CalTRACK Documentation

Release 2.0

Phil Ngo

May 18, 2023

Contents:

1 CalTRACK Methods **1**

1.1 Section 1: Overview . . . 1
1.2 Section 2: Data Management . . . 3
1.3 Section 3(a): Modeling - Billing and Daily Methods . . . 6
1.4 Section 3(b): Modeling - Hourly Methods . . . 11
1.5 Section 4: Aggregation . . . 14

2 Technical Appendix **17**

2.1 2.2.1.2. For fitting baseline models using the hourly methods, no minimum baseline period length is required. . . . 17
2.2 2.2.4.1. Temperature data may not be missing for more than six consecutive hours. . . . 17
2.3 2.4.1. Weather station to be used is closest within climate zone that meets CalTrack data sufficiency requirements. . . . 17
2.4 3.1.3. Models are fit to baseline data in the 365 days immediately prior to the intervention start date. . . 19
2.5 3.1.4.1. Select and qualify balance points for candidate models for each period for each meter. . . . 21
2.6 3.2.1. A grid search of models is performed using a wide range of candidate balance points. . . . 23
2.7 3.2.3. Maximum gap between candidate balance points in the grid search is 3 degrees F or the equivalent in degrees C. . . . 25
2.8 3.3.1.2. Independent variables . . . 26
2.9 3.4.1. Models using daily data are fit using ordinary least squares. . . . 32
2.10 3.4.3.2. Candidate model qualification. . . . 33
2.11 3.4.3.3. The model with highest adjusted R-squared will be selected as the final model. . . . 35
2.12 3.6.5. Baseline periods. . . . 39

CHAPTER 1

CalTRACK Methods

(Version 2.0)

1.1 Section 1: Overview

1.1. Introduction.

1.1.1. CalTRACK is a set of methods for estimating avoided energy use (AEU), related to the implementation of one or more energy efficiency measures, such as an energy efficiency retrofit or a consumer behavior modification. CalTRACK methods yield whole building, site-level savings outputs. Portfolio-level savings confidence is measured by aggregating the performance of a number of individual sites and calculating portfolio fractional savings uncertainty. The primary CalTRACK use case is energy efficiency procurement (such as Pay-for-Performance and Non-Wires Alternatives). As such, key considerations are replicability and availability of data. The methods described in this document require only commonly-available site-level meter and weather data.

1.1.2. CalTRACK methods will be familiar to energy efficiency measurement and verification (M&V) practitioners and program evaluators (EM&V). These methods are similar to ASHRAE Guideline 14, IPMVP Option C, and the Uniform Methods Project (UMP). In addition to a long history of use in project and program evaluation, these approaches draw on a methodological foundation developed in the more general statistical literature on piecewise linear regression or segmented regression for policy analysis and effect estimates that is used in fields as diverse as public health, medical research, and econometrics.

1.1.3. CalTRACK methods are developed through a consensus process, organized as a series of "sprints" and versioned (e.g., 1.0, 2.0, etc.). Older versions are archived for reference, but new versions should be considered to reflect the consensus guidance of the working group. As a rule, CalTRACK methods are designed to inform a technical audience of practitioners, however, where useful, a pragmatic effort is made to provide specific policy guidance when doing so facilitates reproducibility.

1.1.4. In the methodological appendix, this document also references more detailed descriptions of testing procedures and alternative approaches that have already or have not yet been tested. In all cases, the

CalTRACK technical working group has sought to justify technical and procedural choices using referenceable and reproducible empirical analysis.

1.1.5. This is a "living" document, in the sense that the working group continues to meet to propose, test, discuss, and enact changes to the methods. As such, recommendations may change as additional evidence arises supporting new approaches or challenging current approaches. The CalTRACK methods are version controlled to allow users to pin their programs, policies, or analysis to a particular instantiation of the methods. However, given that the latest version will always reflect what the working group believes to be the most complete and well-tested approaches, users of the CalTRACK methods are encouraged to create policies which make room for continued incorporation of methodological improvements.

1.1.6. To further assist with referenceability and versioning, this document provides a numbering scheme to facilitate referencing the methods. As the numbering scheme may change as the methods are tested and refined, it should should be used in combination with a document version to prevent ambiguity.

1.1.7. The CalTRACK methods are designed to be general enough to implement using any general-purpose programming language and are not required to be deployed with any particular software implementation. To the extent that code samples do appear in this document, they are intended as to be used as pseudo-code.

1.2. Participation in methods development.

1.2.1. The technical working group meetings are open to all and technical contribution is highly encouraged. Efforts to propose changes which are in-scope (see 1.3.2), and empirically and replicably tested will generally be met with interest and engagement. Efforts to propose changes which are out-of-scope (see 1.3.3), are discouraged.

1.2.2. Ongoing discussions within the technical working group can be followed in the CalTRACK Github Issues section, which should reference particular proposed changes to this document. These are also noted and linked where applicable in this document.

1.2.3. Interested parties may sign up to participate in the CalTRACK technical working group proceedings by registering at http://www.caltrack.org.

1.3. Scope.

1.3.1. Some items have been considered to be generally in-scope for CalTRACK methods development.

1.3.1.1. Technical clarifications or improvements to existing methods.

1.3.1.2. Methods for calculating uncertainty or establishing criteria for use cases.

1.3.1.3. Empirical tests to evaluate methodological choices and assumptions.

1.3.2. Some items have, to date, been considered out-of-scope for CalTRACK methods development:

1.3.2.1. Pooled analysis.

1.3.2.2. Data sources which prove to be impossible to entirely standardize, such as measures performed.

1.3.2.3. Measure attribution.

1.3.2.4. Programming language or implementation-specific constraints.

1.3.2.5. Proprietary, closed-source, or restrictively-licensed algorithms, procedures, or content.

1.3.2.6. Analysis or evidence based in datasets which cannot be otherwise replicated by working group members.

1.4. Definitions.

1.4.1. Project. An event, action, or set of measures implemented which is expected to have some effect on building energy consumption, such as a retrofit, performance or installation of a measure, or a behavioral intervention.

1.4.2. Baseline period. A period of time during which data is gathered to calculate the relationship between metered consumption and weather conditions prior to a project.

1.4.3. Reporting period. A period of time during which meter and weather data is gathered following a project to calculate the avoided energy use.

1.4.4. Intervention period. A period of time between the end of the baseline period and the beginning of the reporting period in which a project is being installed. The intervention period is typically removed from the analysis because of erratic energy use during project installation. If a behavior intervention, the intervention period is typically one day. If multiple project dates are given, e.g., for multiple measure installations, use the earliest intervention date as project start date and the latest date as the project completion date. Recommend flagging for review if dates are more than one year apart.

1.4.5. Avoided Energy Use. The difference between (1) energy use predictions derived from a baseline energy model in conjunction with weather data during the reporting period, and (2) actual energy use observed in a reporting period.

1.4.6. Procurer. The party that buys energy savings stemming from energy efficiency interventions, usually a utility.

1.4.7. Aggregator. The party that supplies energy savings projects to a procuring entity and generally takes on the associated performance risk.

1.4.8. Site. An endpoint on the grid where energy consumption is monitored by one or more physical meters.

1.4.9. Energy consumption. A value derived from a physical meter based on the amount delivered over particular intervals of time.

1.2 Section 2: Data Management

2.1. Data Inputs.

The data requirements to apply CalTRACK methods to a single metered site are listed in this section. These represent the "ideal". Additional constraints and sufficiency requirements follow in section (2.2) and considerations for handling missing or incomplete data follow.

2.1.1. Energy consumption data (meter data). This data must have the following qualities:

2.1.1.1. Periods of usage, usage during those periods. Can be provided as billing data or as AMI data.

2.1.1.2. May be combined from multiple sources or accounts.

2.1.1.3. Must be converted to units of energy consumption, not supplied volume. This can be achieved, for example, by using a therm factor conversion.

2.1.1.4. Must be subject to the constraints outlined in 2.2.

2.1.1.5. Flag or directional indicator for presence of net metering.

2.1.2. Set of candidate weather station sites. Each weather station should have the following data available:

2.1.2.1. Weather station site latitude and longitude coordinates.

2.1.2.2. Climate zones information if needed in weather station matching (see 2.4).

2.1.2.3. IECC Climate Zone.

2.1.2.4. IECC Moisture Regime.

2.1.2.5. Building America Climate Zone.

2.1.2.6. California Building Climate Zone Area (if site is in the state of California).

2.1.2.7. Observed dry-bulb temperature data, subject to the data requirements outlined in 2.2.

2.1.3. Project data.

2.1.3.1. Date(s).

2.1.3.1.1. Project start date. The date of the beginning of the intervention period (see 1.4.4) during which energy use will be ignored. If the exact start date is not known, an estimate may be used in place of a recorded start date. The estimate should err on the side of an earlier start date.

2.1.3.1.2. Intervention completion date. The date of the end of the intervention period. This date marks the beginning of the reporting period (see 1.4.3). If the exact completion date is not known, an estimate may be used in place of a recorded completion date. The estimate should err on the side of a later start date.

2.1.3.1.3. Intervention active date. For certain interventions without a defined "project start" (e.g. behavioral interventions), the date when a behavioral intervention(s) became active. Use this as the intervention completion date as well.

2.1.3.1.4. Baseline period end. Either the project start date or the intervention active date, depending on intervention type.

2.1.4. Building site data.

2.1.4.1. Latitude and longitude coordinates. Should be four decimal places or more.

2.1.4.1.1. In the absence of a high quality geocode, the latitude and longitude coordinates of the centroid of the ZIP Code Tabulation Area (ZCTA) may be used instead. ZCTA should be used in preference to ZIP code, as ZIP codes are not associated with geographic boundaries (see https://www.census.gov/geo/reference/zctas.html).

2.1.5. Climate zone (see 2.1.2.2).

2.1.6. Time zone.

2.2. Data constraints.

2.2.1. Missing values and data sufficiency for baseline period.

2.2.1.1. Consumption and temperature data should be sufficient to allow for a 365-day baseline period.

2.2.1.2. Number of days of consumption and temperature data missing should not exceed 37 days (10%) for billing and daily methods. For fitting baseline models using the hourly methods, no minimum baseline period length is required. However, baseline consumption data must be available for over 90% of hours in the same calendar month as well as in each of the previous and following calendar months in the previous year. *See Appendix for Details.*

2.2.1.3. Data is considered missing if it is clearly marked by the data provider as NULL, NaN, or similar.

2.2.1.4. Values of 0 are considered missing for electricity data, but not gas data.

2.2.2. Daily data is considered sufficient for baseline period under the following conditions:

2.2.2.1. If summing to daily usage from higher frequency interval data, no more than 50% of high-frequency values should be missing. Missing values should be filled in with average of non-missing values (e.g., for hourly data, 24 * average hourly usage).

2.2.2.2. Although this is more common in billing data than in interval data, if periods are estimated they should be combined with subsequent periods.

2.2.2.3. If daily average temperatures are calculated by averaging higher frequency temperature data, no more than 50% of high-frequency temperature values should be missing.

2.2.3. Billing data is considered sufficient for baseline period under the following conditions:

2.2.3.1. Estimated periods values should be combined with next period up to a 70-day limit. Estimated periods are counted as missing data for the purpose of determining data sufficiency to limit the number of estimated reads used for analysis.

2.2.3.2. If average temperatures for billing periods are calculated by averaging higher frequency temperature data, the high-frequency temperature data must cover 90% of each averaged billing period.

2.2.3.3. If daily average temperatures are calculated by averaging higher frequency temperature data, no more than 50% of high-frequency temperature values should be missing.

2.2.3.4. Off-cycle reads (spanning less than 25 days) should be dropped from analysis. These readings typically occur due to meter reading problems or changes in occupancy.

2.2.3.5. For pseudo-monthly billing cycles, periods spanning more than 35 days should be dropped from analysis. For bi-monthly billing cycles, periods spanning more than 70 days should be dropped from the analysis.

2.2.4. Hourly temperature data is considered sufficient under the following conditions:

2.2.4.1 Temperature data may not be missing for more than six consecutive hours. Missing temperature data may be linearly interpolated for up to 6 consecutive missing hours. *See Appendix for Details.*

2.2.5. Data spanning beyond the period should not be used in analysis.

2.2.6. Projects should be excluded if net metering (i.e., for photovoltaics or other on-site generation) status changes during the baseline period.

2.2.6.1. Exception: Future efforts may provide the ability to access sub-meter data that may allow for backing out onsite generation and storage to arrive at savings. Currently, this data is not readily obtained.

2.2.7. Projects should be flagged if electric vehicle charging is installed during the baseline period.

2.3. Guidelines for handling data quality issues.

In many cases, data quality issues can be resolved by going back to the source to resolve issues in export or transfer. This guidance is a second line of defense for handling or correcting for common data issues, and are provided in the hope of mitigating the myriad issues and discrepancies which arise using different methods for data cleaning.

2.3.1. Impossible dates.

2.3.1.1. If conducting billing analysis, and if day of month is impossible (e.g., 32nd of Jan), use first of month.

2.3.1.2. If month (e.g., 13) or year (e.g. 2051) is impossible flag the date and remove it from the dataset. Check for mis-coding, such as 2015 -> 2051.

2.3.2. Duplicated meter or temperature records.

2.3.2.1. Combine available versions into a single time series by dropping duplicate records, using the most complete version possible. If a record for a particular timestamp conflicts with another version, flag the project for possible existence of multiple meters or submeters. If this is confirmed, the usage from multiple meters may be aggregated.

2.3.3. Ensure that meter and temperature data is using matching and correct timezone and daylight-savings handling across all data sources.

2.3.4. NOAA weather is sampled roughly hourly with minute-level timestamps. This should be converted to hourly by first computing a minute-resolution time series using near interpolation of data points with a limit of 60 minutes, then downsampling to hourly temperature by taking mean of linearly-interpolated minute-level readings.

2.3.5. Negative meter data values should be flagged for review as they indicate the possible unreported presence of net metering.

2.3.6. Extreme values: Usage values that are more than three interquartile ranges larger than the median usage should be flagged as outliers and manually reviewed.

