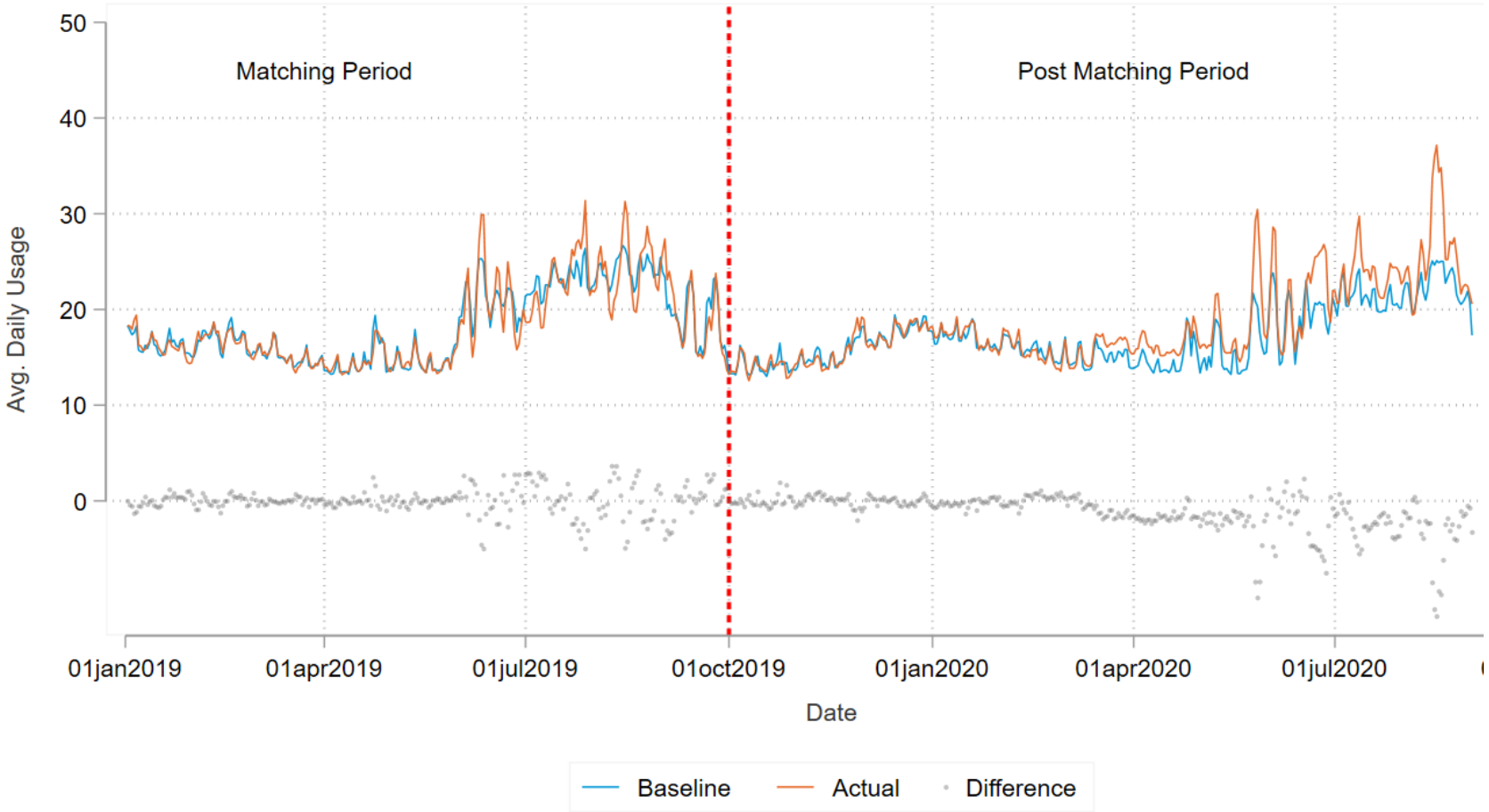
			% Bias		Accuracy (RRMSE)	
Regression Method	Matching Method	Matching Pool	Pre-COVID	COVID	Pre-COVID	COVID
Individual Regressions	-	-	0.290%	-4.553%	0.049	0.059
Individual Regressions with Synthetic Controls	-	-	-0.001%	-0.508%	0.016	0.016
Individual Regressions	Stratified Matching	Non-Participants from PG&E Population	-0.041%	-0.032%	0.029	0.029
Individual Regressions	Stratified + PSM	Non-Participants from PG&E Population	0.102%	-0.486%	0.027	0.016
Individual Regressions	Stratified Matching	Former EE Participants	0.110%	-1.605%	0.025	0.027
Individual Regressions	Stratified + PSM	Former EE Participants	-0.105%	-0.256%	0.032	0.026