2.3.7. Generally recommend an audit for dataset completeness using expected counts of sites, meters, and projects.

2.3.8. Roll up data if not given with expected frequency.

2.4. Matching a site to a weather station.

2.4.1. Weather station to be used is closest within climate zone that meets CalTrack data sufficiency requirements. *See Appendix for Details.*

2.4.1.1. If there are no weather stations within that climate zone, fallback to closest weather station that has complete data.

2.4.2. Matches further than 200 km should be flagged for review, as these distant matches may sacrifice interpretability of the model.

1.3 Section 3(a): Modeling - Billing and Daily Methods

3.1. Overview of usage per day model strategy.

3.1.1. Model intuition.

3.1.1.1. Building is modeled as base load, heating load, and cooling load. Heating load and cooling load are assumed to have a linear relationship with heating and cooling demand, as approximated by heating and cooling degree days, beyond particular heating and cooling balance points.

3.1.2. Model foundations in literature. Modeling does not strictly adhere to these methods, but draws from them for inspiration.

3.1.2.1. PRISM.

3.1.2.2. Uniform Methods Project for Whole Home Building Analysis.

3.1.2.3. California Evaluation Project.

3.1.3. Models are fit to baseline data in the 365 days immediately prior to the intervention start date, provided the data sufficiency criteria are met. *See Appendix for Details.*

3.1.4. Follow the process outlined below and detailed in subsequent sections:

3.1.4.1. Select and qualify balance points for candidate models for each period for each meter. *See Appendix for Details.*

3.1.4.2. Use hourly temperature from the matched weather station (11).

3.1.4.3. Compute design matrixes, fit, and qualify all candidate models.

3.1.4.4. Select best candidate model.

3.1.4.5. Compute estimated values.

3.1.4.6. Compute measured values.

3.1.4.7. Compute savings.

3.1.4.8. Aggregate across sites.

3.2. Select and qualify balance points.

3.2.1. A grid search of models is performed using a wide range of candidate balance points. *See Appendix for Details.*

3.2.1.1. Recommended cooling balance point range is from 30 to 90 degrees F. For analysis of natural gas consumption, models using cooling degree days are not considered.

3.2.1.2. Recommended heating balance point range is from 30 to 90 degrees F.

3.2.2. Constraints and qualification. Only model balance points or balance point combinations for which:

3.2.2.1. Cooling balance point $\geq$ heating balance point.

3.2.2.2. Have enough numbers of non-zero degree days. This is in order to avoid overfitting in the case where only a few days exist with usage and nonzero degree-days, and the usage happens by chance to be unusually high on those days.

3.2.2.2.1. At least 10 days with non-zero degree days per year. This requirement does not apply when using billing data.

3.2.2.2.2. At least 20 degree days per year.

3.2.3. Maximum gap between candidate balance points in the grid search is 3 degrees F or the equivalent in degrees C. *See Appendix for Details.*

3.3. Computing design matrix for each model.

3.3.1. Basic structure applies to analysis using both daily and billing periods.

3.3.1.1. Dependent variable: average usage per day for a usage period.

3.3.1.2. Independent variables. *See Appendix for Details.*

3.3.1.2.1. Average cooling degree days per day for a usage period.

3.3.1.2.2. Average heating degree days per day for a usage period.

3.3.1.3. Fitted model parameters.

3.3.1.3.1. intercept (interpreted as daily base load).

3.3.1.3.2. $_H$ is the slope.

3.3.1.3.3. $_C$ is the slope.

3.3.2. Equation: $UPD_{p,i} = +_{H,i} * HDD_p +_{C,i} * CDD_p +_{p,i}$, where:

3.3.2.1. $UPD_{p,i}$ is average use (gas in therms, electricity in kWh) per day during period p for site i.

3.3.2.2. is the mean use for site i, or intercept.

3.3.2.3. $_{H,i}$ is the heating coefficient for site i. It represents the incremental change in energy use per day for every additional heating degree day.

3.3.2.4. $_{C,i}$ is the cooling coefficient for site i. It represents the incremental change in energy use per day for every additional cooling degree day.

3.3.2.5. HDD_p is the average number of heating degree days per day in period p, which is a function of the selected balance point temperature, the average daily temperatures from the weather station matched to site i during the period p, and the number of days in period p with matched usage and weather data for site i.

3.3.2.6. CDD_p is the average number of cooling degree days per day in period p, which is a function of the selected balance point temperature, the average daily temperatures from the weather station matched to site i during the period p, and the number of days in period p with matched usage and weather data for site i.

3.3.2.7. n is the site specific random error term for a given period.

3.3.3. Computing average usage per day (UPD) for each period.

3.3.3.1. $UPD_p = \frac{1}{n_p} * (U_d)$, where

3.3.3.2. UPD_p is the average use per day for a given period p.

3.3.3.3. (U_d) is the sum of all daily use values U_d for a given period p.

3.3.3.4. n_p is the total number of days for which daily use values U_d were available in period p.

3.3.3.5. Boundaries between days should occur at midnight of the local time zone.

3.3.4. Cooling degree days for each particular balance point.

3.3.4.1. CDD values are calculated as follows:

3.3.4.1.1. $CDD_p = \frac{1}{n_{d,p}} * (max(avg(T_d) - CDD_b, 0))$, where

3.3.4.1.2. CDD_p = Cooling degree days for period p.

3.3.4.1.3. CDD_b = the CDD balance point that provides best model fit.

3.3.4.1.4. $n_{d,p}$ is the total number of days elapsed between the start time of the period p and the end time of the period p.

3.3.4.1.5. $()$ = the sum of values in () over each day d in period p.

3.3.4.1.6. $max()$ = the maximum of the two values in ().

3.3.4.1.7. $avg(T_d)$ = the average temperature for day d.

3.3.5. Heating degree days for each particular balance point.

3.3.5.1. HDD values are calculated as follows:

3.3.5.1.1. $HDD_p = \frac{1}{n_{d,p}} * (max(HDD_b - avg(T_d), 0))$, where

3.3.5.1.2. HDD_p = Average heating degree days per day for period p.

3.3.5.1.3. HDD_b = the HDD balance point that provides best model fit.

3.3.5.1.4. $n_{d,p}$ is the total number of days elapsed between the start time of the period p and the end time of the period p.

3.3.5.1.5. $()$ = the sum of values in () over each day d in period p.

3.3.5.1.6. $max()$ = the maximum of the two values in ().

3.3.5.1.7. $avg(T_d)$ = the average temperature for day d.

3.4. Fit candidate models.

3.4.1. Models using daily data are fit using ordinary least squares. *See Appendix for Details.*

3.4.2. Models using billing data are fit using weighted least squares regression. Use the corresponding number of days np as the weight for each billing period.

3.4.3. For each meter at each site, the choice must be made between using one of the single parameter models (just HDD or CDD) or combined models (HDD and CDD). This choice is called model selection. A range of candidate models is fitted for each qualified balance point, then the most appropriate single qualified model, as estimated using the metric below, is used to calculate estimated quantities.

3.4.3.1. Given the selected balance point ranges, all combinations of candidate balance points are tried. Models are as follows:

3.4.3.1.1. HDD and CDD (electricity only):

$$UPD_{p,i} =_i +_{H,i} * HDD_p +_{C,i} *CDD_p +_{p,i}$$

3.4.3.1.2. HDD only:

$$UPD_{p,i} =_i +_{H,i} * HDD_p +_{p,i}$$

3.4.3.1.3. CDD only: (electricity only):

$$UPD_{p,i} =_i +_{C,i} * CDD_p +_{p,i}$$

3.4.3.1.4. Intercept-only:

$$UPD_{p,i} =_i +_{p,i}.$$

In this case, adjusted R-squared is 0 by definition.

3.4.3.2. Candidate model qualification. If each parameter estimate is not negative, then the model qualifies for inclusion in model selection. *See Appendix for Details.*

3.4.3.2.1. $_H > 0$

3.4.3.2.2. $_C > 0$

3.4.3.2.3. $_i > 0$

3.4.3.3. The model with highest adjusted R-squared will be selected as the final model. *See Appendix for Details.* Adjusted R-squared will be defined as:

3.4.3.3.1. $R^2_{adj} = 1 - \frac{\frac{SS_{res}}{df_e}}{\frac{SS_{tot}}{df_t}}$, where

3.4.3.3.2. SS_{res} is the sum of squares of residuals.

3.4.3.3.3. df_e is the degrees of freedom of the estimate of the underlying population error variance, and is calculated using $(P - c - 1)$, where P is the number of periods (e.g. days or billing periods) in the baseline used to estimate the model and c is the number of explanatory variables, not including the intercept term.

3.4.3.3.4. SS_{tot} is the total sum of squares

3.4.3.3.5. df_t is the degrees of freedom of the estimate of the population variance of the dependent variable, and is calculated as $(P-1)$, were P is the number of periods (e.g. days or billing periods) in the baseline used to estimate the model.

3.5. Missing Data in Reporting Period.

3.5.1. Missing temperature values and data sufficiency for reporting period.

3.5.1.1. If a day is missing a temperature value, the corresponding consumption value for that day should be masked.

3.5.1.2. If daily average temperatures are calculated by averaging higher frequency temperature data, no more than 50% of high-frequency temperature values should be missing.

3.5.1.3. Missing values should be filled in with average of non-missing values (e.g., for hourly data, average hourly temperature).

3.5.1.4. Data is considered missing if it is clearly marked by the data provider as NULL, NaN, or similar.

3.5.2. Missing consumption values and data sufficiency for reporting period.

3.5.2.1. If a day is missing a consumption value, the corresponding counterfactual value for that day should be masked.

3.5.2.2. Data is considered missing if it is clearly marked by the data provider as NULL, NaN, or similar.

3.5.2.1. Values of 0 are considered missing for electricity data, but not gas data.

3.5.3. Estimating counterfactual usage when temperature data is missing.

3.5.3.1. Counterfactual usage is not calculated when daily temperature data is missing, pending further methodological discussion.

3.5.4. Estimating avoided energy usage when consumption data is missing.

3.5.4.1. Avoided energy use is not calculated when consumption data is missing.

3.5.5. Billing data in the reporting period.

3.5.5.1. Estimated periods values should be combined with next period up to a 70-day limit.

3.5.5.2. If average temperatures for billing periods are calculated by averaging higher frequency temperature data, the high-frequency temperature data must cover 90% of each averaged billing period.

3.5.5.3. Off-cycle reads (spanning less than 25 days) should be combined with next period up to a 70 day limit. These readings typically occur due to meter reading problems or changes in occupancy.

3.5.5.4. For monthly billing cycles, periods spanning more than 35 days should be flagged for review. For bi-monthly billing cycles, periods spanning more than 70 days should be flagged for review.

3.5.6. Projects should be excluded if net metering (i.e., for photovoltaics or other on-site generation) status changes during the reporting period.

3.5.6.1. Exception: Future efforts may provide the ability to access sub-meter data that may allow for backing out onsite generation and storage to arrive at savings. Currently, this data is not readily obtained.

3.6. Computing derived quantities for billing and daily.

3.6.1. Avoided energy use (AEU) for each time period in the reporting period is calculated as follows.

$$AEU_{p,i} = n_p * ({}_i + {}_{H,i} * HDD_p + {}_{C,i} * CDD_p - UPD_p)$$

3.6.1.1. The coefficients ${}_i$, ${}_{H,i}$, ${}_{C,i}$ are those from the final model.

3.6.1.2. HDD_p and CDD_p are calculated using weather data in the reporting period according to guidelines in Section 3.3.

3.6.1.3. UPD_p is the usage per day calculated for a period p using the same procedure as in Section 3.3.3.

1.4 Section 3(b): Modeling - Hourly Methods

3.7.2. CalTRACK implementation. CalTRACK recommends the use of the TOWT model with standardization of certain user-defined inputs to model hourly load and energy savings.

3.7.3. Model foundation in literature. The model is described in these publications:

1. Mathieu et al., Quantifying Changes in Building Electric Load, With Application to Demand Response. IEEE Transactions on Smart Grid 2:507-518, 2011
2. Price P, Methods for Analyzing Electric Load Shape and its Variability. Lawrence Berkeley National Laboratory Report LBNL-3713E, May 2010.

3.7.4. Models are fit to baseline data immediately prior to the baseline end date, provided the data sufficiency criteria are met.

3.7.5. Baseline periods. *See Appendix for Details.* Instead of using a single baseline model for estimating the counterfactual during all times of the year, predicting the counterfactual during any time period will be done using the baseline model for that calendar month ("month-by-month" models). This implies that there can be up to 12 separate models for a particular building - one for predicting the counterfactual in each calendar month. Each model will be fit using baseline data comprising (i) data from the same calendar month in the 365 days prior to the intervention date. These data points will be given full weight when fitting the model, (ii) data from the previous and subsequent calendar months in the 365 days prior to the intervention date. These data points will be given a weight of 0.5 when fitting the model. For example, for a project installed in March 2018, predicting the counterfactual in July 2018 will be done using a model fit to baseline data from June, July and August 2017, with weights of 0.5, 1 and 0.5 assigned to the data points in those three months.

3.7.5.1. In some cases, building energy use patterns are consistent from month to month and a single model with a 365 day baseline may be used, as long as the normalized mean bias error (NMBE) for each month in the baseline period is calculated separately and no more than two months have NMBE larger than an acceptable threshold (default, 1%).

3.7.5.2. A single model with a 365 day baseline may also be used if temperature coverage in the baseline period is insufficient. In particular, if either of these conditions is satisfied:

$$T_{min,reporting} < T_{min,baseline} - 0.1 * (T_{max,baseline} - T_{min,baseline})$$

or

$$T_{max,reporting} > T_{max,baseline} + 0.1 * (T_{max,baseline} - T_{min,baseline})$$

where, the subscripts min and max refer to the minimum and maximum temperatures in the baseline and reporting periods.

3.7.6. The following procedures (3.7-3.11) will be applied separately for each "month-by-month" model or to the full 365-day model.

3.8. Occupancy estimation.

3.8.1. Overview. The sensitivity of building energy use to temperature may vary depending on the "occupancy" status. This is handled in the Time-Of-Week and Temperature model by segmenting the times-of-week into periods of high load and low load (also referred to as occupied/unoccupied, although the states may not necessarily correspond to occupancy changes). The segmentation is accomplished using the residuals of a HDD-CDD model developed as follows.

3.8.2. Time-of-week. A week is divided into 168 hourly time-of-week intervals starting on Monday. For example, interval 1 is from midnight to 1 a.m. on Monday morning, interval 2 is from 1 a.m.-2 a.m. and so on. Dummy variables TOW_p (and consequently separate coefficients α_i) are included in the model for each time of week.

3.8.3. Regression to determine occupancy status. A single HDD and CDD weighted least squares (WLS) model is fit to the baseline dataset (defined pursuant to 3.6.5) using fixed balance points (50 for heating and 65 for cooling):

$$UPH_{pi} = \mu_i + \beta_{Hi} HDH50_p + \beta_{Ci} CDH65_p + \epsilon_{pi}$$

where

3.8.3.1. UPH_{pi} is the usage per hour for period p.

3.8.3.2. μ_i is the mean use for the site.

3.8.3.3. β_{H_i} is the heating coefficient.

3.8.3.4. $HDH50_p$ is the heating degree-hour with a 50-degree balance point.

3.8.3.5. β_{C_i} is the cooling coefficient.

3.8.3.6. $CDH65_p$ is the cooling degree hour with a 65-degree balance point.

3.8.4. The predictions of this model are calculated for each data point in the baseline period. A prediction is flagged as an underprediction if the actual measured value exceeds the prediction. The data points are then grouped by the time-of-week and the percentage of underpredictions for each time-of week is calculated. If this value exceeds 65%, then the corresponding time-of-week is flagged as "Occupied," otherwise, it is flagged as "Unoccupied." These flags are expressed in a binary variable (0/1) for the unoccupied and occupied modes, respectively.

3.9. Temperature variables.

3.9.1. For each data point in the baseline dataset, the outdoor air temperature is used to calculate up to 7 new binned features using the following algorithm:

3.9.1.1. Six bin endpoints B_n are defined at 30, 45, 55, 65, 75 and 90 degrees Fahrenheit, which define 7 temperature bins (<30, 30-45, 45-55, 55-65, 65-75, 75-90, >90).

3.9.1.2. These bin endpoints are validated for each model by counting the number of hours with temperatures within these bins. Bins with fewer than 20 hours are combined with the next closest bin by dropping the larger bin endpoint, except for the largest bin, where the lower endpoint is dropped. The N valid bin endpoints are then used to develop the binned temperature features.

3.9.1.3. If the temperature $T_p > B_1$, then the first temperature feature $Tc_{1,p} = B_1$ and the algorithm proceeds to the next step. Otherwise, $Tc_{1,p} = T_p$, and $Tc_{n,p} = 0$, for $n = 2...N$, and the algorithm ends.

3.9.1.4. For $n = 2...N$, if the temperature $T_p > B_n$, then $Tc_{n,p} = B_n - B_{n-1}$ and the algorithm proceeds to the next n. Otherwise, $Tc_{n,p} = T_p - B_{n-1}$, and $Tc_{n,p} = 0$, for $n = (n+1)...N$, and the algorithm ends.

3.9.1.5. If the temperature $T_p > B_N$, then the last temperature feature $Tc_{N+1,p} = T_p - B_N$, and equal to zero otherwise.

3.9.2. Example of temperature binning outputs using the default temperature bin endpoints:

Bin:	<30	30-45	45-55	55-65	65-75	75-90	>90
T_p	$T_{1,p}$	$T_{2,p}$	$T_{3,p}$	$T_{4,p}$	$T_{5,p}$	$T_{6,p}$	$T_{7,p}$
20	20	0	0	0	0	0	0
40	30	10	0	0	0	0	0
50	30	15	5	0	0	0	0
60	30	15	10	5	0	0	0
70	30	15	10	10	5	0	0
80	30	15	10	10	10	5	0
100	30	15	10	10	10	15	10

3.10. Compute Design Matrix.

3.10.1. The following structure applies to the hourly design matrix.

3.10.1.1. Dependent variable: total energy consumption per hour.

3.10.1.2. Independent variables:

3.10.1.2.1. Seven (or fewer) temperature features developed according to Section 3.9.

3.10.1.2.2. 168 binary dummy variables indicating the time-of-week.

3.10.1.2.3. An occupancy binary variable developed according to Section 3.8 interacted with the temperature and time-of-week variables.

3.10.1.3. Fitted model parameters.

3.10.1.3.1. Seven (or fewer) temperature coefficients for the occupied mode and seven (or fewer) temperature coefficients for the unoccupied mode.

3.10.1.3.2. 168 time-of-week coefficients for the occupied mode and 168 time-of-week coefficients for the unoccupied mode.

3.10.1.3.3. No separate intercept term is used in this regression.

3.11. Fit Models.

3.11.1. Model specification. One weighted least squares (WLS) regression is fit to the data with the following specification:

$$UPH_{pi} = \sum \alpha_t TOWp + \sum \beta_{T,n} Tc_{n,p} + \sum occupied \alpha_t TOWp + \sum occupied \beta_{T,n} Tc_{n,p} + \epsilon_{pi}$$

3.11.2. Weights are assigned to the data points from different calendar months in accordance with Section 3.7.5.

3.12. Computing derived quantities for hourly methods.

3.12.1. Avoided energy use (AEU) for each time period in the reporting period is calculated as follows:

$$AEU_{pi} = \sum \alpha_t TOWp + \sum \beta_{T,n} Tc_{n,p} + \sum occupied \alpha_t TOWp + \sum occupied \beta_{T,n} Tc_{n,p} - UPH_p.$$

1.5 Section 4: Aggregation

4.1. Aggregating results for individual time periods.

4.1.1. In many cases, it may be desired to report results at an aggregated time scale (e.g. annual energy savings), rather than for specific time periods. This may be done by simply adding the period specific results.

$$AEU_{total,P} = \sum_{p=1}^{P}(AEU_{p,i})$$

4.1.2. CalTRACK does not explicitly support annualizing results. For example, if avoided energy use $AEU_{total,8}$ is calculated for 8 monthly periods, then the annual savings cannot be estimated as $AEU_{total,12} = (AEU_{total,8} * \frac{12}{8})$, as this may yield biased estimates for interventions that yield seasonal savings. However, at the discretion of the procurer, such values may be used for intermediate reporting.

4.2. Aggregating multiple site-level results.

4.2.1. Multiple site-level results may be aggregated by adding metered savings that occurred during the same time periods.

$$AEU_{p,S} = \sum_{i=1}^{S}(AEU_{p,i})$$

4.3. Portfolio Uncertainty.

4.3.1. CalTRACK recommends approaching the uncertainty in avoided energy use on a case-by-case basis, depending on the objectives of the procurement or program.

4.3.2. Portfolio use case.

4.3.2.1. For use cases where confidence in portfolio-level performance is required (e.g. aggregator-driven pay-for-performance, non-wires alternatives (NWA) procurements), a building-level Coefficient of Variation of the Root Mean Squared Error (CV(RMSE)) threshold of 100% is recommended as a default, but this requirement may be waived at the discretion of the procurer.

4.3.2.2. CV(RMSE) is calculated as follows:

$$CV(RMSE) = \frac{\sqrt{\frac{\sum_{p=1}^{P}(U_p - \hat{U_p})^2}{P-c}}}{\overline{U}}$$

4.3.2.2.1. U_p is the total measured energy use during period p.

4.3.2.2.2. $\hat{U}$ is the predicted energy use during period p.

4.3.2.2.3. $\overline{U}$ is the mean energy use during the baseline period.

4.3.2.2.4. P is the total number of periods (e.g. days or billing periods) in the baseline used to estimate the model.

4.3.2.2.5. c is the number of explanatory variables in the baseline model (not including the intercept).

4.3.2.3. The portfolio-level fractional savings uncertainty (FSU) should be reported when using portfolio aggregation. Fractional savings uncertainty thresholds may be set by the procurer depending on the use case. For example, an NWA procurement may require less than 15% uncertainty, while a pay-for-performance program may require 25%. An alternative approach could use a discount rate based on the uncertainty of a portfolio.

4.3.2.4. Site-level FSU is calculated as follows using a modified version of the ASHRAE Guideline 14 formulation:

$$FSU_i = \frac{\Delta U_{save,Qi}}{U_{save,Qi}} = \frac{t(aM^2 + bM + d)CV(RMSE) * \sqrt{\frac{P}{P'}(1 + \frac{2}{P'})\frac{1}{Q}})}{F}$$

Where:

4.3.2.4.1. FSU_i is the fractional savings uncertainty in the baseline model predictions for the reporting period (this is also the fractional savings uncertainty of the avoided energy use, assuming that the metered consumption is accurate).

4.3.2.4.2. t is the t-statistic, which is a function of the required confidence level (usually 90%) and the degrees of freedom of the baseline model $(P - c)$.

4.3.2.4.3. M is the number of months in the reporting period.

4.3.2.4.4. Q is the number of periods (e.g. days or billing periods) in the reporting period.

4.3.2.4.5. F is the savings fraction, defined as the energy savings during q periods in the reporting period divided by the predicted baseline usage during that same period:

$$\frac{U_{save,Qi}}{U_{baseline,Qi}}$$

4.3.2.4.6. a, b and d are empirical coefficients proposed by Sun and Baltazar [2013] to handle problems with autocorrelated residuals in time series energy use data.

For billing data, $a = -0.00022, b = 0.03306, d = 0.94054$.

For daily data, $a = -0.00024, b = 0.03535, d = 1.00286$.

4.3.2.5. Site-level FSU from multiple projects can be aggregated to portfolio-level FSU as follows:

$$FSU_{portfolio} = \frac{\sqrt{\sum_{i=1}^{N}(\Delta U_{save,Qi})^2}}{\sum_{i=1}^{N} U_{save,Qi}}$$

4.3.2.6. Bias. While aggregation can dramatically reduce portfolio-level savings uncertainty, it does not eliminate inherent systemic biases due to the use of non-linear models, implementation variance, imbalanced application of non-routine adjustments, unaccounted for independent variables, or population trends.

4.3.2.6.1. Portfolio-level bias from modeling should be reported using the fractional bias error defined as follows:

Mean bias for a single site:

$$MB_i = \frac{1}{P}\sum_{p=1}^{P}(y_p - \widehat{y_p})$$

Portfolio-level bias error expressed as a percent of portfolio savings:

$$FBE_{portfolio} = \frac{\sum_{i=1}^{N}(MB_i)^2}{\sum_{i=1}^{N} U_{save,Qi}}$$

CHAPTER 2

Technical Appendix

2.1 2.2.1.2. For fitting baseline models using the hourly methods, no minimum baseline period length is required.

The baseline period for hourly methods is not set according to a particular time period – one year, for example – but is instead defined as sufficient when the full range of independent variables are observed. This is referred to as "data coverage" in Hourly Methods Documentation. This process is described in greater detail in LBNL's video on Time-of-Week Temperature models.

2.2 2.2.4.1. Temperature data may not be missing for more than six consecutive hours.

The decision to linearly interpolate up to 6 consecutive missing hours of weather data is adapted from Mathieu et al's Quantifying Changes in Building Electricity Use, with Application to Demand Response (Section 2.2).

2.3 2.4.1. Weather station to be used is closest within climate zone that meets CalTrack data sufficiency requirements.

2.3.1 Test 1: Weather station accuracy

Github issue: https://github.com/CalTRACK-2/caltrack/issues/65

Background:

Two weather station mapping methods were considered for CalTRACK 2.0's data sufficiency requirements. Each proposed weather mapping method is described below:

Method A:

1. Determine the candidate weather stations.
2. Determine the climate zone standards. The following climate zone standards were considered: a. IECC Climate Zone b. IECC Moisture Regime c. Building America Climate Zones d. CEC California Building Climate Zone Areas
3. Establish the site's climate zone inclusion for each considered climate zone standard.
4. Determine each site's closest candidate weather station.
5. Establish the weather station's climate zone inclusion for each considered climate zone standard.
6. Reject the candidate weather station if any of the following are true: a. The candidate station is more than 150 km from the site. b. The candidate weather station's climate zone does not match the site's climate zone for all considered climate zone standards. c. The candidate station does not have sufficient data quality when matched with site meter data.
7. If criteria in (6) are met, use the candidate weather station. If criteria are not met, move back to step 3 and test the next closest candidate weather station.
8. If no stations meet the criteria above, use method B.

Method B:

1. Determine the candidate weather station to be considered.
2. Determine the (next) closest weather station.
3. Reject the candidate station if it does not have sufficient data quality when matched with site meter data
4. If no stations meet the criteria above, consider the site unmatched.

Data:

Hourly temperature data from all California weather stations from January 2014 to January 2018. Tested parameters: The weather stations selected based on Method A and Method B were compared against the "ground-truth" weather stations established by a similarity metric.

Testing methodology:

It was problematic to empirically test weather mapping methods on buildings because true temperature values were not available at each site. Although there was no accurate data at the site level, there was accurate data at the location of each weather station. To compare mapping methods, each weather station was considered as a site and weather station selection methods were tested on the weather stations.

The results were compared against a "ground-truth" of weather similarity between weather stations. The "ground-truth" ranking of weather station similarity was determined by a metric that ranked the similarity of a weather station to other weather stations. The similarity metric was constructed with three distance metrics: 1) root mean squared error, 2) kilometers, and 3) cosine distance to the weather station of interest. Weighted equally, these rankings were combined to provide a similarity ranking of all other weather stations for each weather station.

These rankings were the "ground-truth" that compared Methods A and B. After weather stations were selected based on Methods A and B, their accuracy was compared to the "ground-truth" ranking defined by the similarity metric.

Results:

Our results show that Method A produced the best possible weather station 56% of the time and Method B produced the best possible weather station 53% of the time. This indicates that Method A is more accurate than Method B by a small margin.