- To examine accuracy during COVID, analyzed loads for customers that enrolled late in 2020 (after August 2020)
  - Trained using 2019 data
  - Predicted usage during 2020 pre-treatment period
- Compared model accuracy before & after COVID shutdowns
- Overall, models with comparison groups perform better both before and during COVID, substantially reducing % bias during COVID period

Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

P4P PREDICTED VS. ACTUAL LOAD – NO COMPARISON GROUP

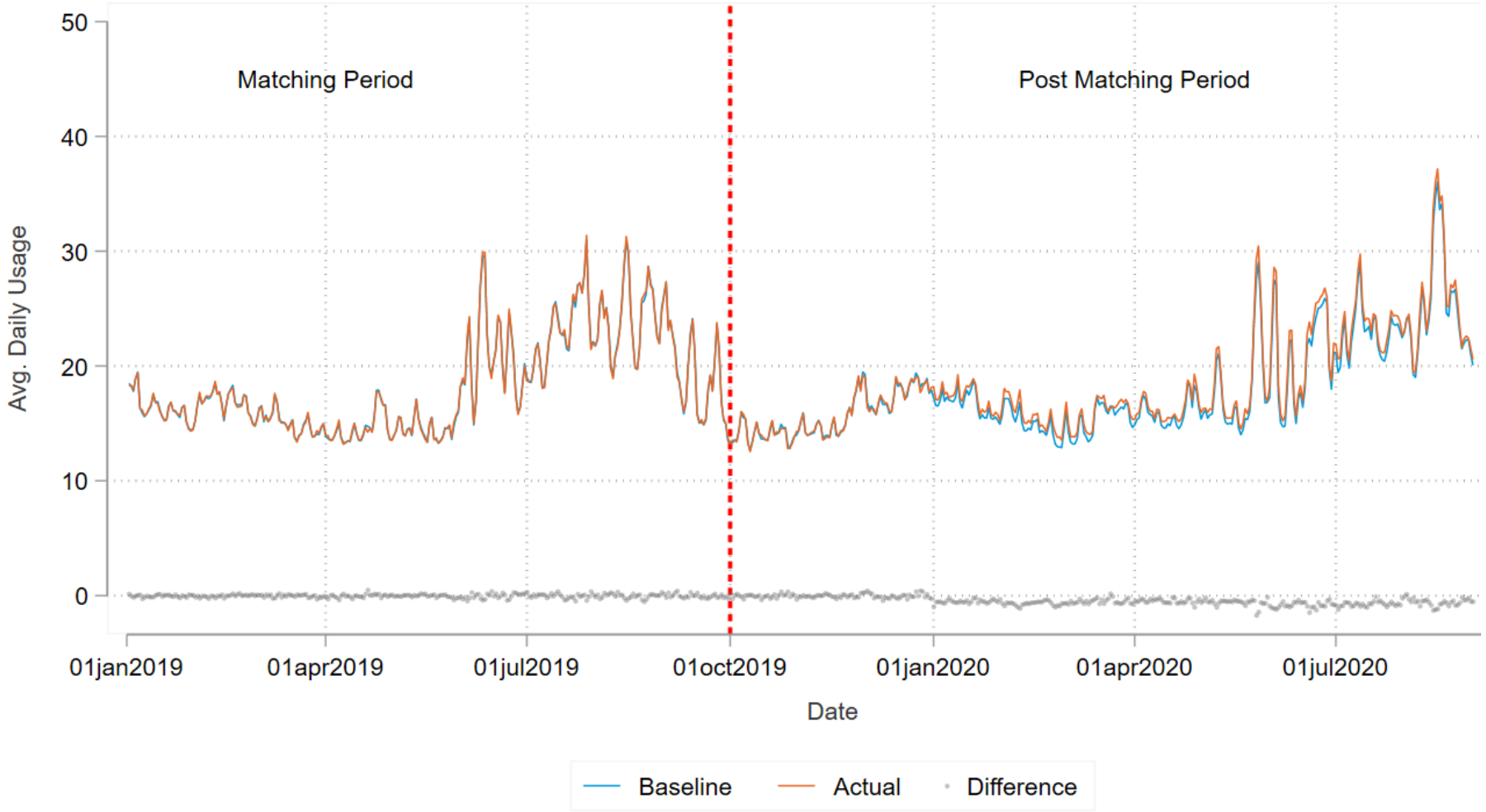
- Counterfactual models load as a function of temperature and time of year using pre-treatment data
- Generally under-predicts load during COVID



Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

P4P PREDICTED VS. ACTUAL LOAD – WITH SYNTHETIC CONTROLS

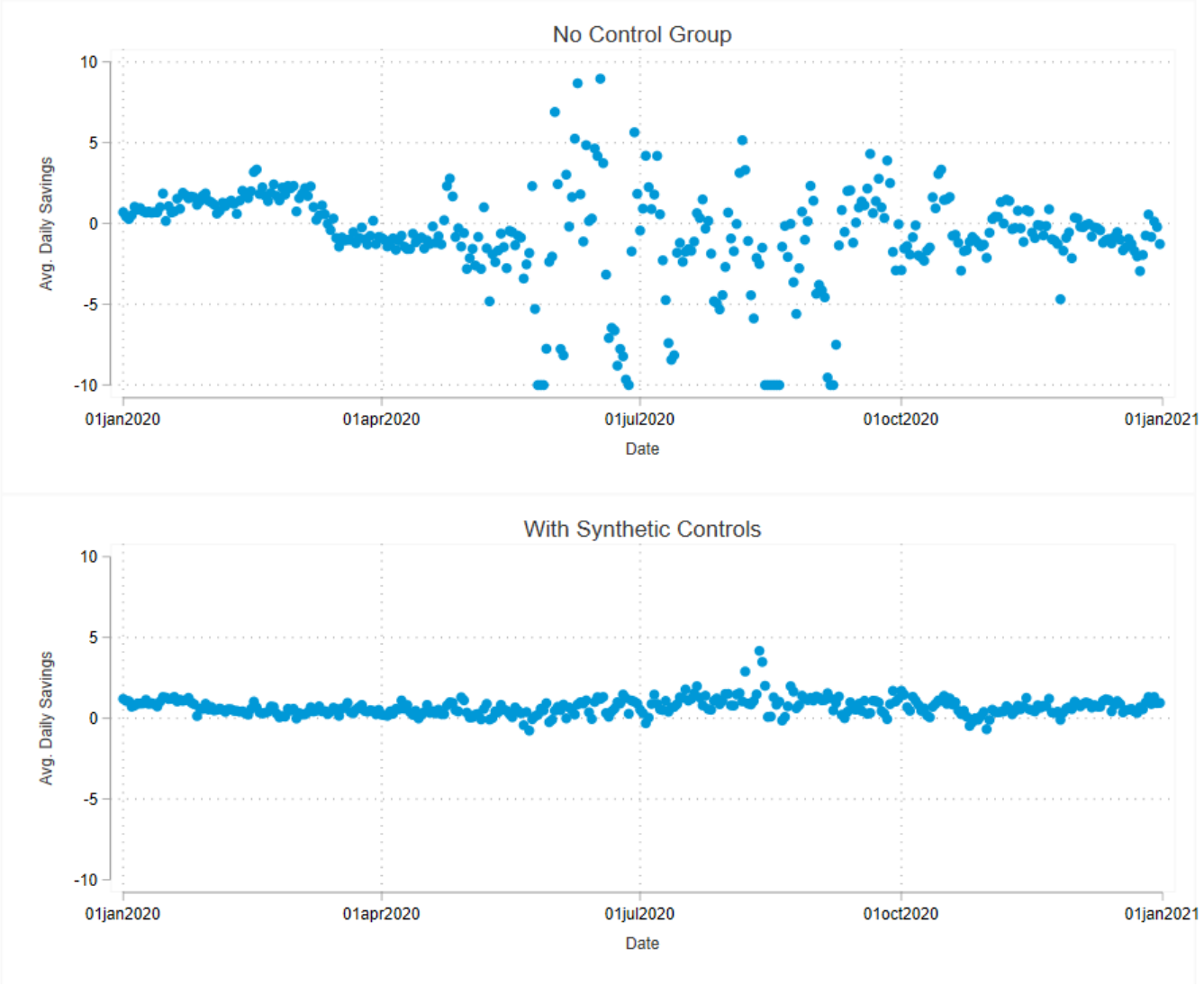
- Counterfactual takes into account observed behavior by comparison group
- Generally much higher accuracy during COVID period



Demand Side Analytics
DATA DRIVEN RESEARCH AND INSIGHTS

IMPACT ON SAVINGS*

- Savings negative and noisy without comparison group once COVID begins
- Once comparison group is added, we see relatively consistent and positive savings before and after COVID



*Savings shown are only for a subset of participants and do not represent overall P4P savings

QUESTIONS?

Adriana Ciccone, Principal
Demand Side Analytics
aciccone@demandsideanalytics.com

Stephanie Bieler, Senior Consultant
Demand Side Analytics
sbieler@demandsideanalytics.com