2.3.2 Test 2: Importance of weather data

Github issue: https://github.com/CalTRACK-2/caltrack/issues/65

Background:

CalTRACK 2.0 methods were designed to be simple and universally applicable. We empirically tested the effect of inaccurate weather data on our model prediction error to determine the importance of weather data accuracy.

Data:

Weather data was collected from two weather stations in each of the 50 states in the United States.

Tested parameters:

Model fit, measured by CVRMSE, was calculated with CalTRACK methods using increasingly inaccurate temperature data.

Testing methodology:

The models from CalTRACK 1.0 were fit with weather data collected from two weather stations in each of the 50 states in the United States. The weather values provided by the geographically and climatically diverse weather stations were largely inaccurate for the buildings observed. This provided an opportunity to analyze the effect of inaccurate temperature data on model fit.

Results:

In the figure below, in-sample prediction error slightly increased as temperature data error increased. These results suggest that small inaccuracies in the weather data have a small effect on model prediction error.

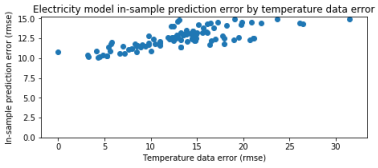

Fig. 1: *Figure: Model error vs. temperature difference from nearest weather station*

Conclusion:

CalTRACK 2.0 will employ Method A as the preferred weather station mapping method, despite the implication that weather station mapping misalignment leads to only a minimal increase in model error. If Method A is impractical or impossible, Method B is a suitable alternative.

2.4 3.1.3. Models are fit to baseline data in the 365 days immediately prior to the intervention start date.

Github issue: https://github.com/CalTRACK-2/caltrack/issues/68

Background:

The length of the baseline period in energy savings models may affect energy savings calculations in two ways:

1. Periods that are too short may not capture the full range of input conditions, such as weather or occupancy patterns, that are typically experienced by a building.
2. Periods that are too long increase the chances of unexpected changes in a building's energy use. For example, energy efficient equipment unrelated to the intervention is more likely to be added during longer baseline periods. This will affect estimated energy savings.

CalTRACK methods adopt the Uniform Methods Project's (UMP) minimum baseline period of 365 days (see UMP guidelines in 6.4.1 Analysis Data Preparation). The obvious justification for a 365 day baseline is the value of fitting a model over a wide range of possible temperatures. Hourly methods may require a different baseline length assumption. The UMP does not provide guidance for the maximum length of the baseline period, so empirical testing was conducted to determine the optimal maximum baseline period length.

Data:

Billing period data from approximately 1000 residential buildings in Oregon and daily data from 1000 residential buildings in California.

Tested parameters:

The effect of increasing the length of the baseline period on prediction error.

Testing methodology:

The CalTRACK methods were applied to the full dataset five times using baseline periods of 12, 15, 18, 21, and 24 months for each iteration. The length of the baseline period was the only change between iterations.

Testing was conducted as follows:

1. CalTRACK models were fit to each of the candidate baseline periods.
2. Total energy consumption predictions for each baseline model we calculated for a 12-month reporting period.
3. The CVRMSE and NMBE error metrics were calculated for these predictions.
4. This test was conducted separately for daily and monthly data.

Acceptance criteria:

Energy consumption trends and error metrics were compared for different baseline period lengths. The baseline period length that did not inflate out-of-sample errors was recommended as the maximum baseline period length in CalTRACK methods.

Results:

The figure below shows that baseline normalized annual consumption (NAC) increases as the baseline period length increases. This implies that using baseline periods longer than 12 months may unjustifiably inflate estimated savings.

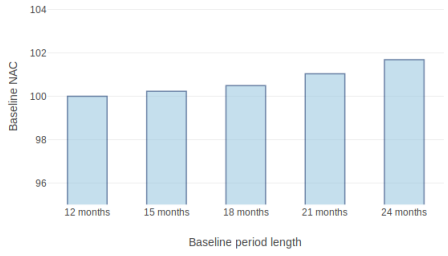

Fig. 2: *Figure: Effect of baseline period length on normalized annual consumption using daily data*

The figure below demonstrates that increasing baseline period length worsened model fit. This may occur because increased baseline periods are more likely to include non-routine events that affect energy use in unpredictable ways.

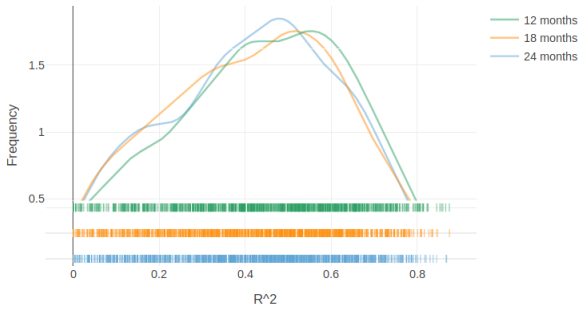

Fig. 3: *Figure: Effect of baseline period length on model R-squared distribution*

Conclusion:

The results from these empirical results indicate that baseline periods longer than 1 year may have increased baseline energy consumption and poorer model fit than the minimum 12-month baseline. We recommend a maximum baseline period length of 12 months for both billing and daily models.

2.5 3.1.4.1. Select and qualify balance points for candidate models for each period for each meter.

Github issue: https://github.com/CalTRACK-2/caltrack/issues/69

Background:

CalTRACK 1.0 methods recommends that fixed balance point temperatures are used for degree-day covariates in billing period methods. The UMP recommends fixed balance point temperatures 60 F for heating degree days and 70 F for cooling degree days for billing period methods.

For daily methods, the UMP recommends that variable balance points are used for degree-day covariates.

It is possible that billing period models will have improved model fit with variable balance points instead of the fixed balance points suggested by the UMP.

Data:

Electricity and gas billing data from approximately 1000 residential buildings that had undergone home performance improvements in Oregon.

Tested parameters:

The R-squared of a billing period model using fixed balance points of 60 F for HDD and 70 F for CDD was compared to a model with variable balance point ranges of 40-80 for HDD and 50-90 for CDD.

Testing methodology:

1. CalTRACK billing period models were fit to baseline period usage data with fixed balance point temperatures.
2. The fitting process was repeated with a grid search for the balance point temperatures with a grid search range of 40-70 F for heating degree days and 60-90 F for cooling degree days. A 3 F search increment was used to determine variable balance point temperatures.

3. The error metrics of CVRMSE and NMBE were calculated for each model with fixed and variable balance points.

Acceptance criteria:

Variable balance point temperatures for billing period models were accepted into the CalTRACK 2.0 specification if the variable balance point models did not cause the average model performance to deteriorate. The average model performance did not deteriorate if average model fit improved or a paired t-test of model fit metrics showed no significant difference.

Results:

For the 1077 buildings tested with fixed balance points, 479 were fit using intercept-only models. When the same 1077 buildings were tested with variable balance points, there were 357 intercept-only models. These results indicate that the weather-sensitivity of some buildings was not being modelled with fixed balance point temperatures.

The performance of the remaining weather-sensitive buildings were compared when fit with fixed and variable balance point temperatures. The mean R-squared for fixed balance point models was 0.480, while the mean R-squared for variable balance point models was 0.495. The figure below shows slight improvements in model fit for variable balance point models.

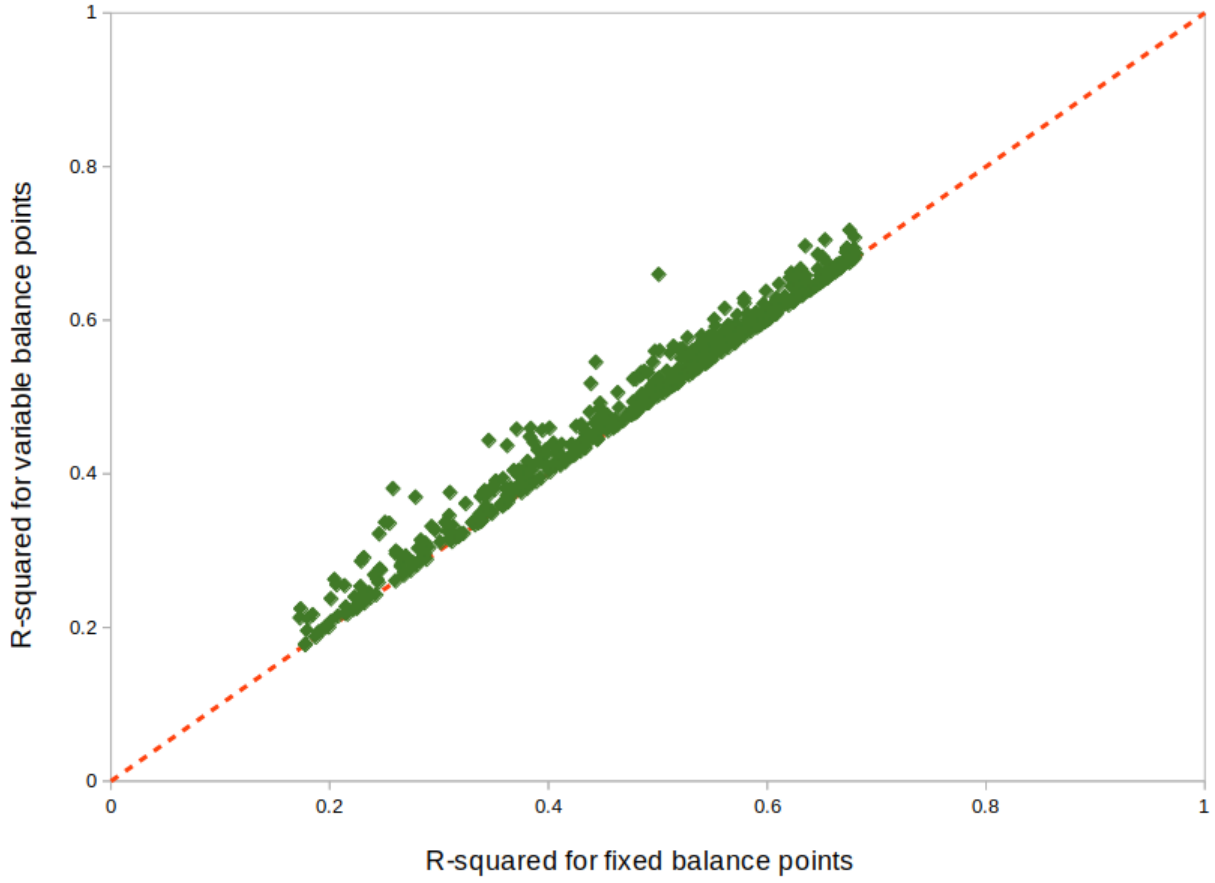

Fig. 4: *Figure: R-squared for fixed and variable balance points*

Conclusion:

Our empirical results indicate that variable balance points generated fewer intercept-only models and had a slightly

improved R-squared than fixed balance points in billing period methods. Therefore, we recommend using variable balance points in billing period methods.

2.6 3.2.1. A grid search of models is performed using a wide range of candidate balance points.

Github issue: https://github.com/CalTRACK-2/caltrack/issues/72

2.6.1 Test 1: Distribution of selected balance points with different grid search ranges

Background:

Daily and Billing Period methods in CalTRACK 2.0 use variable degree-day regression to model baseline and reporting period energy consumption. In variable degree-day regression, the analyst must establish a search range that contains the optimal balance point temperatures for each degree-day covariate. Excessively large search ranges have high computation requirements. However, overly constrained grid search ranges may lead to suboptimal balance point temperatures and poor model fit. The testing protocol below was used to define the optimal grid search ranges for HDD and CDD covariates.

Data:

Billing period data from approximately 1000 residential buildings in Oregon and daily data from 1000 residential buildings in California.

Tested parameters:

HDD and CDD balance points were calculated with different grid search ranges using Caltrack methods.

Testing methodology:

Caltrack models were fit to the Oregon building usage dataset using 4 balance point search ranges:

1. 10-degree range: 55-65 F HDD and 65-75 F CDD
2. 20-degree range: 45-65 F HDD and 65-85 F CDD
3. 30-degree range: 40-70 F HDD and 60-90 F CDD
4. 40-degree range: 40-80 F HDD and 50-90 F CDD

Results:

The bar chart below shows the distribution of best-fit HDD balance points for three of the four tested grid search ranges. These results show that when the grid search is constrained, models tend to select balance points at the end of the grid search range. For the 10-degree grid search range, almost 30% of the buildings have an HDD balance point of exactly 65 F. But when the grid search range is expanded to the 30 F or 40 F ranges, the distribution of best-fit balance points tends towards a Gaussian distribution centered around 63 F. These results indicate that overly constrained grid search ranges may result in suboptimal best-fit balance points and, thereby, suboptimal model fits.

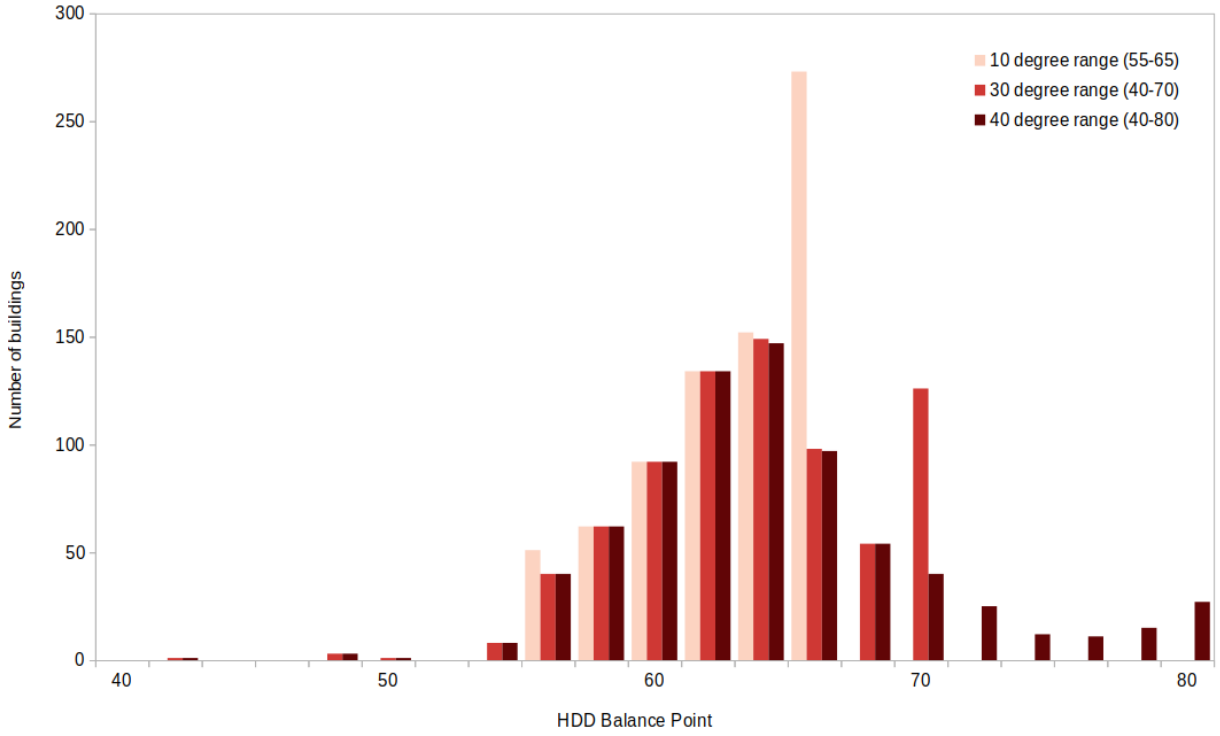

Fig. 5: *Figure: HDD balance point frequency by grid search range*

2.6.2 Test 2: Importance of optimal balance points on estimated savings

Github issue: https://github.com/CalTRACK-2/caltrack/issues/72

Background:

The robustness of estimated savings to different balance point ranges provides more confidence in our estimated energy savings calculations.

Data:

Billing period data from 1000 residential buildings in Oregon and daily data from 1000 residential buildings in California.

Tested parameters:

CalTRACK methods estimated energy savings for all program participants for five different grid search ranges.

Testing methodology:

Baseline models were fit to the full set of program participants five times, varying the search range for the HDD balance point and keeping all other parameters constant. The annualized estimated savings were calculated for each grid search range.

Results:

The box plots below show that estimated energy savings were similar across different grid search ranges. This indicates that estimated savings with CalTRACK methods are robust to varying grid search ranges.

Conclusion:

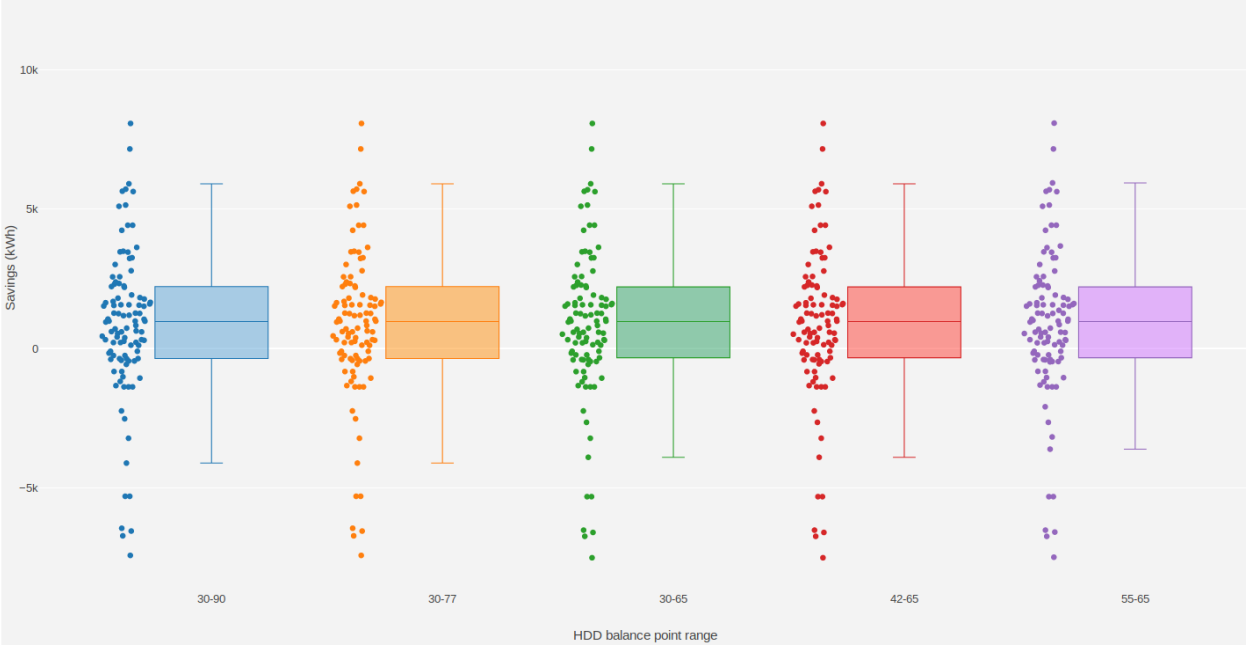

Fig. 6: *Figure: Estimated savings with different grid search ranges*

Expand balance point search range to 30-90 F for heating balance points and 30-90 F for cooling balance points.

2.7 3.2.3. Maximum gap between candidate balance points in the grid search is 3 degrees F or the equivalent in degrees C.

Github issue: https://github.com/CalTRACK-2/caltrack/issues/72

Background:

The grid search algorithm selects balance points by, first, estimating a model with each set of candidate HDD and CDD balance points and, second, choosing the balance points that generate the best-fit model.

The analyst determines the search increments, or "steps", that the algorithm uses to choose models that are tested for the optimal balance point. Small search increments, such as 1 degree, estimate a model for each degree in the HDD and CDD grid search ranges. This is computationally intensive. Larger search increments have lower computational demands, but could provide less accurate balance points temperatures.

Data:

Billing period data from 1000 residential buildings in Oregon and daily data from 1000 residential buildings in California.

Tested parameters:

The selected balance point temperature with search increments of 1, 2, 3, and 4 degrees.

Testing methodology:

CalTRACK methods were used to estimate models with balance point temperatures selected with search increments of 1, 2, 3, and 4 degrees. The results were compared to determine the optimal search increment.

Results:

The empirical results show that model fit did not change significantly when balance points were off by 1 F. This implies that a search increment of 3 F is acceptable because the optimal balance point temperature can only be 1 F above or below the optimal balance point with a 3 F search increment.

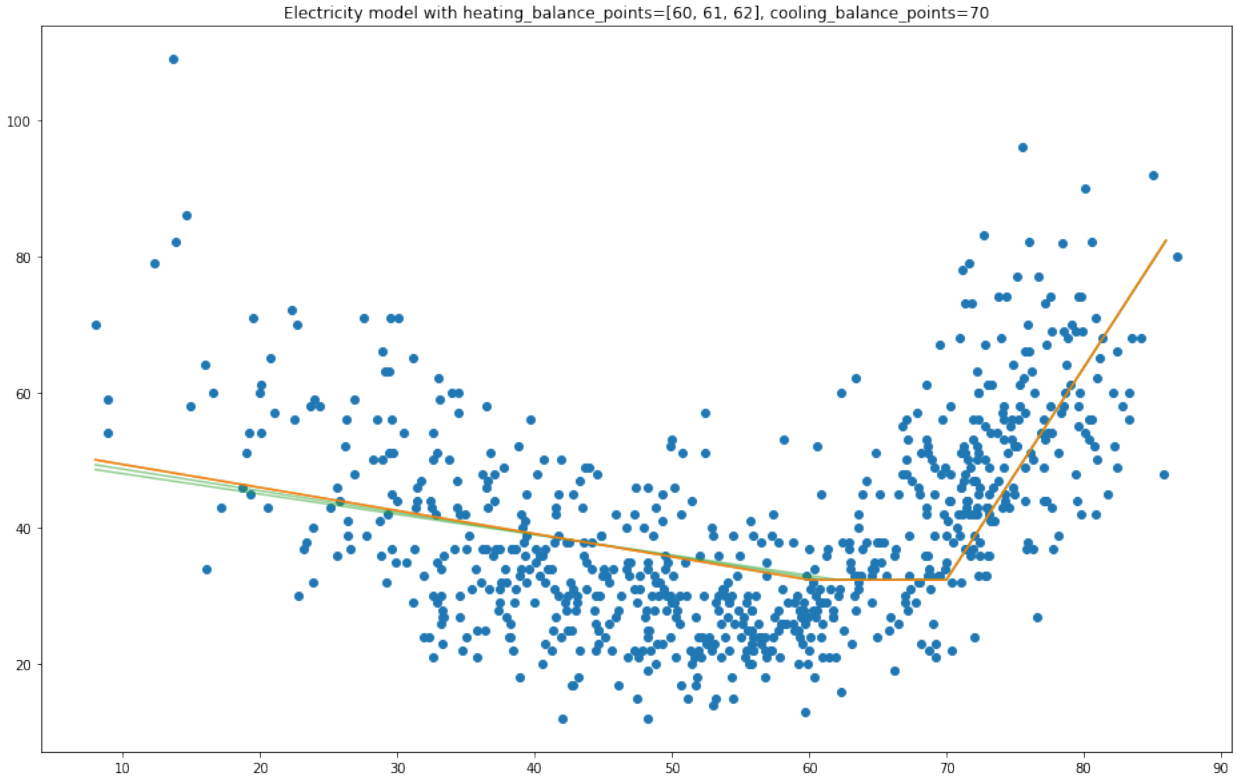

Fig. 7: *Figure: HDD balance points with different search increments*

Conclusion:

CalTRACK's Billing Period and Daily methods will use a 3 F search increment in the grid search algorithm.

2.8 3.3.1.2. Independent variables

2.8.1 Test 1: Calendar effects and error structure

Github issue: https://github.com/CalTRACK-2/caltrack/issues/57

Background:

CalTRACK models are specified with only HDD and CDD covariates. However, there are a priori reasons to expect that energy consumption could be correlated with calendar effects, such as day-of-week, day-of-month, month-of-year, or holidays. If calendar effects are significantly correlated with energy consumption and excluded from the model, it may cause less accurate energy savings estimates with poorer model fit. .

Calendar effects can be added to the model as categorical variables for day-of-week, day-of-month, month-of-year, or holidays. Including these variables will control for variation in building-level energy consumption that is correlated with each respective calendar effect. If calendar effects variables have significant explanatory power for building-level energy consumption, including them will improve the accuracy of our energy savings estimates and model fit. However, the introduction of calendar effects complicates our model and demands additional data sufficiency requirements. The following test was conducted to determine which, if any, calendar effects should be included in CalTRACK model specifications.

Data:

A 100-home sample with temperature and AMI electricity data.

Tested parameters:

The error structure of models with respect to temperature, day-of-week, day-of-month, month-of-year, and holidays were examined to detect non-stationary structures in the residuals.

Testing methodology:

For each of the 100 buildings, daily usage was normalized by dividing all the daily values for each building by the mean energy consumption for that building. The HDD and CDD variables were defined by fixed balance points.

CalTRACK methods estimated models for each building in the sample, which generated normalized residuals for each of the 100 buildings in the sample. The error structure of models with respect to temperature, day-of-week, day-of-month, month-of-year, and holidays were examined.

Acceptance criteria:

If significant normalized error structure is not observable for any of the proposed calendar effects, the model specification with only HDD and CDD covariates is sufficient.

Results:

The average residuals vs. temperature graph indicates that there was not a strong trend in the error structure at given temperatures in the data.

The average residuals vs. day-of-month graph did not show strong correlation between residuals and a particular day of the month.

However, there did appear to be correlation between month-of-year and day-of-week with average residuals. The average residuals vs. month-of-year residuals graph shows positive and large residuals in the June, July, August, and December. The HDD and CDD covariates included in CalTRACK models control for temperature, which means this residual trend was not a first-order temperature effect. A possible explanation is that the high residual months coincided with months when school was not in session. School vacations likely result in higher household occupation and, thereby, higher energy consumption. This supports the inclusion of a month-of-day category variable.

The average residuals vs. day-of-week graph shows a trend of large, positive residuals during the weekend days. It is reasonable to expect higher energy consumption during weekends because residents are more likely to be occupying the house. This supports the inclusion of a day-of-week category variable.

Finally, the average residuals were 0.061 for holidays and -0.001 for non-holidays. This indicates that energy consumption was higher during holidays than non-holidays for residential customers. This also supports the inclusion of a holiday indicator variable.

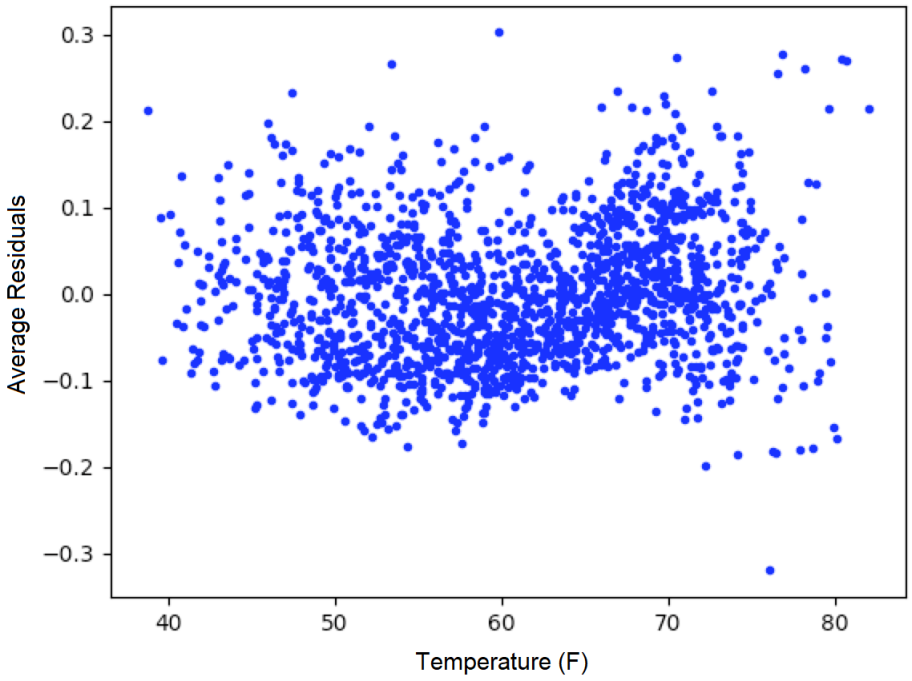

Fig. 8: *Figure: Average residuals vs. temperature*

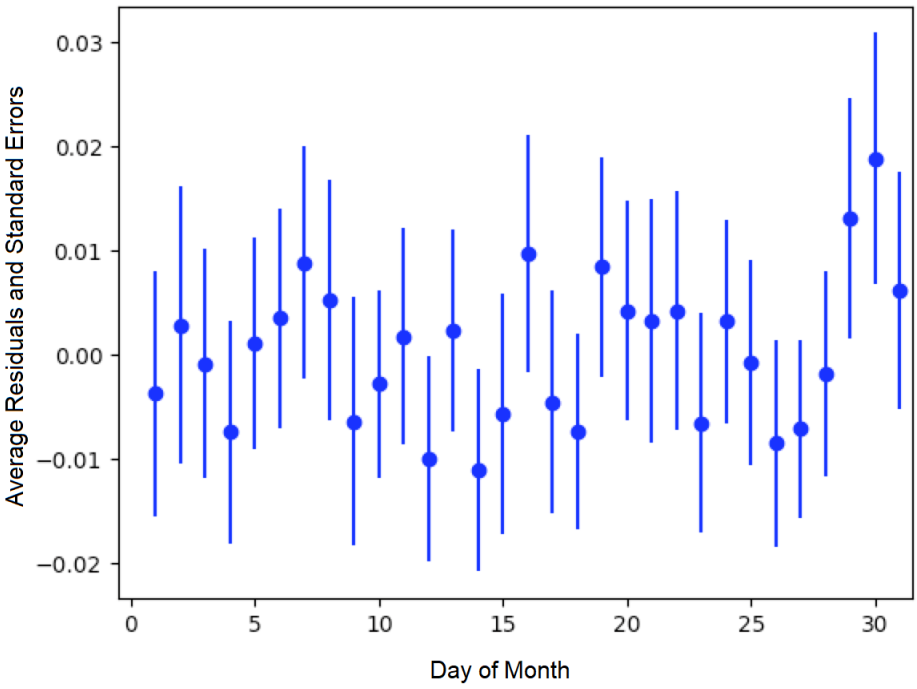

Fig. 9: *Figure: Average residuals vs. day-of-month*

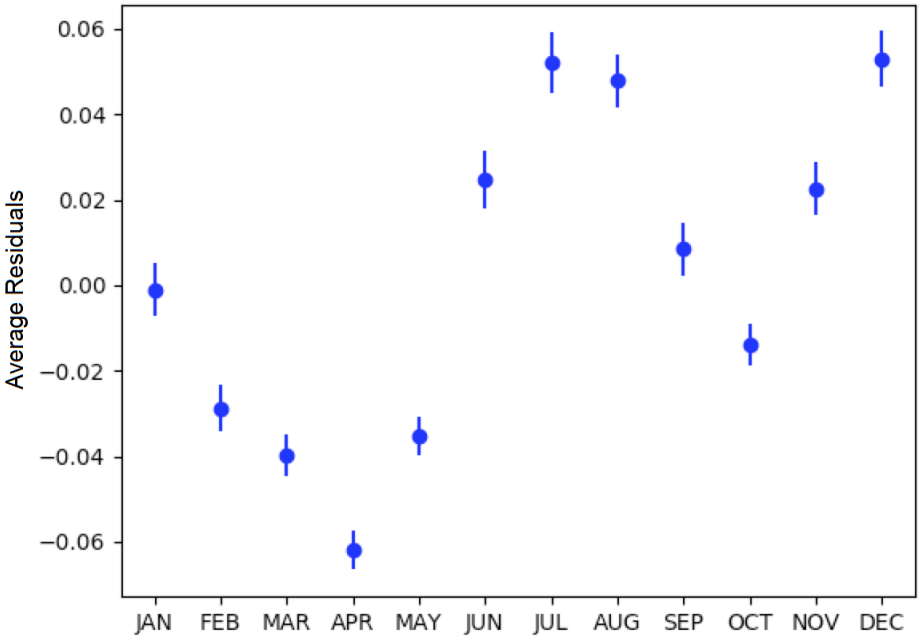

Fig. 10: *Figure: Average residuals vs. month-of-year*

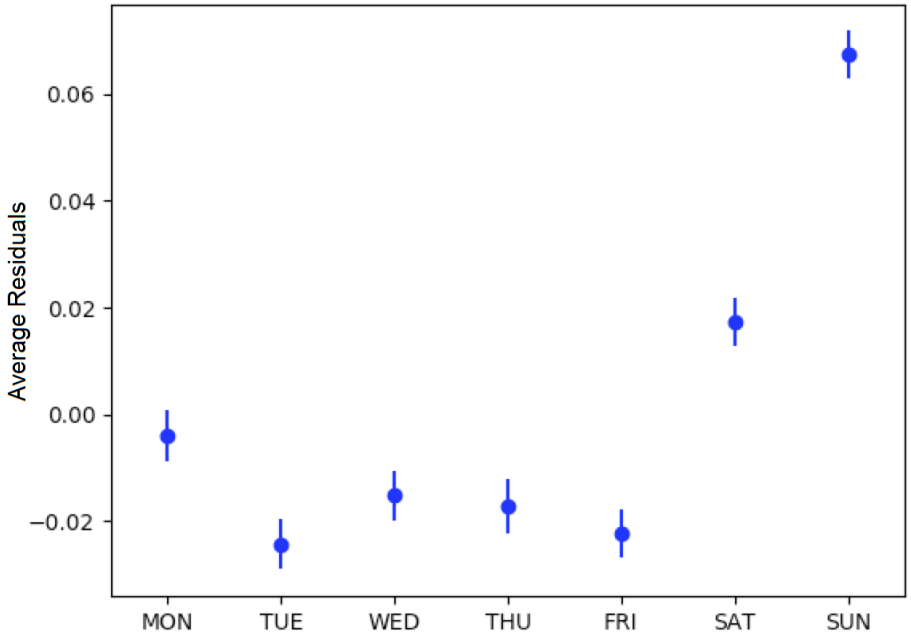

Fig. 11: *Figure: Average residuals vs. day-of-week*

2.8.2 Test 2: Calendar effects and aggregated energy savings

Github issue: https://github.com/CalTRACK-2/caltrack/issues/57

Background:

Previous analysis showed temporal patterns in the residual structure for daily and billing period models, which suggests that calendar effects should be included in the billing period and daily CalTRACK model specifications. However, at the time of this testing, CalTRACK methods were designed to calculate only annual savings by aggregating building-level savings estimates over a year. For this reason, it is only necessary to include calendar effects in CalTRACK model specifications if they have a significant effect on annual savings calculations. The effect of adding calendar effects on annual energy savings was empirically tested as follows.

Tested parameters:

The empirical testing compared different model specifications with the following metrics:

1. CVRMSE (Coefficient of Variation Root Mean Squared Error)
2. NMBE (Normalized Mean Bias Error)

These are labels for the tested model specifications:

1. **M0:** CalTRACK model with only HDD and CDD covariates
2. **M0.1:** M0 but with a wide range of possible CDD/HDD balance points.
3. **M0.2:** M0 but with robust regression and using the Huber loss function.
4. **M1:** M0 plus a categorical day-of-week variable.
5. **M2:** M0 plus a categorical variable distinguishing weekdays versus weekend days only.
6. **M3:** M0 plus categorical day-of-week plus categorical month-of-year.
7. **M4:** M1 with elastic net regularization (L1 = 0.5, L2 = 0.5).
8. **M5:** M0.2 plus a categorical variable distinguishing weekdays versus weekend days only.

Testing methodology:

The testing methodology used out-of-sample data to estimate prediction error associated with each model specification of interest. Models with lower CVRMSE and a NMBE closer to 0 were preferred.

CalTRACK's objectives prioritize simplicity in model specification decisions, so less complex model specifications were desired if they did not significantly detract from the quality of aggregated annual energy savings calculations.

Results:

The graph below presents the median CVRMSE and NMBE for each model specification. The results indicate that model M1 and M2, which included day-of-week and weekday or weekend indicator variables, respectively, generated only slightly lower CVRMSE and NMBE than the M0.1 model.

Additionally, it is clear that the two models with robust regression specifications generated much lower NMBE than the non-robust regression results. This is an important finding and a further discussion of robust regression is found in appendix 3.4.1.

Conclusion:

Although adding calendar effects may be significant when estimating daily models, their effect is reduced when aggregated over a year. The small reductions in CVRMSE and NMBE gained by adding calendar effects do not justify their added complexity to CalTRACK models.

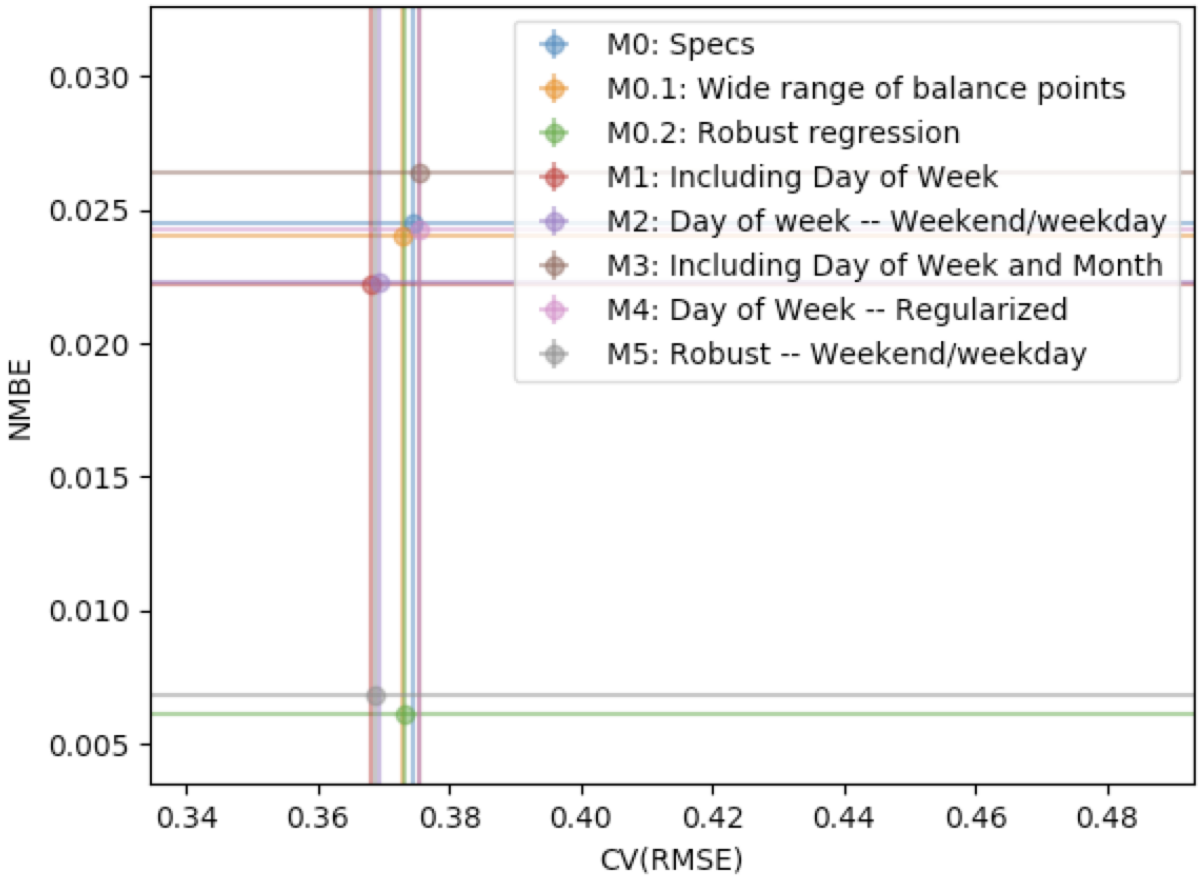

Fig. 12: *Figure: CVRMSE vs. NMBE*

2.9 3.4.1. Models using daily data are fit using ordinary least squares.

Github issue: https://github.com/CalTRACK-2/caltrack/issues/57

Background:

In the context of home energy modelling, robust regressions tend to generate a normalized mean bias error (NMBE) that is closer to zero than ordinary least squares (OLS). This trend is apparent in the figure below, which presents the distribution of NMBE for robust regression and OLS with no calendar effects. The results show that OLS tended to predict higher energy usage, while robust regression tended to predict lower energy usage.

Robust regressions are computationally intensive and may be difficult to replicate across statistical software packages and CalTRACK methods value simplicity and replicability. This makes OLS preferable unless robust regression provides significantly better results than OLS.

Tested parameters:

The NMBE for CalTRACK models with robust regression and OLS were compared. Additionally, the computation requirements for robust regression and OLS were recorded and compared to further inform the decision.

Testing methodology:

Robust and OLS regressions were estimated and their NMBE was calculated. The computational requirements for each of these methods was also analyzed.

Results:

The empirical results below show that robust regression generates NMBE that are slightly closer to zero than OLS. The OLS models tend to have positive NMBE, while the robust regression has negative NMBE. Although the model fit was slightly better for robust regression, the robust regression took nearly 3 times longer to calculate than OLS. These were significant additional computation requirements for robust regression.

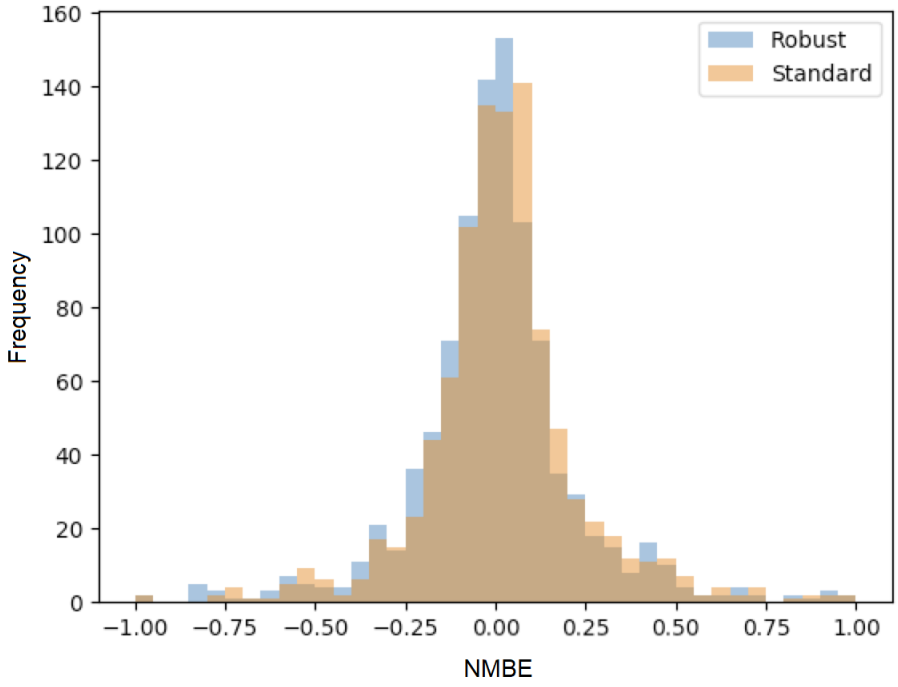

Fig. 13: *Figure: Distribution of NMBE for robust regression and ordinary least squares*

Conclusion:

CalTRACK recommends using an OLS modelling approach instead of robust regression. The additional computation requirements and difficulties replicating robust regression across statistical packages are not justified by the small improvements to model fit from robust regression.

2.10 3.4.3.2. Candidate model qualification.

Github issue: https://github.com/CalTRACK-2/caltrack/issues/76

Background:

After HDD + CDD, HDD-only, CDD-only, and intercept-only model candidates are estimated, the best-fit model is selected through a two-stage process.

First, for each model specifications with covariates (HDD + CDD, HDD-only, and CDD-only), the estimated model's HDD and CDD covariates that are statistically insignificant (p-value > .10) are removed from consideration. This is referred to as the p-value screen.

Second, the remaining model with the highest R-squared value is selected.

The p-value screen is conducted because statistically insignificant coefficients may lead to poor out-of-sample prediction. However, this procedure may eliminate best-fit candidate models. Additionally, models with estimates for HDD and CDD have more interpretation value than models with only HDD, CDD, or an intercept.

The effect of the p-value screen on out-of-sample error was analyzed to determine if the p-value screen improved model selection.

Data:

Billing period data from approximately 1000 residential buildings in Oregon.

Tested parameters:

The out-of-sample prediction errors were calculated for models selected with and without the p-value criterion.

Testing methodology:

1. Caltrack monthly models were fit to the baseline period usage data using the p-value screen.
2. The fitting process was repeated without a p-value screen.
3. A comparison was performed using 24-month electric traces, split into 12 months of training and 12 months of test data. Mean absolute prediction error was used as the metric to compare performance.

Acceptance criteria:

This update was accepted if removing the p-value screen did not cause average model performance to deteriorate.

Results:

The graph below shows that the distribution of selected model types for the 1,000 building sample changed slightly when models were selected without a p-value screen. In almost 90% of the buildings, the fit did not change when the p-value criterion was removed. For the remainder that did change, the model shifted from an intercept-only model to a weather-sensitive model.

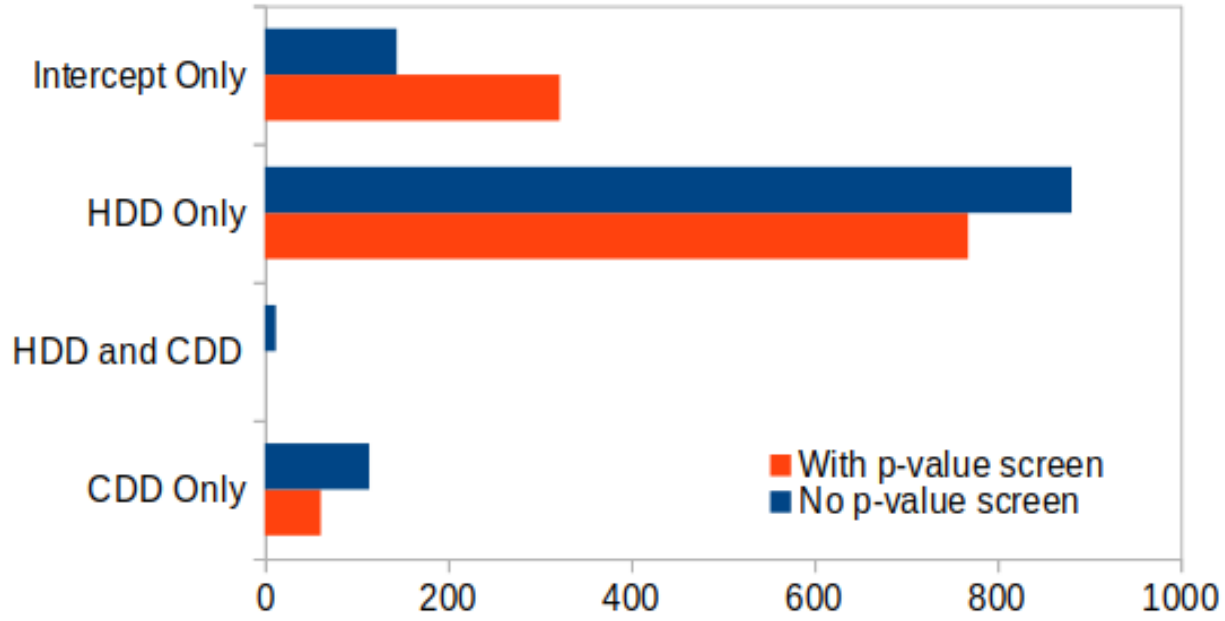

Fig. 14: *Figure: Best-fit model with p-value screen*

The out-of-sample prediction for models selected with and without the p-value screen is graphed below. The average Mean Absolute Error (MAE) was 8.20 when the p-value screen was removed and 8.34 with the p-value screen. This indicates a slight improvement in prediction error when the p-value screen was eliminated. Additionally, over two times more models had improved model fit when the p-value screen was eliminated. Of the models that degraded when the p-value screen was eliminated, none of the degradations were catastrophic.

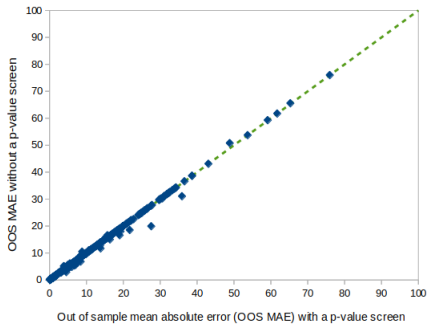

Fig. 15: *Figure: Mean absolute error with and without p-value screen*

Conclusion:

Our tests indicate that requiring a P-value screen was superfluous at best and marginally counterproductive at worst. Therefore, the required P-value screen for candidate models will be removed from CalTRACK model selection criteria.

2.11 3.4.3.3. The model with highest adjusted R-squared will be selected as the final model.

Github issue: https://github.com/CalTRACK-2/caltrack/issues/62

Background:

CalTRACK grid search algorithm determines the optimal balance point temperatures by estimating a model for each HDD and CDD balance point in the grid search range and selecting the best-fit model. The definition of "best-fit" depends on the loss function. There are a variety of loss functions that are more or less suitable for different data structures and modelling methods.

The loss function candidates analyzed were:

1. Quadratic Loss Function (referred to in the figure below as "least squares")
2. Linear Loss Function (referred to in the figure below as "absolute value")
3. Huber Loss Function
4. Tukey's Bi-square Loss Function

The distributions of each candidate loss function are visualized in the graph below. Of the candidate loss functions, the "Least Squares" or Quadratic Loss Function is the most common. It evaluates "best-fit" by selecting the model that minimizes the sum of squared residuals, which will also result in the highest R-Squared model. The Quadratic Loss Function is not robust to outliers.

The Tukey Bi-Square Loss Function is the most robust to outliers, but generates larger model variance in the absence of outliers.

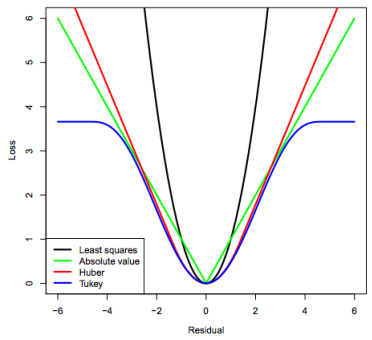

Fig. 16: *Figure: Loss function distributions*

Empirical testing was conducted to determine the loss function that resulted in the best-fit balance point temperature models.

Tested parameters:

The HDD balance points were analyzed for each candidate loss function. They were evaluated based on the standard deviations of their estimated HDD balance points.

Testing methodology:

From the 1,000 residential building data set, the 263 buildings that were best fit by an CDD + HDD model were filtered for closer examination. Excluded models either had no significant heating or cooling component or were intercept-only models.

For each of the 263 selected buildings, the balance point algorithm was conducted 25 times. Before each test run, 10% of days in the baseline period were randomly removed. The selected balance point temperatures should not change drastically when 10% of days are removed from the baseline period. If the balance points do show large deviations between test runs, then it is possible that outliers are driving the fit.

The algorithm for selecting balance points operated as follows:

1. Choose relevant balance points to test, which is determined by the grid search range and search increment
2. Run a linear regression using the HDD and CDD values for each candidate balance point combination. Qualifying models must have non-negative intercept, heating, and cooling coefficients, and the heating and cooling coefficients must be statistically significant.
3. Calculate the loss function.
4. Choose new balance points and repeat.
5. Select the balance points with the lowest loss function.

Our testing methodology used the balance point selection algorithm for each qualifying building 25 times with a different 10% of days removed from the baseline period. The mean and standard deviations for each home across the 25 test runs were calculated and recorded. This was repeated for each of the 4 candidate loss functions.

Results:

The frequency of selected balance points for each candidate loss function are shown in the graph below. It is apparent that the quadratic loss function was concentrated near the 65 F balance point. The high frequency of balance points at one temperature indicates that the quadratic loss function was not heavily skewed by outliers in this dataset.

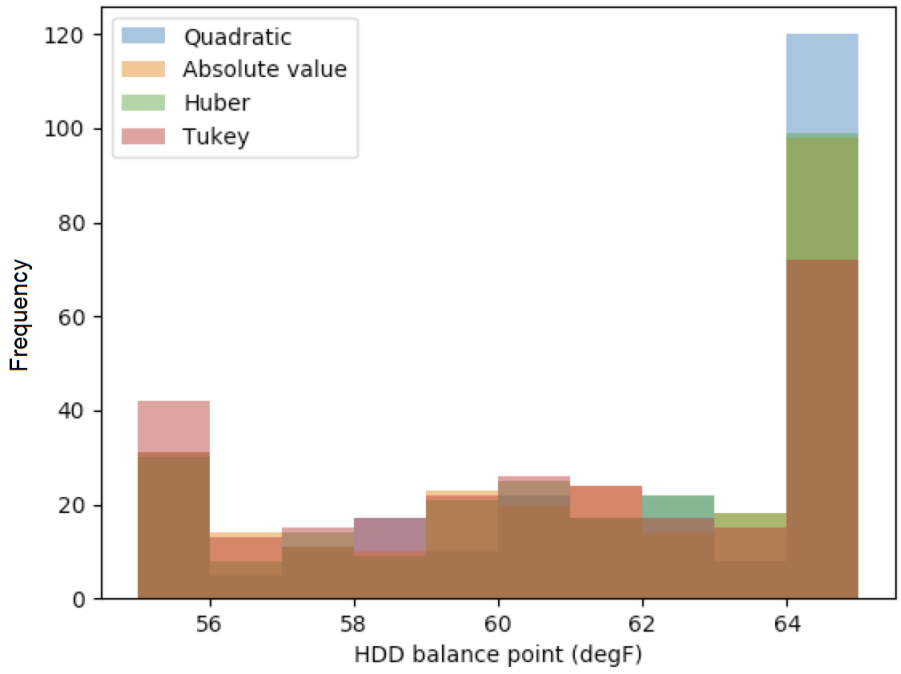

Fig. 17: *Figure: Frequency of balance point temperatures by loss function*

These results are confirmed by analyzing standard deviations of balance points for each loss function. The empirical results below indicate that the quadratic loss function is the most stable among candidate loss functions.

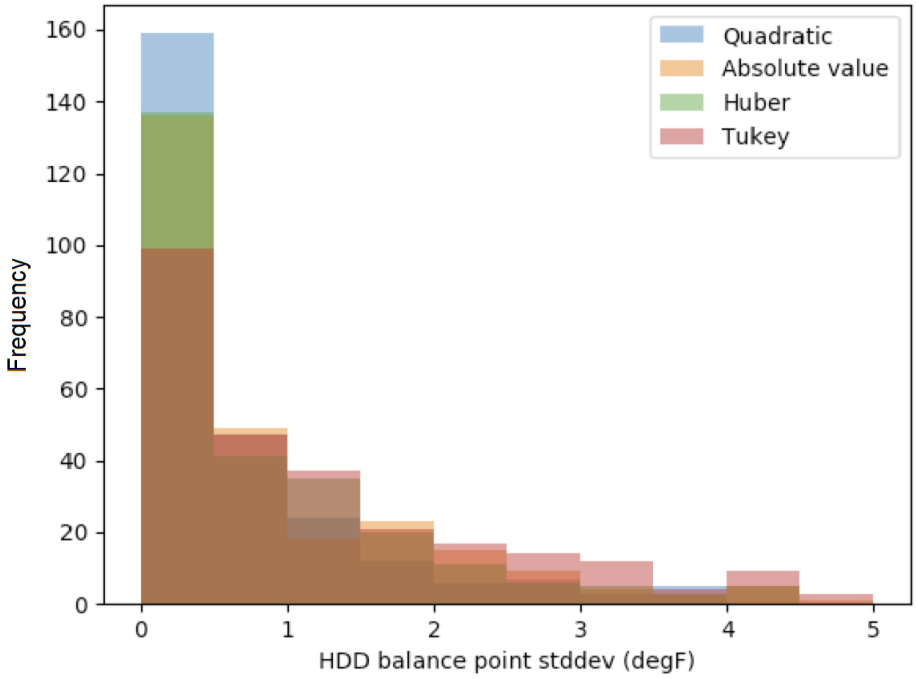

Fig. 18: *Figure: Balance point standard deviation by loss function*

The "Frequency of balance point temperatures by loss function" bar chart shows that the quadratic loss function produced balance point temperatures concentrated at 65 F. It is possible that the grid search range was overly constrained, which caused HDD temperatures above 65 F to bunch on the edge of the grid search range.

The figure below presents the standard deviations of each candidate loss function with balance point temperatures at either 55 F or 65 F removed from the sample. This eliminated HDD balance points that were bunched at either edge of the grid search range. The results still showed that the quadratic loss function generated balance points with the smallest standard deviations.

Figure: Balance point standard deviation by loss function (edges removed)

The quadratic loss function performed the best among candidate loss functions. The bar chart below shows the standard deviation of balance point temperatures across test runs evaluated with the quadratic loss function. These results indicate that 80% of the sample has balance point temperature standard deviations of less than 1 degree.

Conclusion:

Our results indicate that the quadratic loss function produced the most stable results across the test data set. Under the quadratic loss function, 80% of buildings had standard deviations of 1 F or less. This is the recommended loss function for CalTRACK methods.

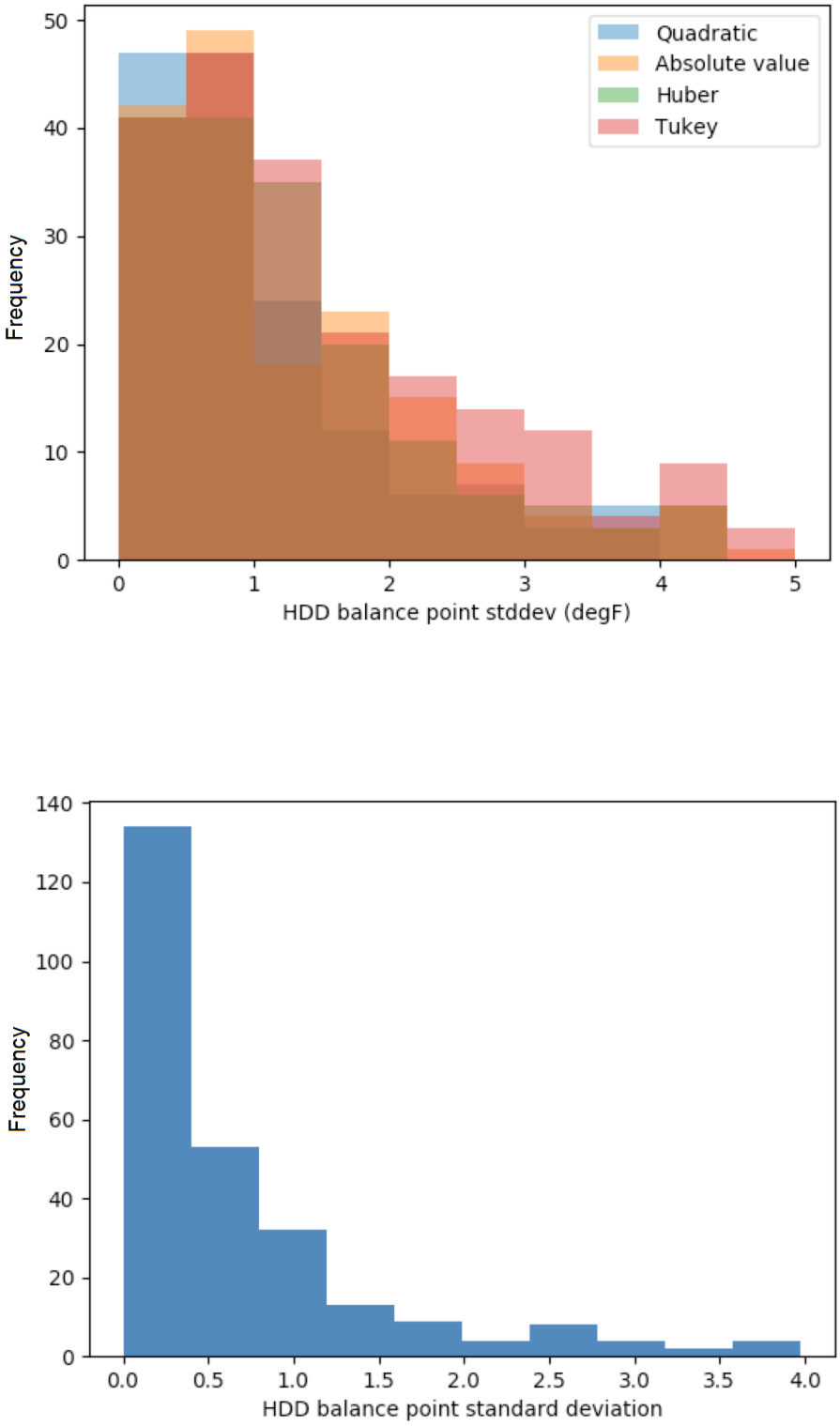

Fig. 19: *Figure: Quadratic loss function balance point standard deviations*

2.12 3.6.5. Baseline periods.

2.12.1 Test 1: Adequacy of single estimated regression for baseline period

Github issue: https://github.com/CalTRACK-2/caltrack/issues/103

Background:

The Time-Of-Week and Temperature model for hourly methods uses a single model to fit the entire baseline period, which could be up to 12 months long. The single, yearly regression approach assumes that baseline and weather-sensitive energy consumption is constant throughout the year.

However, it is possible that baseline and weather-sensitive energy consumption is not constant across the entire baseline period. For example, the baseline energy consumption may be higher during summer months than spring months because children do not spend daytime hours at school during the summer. If this is the case, then estimated parameters for a single, yearly regression will not reflect the true energy consumption during each month of the baseline period. This results in higher uncertainty of CalTRACK energy savings estimates.

Data:

Residential, daily electricity data from 80 buildings. Data was supplied by Home Energy Analytics.

Tested parameters:

The average CVRMSE from sampled buildings with different modelling approaches.

Testing methodology:

Energy savings for the 80 sampled households were calculated and compared when:

1. A single, yearly regression was estimated for the entire baseline period
2. 12 separate regressions were estimated for each month of the baseline period

The methods were evaluated based on the average CVRMSE of sampled households.

Results:

The results shows that CVRMSE was improved when regressions were estimated for each month of the baseline period. There were CVRMSE improvements of 33% and 19% for electric and gas respectively when regressions were estimated for each month of the baseline period.

The figure below shows a distinct difference between monthly and annual regressions. The baseline region, which is the green portion of the graphs, in the annual regression was constant across the entire year. However, when regressions are estimated for each month in the baseline period, it is clear that baseline energy consumption varied in different months of the year.

2.12.2 Test 2: Optimal number of regressions in baseline period

Github issue: https://github.com/CalTRACK-2/caltrack/issues/85

Background:

In the context of hourly methods, it was established that a single estimated regression for the entire baseline period likely increases CVRMSE. One possible strategy to address this problem is estimating a regression for each month of the baseline period, which results in 12 estimated regressions. Unfortunately, models fit from limited time periods without enough data points may become overfit.

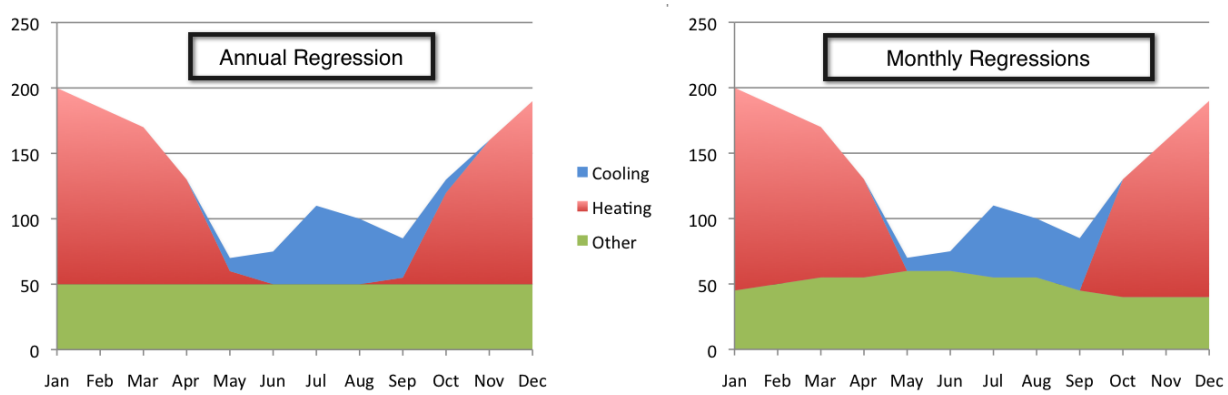

Fig. 20: *Figure: Annual vs. monthly regressions*

To determine the optimal number of regressions in the baseline period, the following regression intervals were tested and compared based on their in-sample and out-of-sample CVRMSE. The out-of-sample CVRMSE indicates if the model is overfit.

Regression intervals:

1. Year-long baseline: One regression was estimated for the entire baseline period.
2. 3-month baseline: A regression was estimated every 3 months of the baseline period.
3. 3-month weighted baseline: A regression was estimated for each month of the baseline period, but the months before and after were included and weighted down by 50%.
4. 3-month weighted baseline with holiday flags: This was the same as the 3-month weighted baseline but indicator variables for holidays were included in the regression specification.
5. 1-month baseline: A regression was estimated for each month of the baseline period.

Data:

Hourly data from residential buildings in California.

Tested parameters:

The in-sample and out-of-sample CVRMSE were calculated for different numbers of regressions estimated in baseline period.

Testing methodology:

Models with the five candidate regression intervals were estimated and the CVRMSE for each of these was calculated with in-sample and out-of-sample data.

Results:

The graph below presents the in-sample and out-of-sample CVRMSE for each regression interval candidate. The optimal regression interval had the lowest in-sample CVRMSE without overfitting the model. The out-of-sample CVRMSE increases when the model was overfit.

The results show that the in-sample CVRMSE with a year-long baseline was much larger than the other regression intervals. The 1-month interval had the lowest in-sample CVRMSE, but had a larger out-of-sample CVRMSE than the 3-month and weighted 3-month baselines. This signals that the 1-month interval may overfit limited data.

Conclusion:

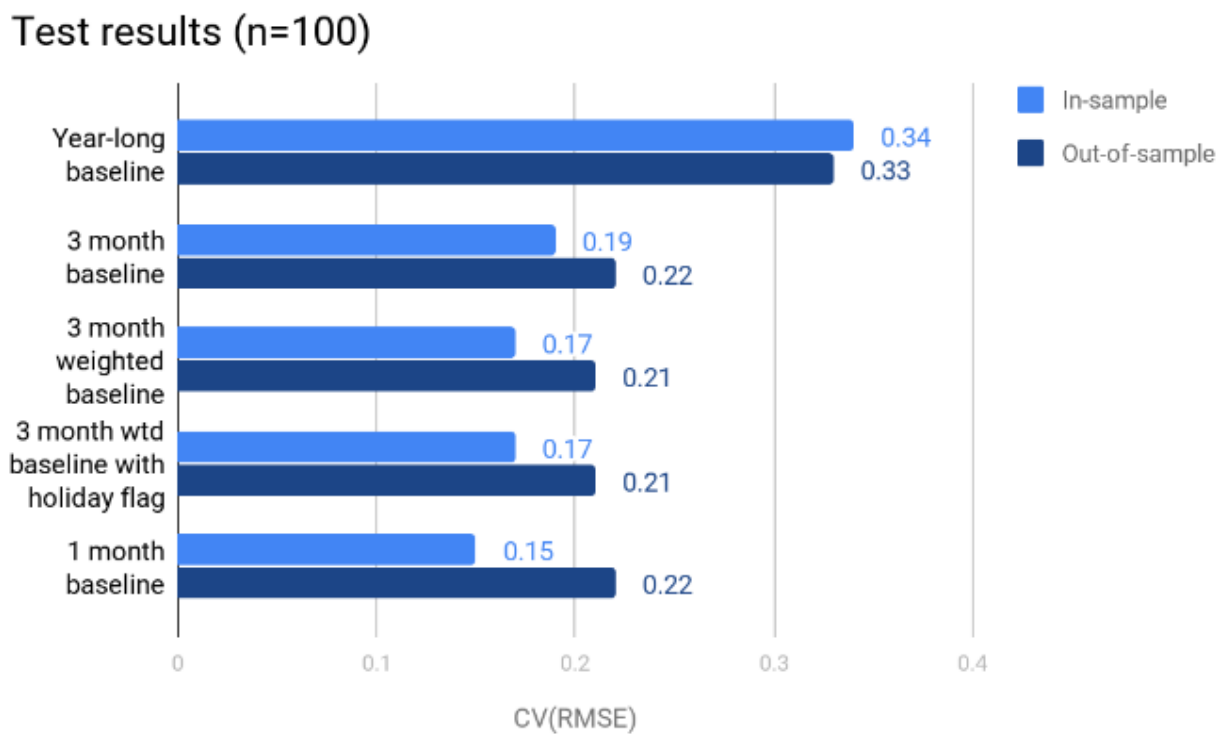

Fig. 21: *Figure: In- and out-of-sample CVRMSE*

The results showed that estimating a single regression for the entire baseline period likely increased the CVRMSE, which is not ideal for CalTRACK methods.

The 3-month weighted baseline is the preferred number of regressions for the baseline period. This approach accounts for variation in baseline energy consumption across months without overfitting the model to limited data.