Outdated Comment directly in Github

(https://github.com/impactlab/caltrack-betatest/blob/mcgeeyoung-patch-2/data-prep/README.md)

CalTRACK v1 Monthly Billing Analysis Data Preparation Guidelines

The CalTRACK Beta test worked primarily with hourly usage data, even for monthly billing analysis,[1] and as a result did not formally test data preparation guidelines for monthly billing analysis. Because of group members' extensive experience working with monthly billing data, the working group recommends the following processes be followed when preparing monthly consumption, weather, and project data for performing the monthly billing analysis specified in the CalTRACK v1 monthly methods.

Data Preparation Overview

A full accounting of the data cleaning and processing steps that were required for the purposes of the Beta test can be found here. General guidance suggests that data cleaning processes should be well documented and reviewed. There are countless small decisions that must be made as edge cases in the data arise. Thorough documentation ensures that evaluators understand the implications of these choices. Below are guidelines and a general process for addressing the most common issues that arise during data cleaning efforts for monthly billing

[1] It is worth noting that the Beta testers initially received the same data from PG&E (consumption) and Build it Green (projects) and tried to arrive at the same set of outputs for making savings calculations. However, despite multiple attempts at reconciliation, the Beta testers were unable to fully match their processes to produce identical outputs. Instead, for the remainder of the testing process, a subset of data generated by Open Energy Efficiency was used by the Beta testers in order to ensure that methods implementations for monthly analysis of hourly data rolled up to calendar months could be effectively compared.

analysis, based on the experience of the Beta testers doing prior billing analyses as well as lessons learned during the CalTRACK testing process. We recommend doing them in the order they appear because we found through testing that the final combined dataset you end up with is highly sensitive to the order of data preparation steps.

The CalTRACK data preparations guidelines for monthly billing analysis consist of the following steps:

1. **Project data preparation**
    a. **Generate necessary project fields from raw project data**
    b. **Deal with missing and miscoded values**
    c. **Deduplicate project records**
2. **Weather data preparation**
3. **Monthly electric and gas use data preparation**
4. **Link project and electric use files**
5. **Linked project+electric use data preparation**
    a. **Deduplicate records based on combined attributes**
    b. **Drop observations not meeting data sufficiency requirements**
6. **Link project records and gas use files**
7. **Linked project+gas use data preparation**
8. **Link weather data and project records**
9. **Final combined data sufficiency checks**

1. Project Data Preparation

The minimum field requirements for project data under the CalTRACK monthly specification are outlined here. Notably, a prepared project file should consist of one row per project, with a unique ID that can be used to link to gas and/or electric usage data, project start and stop dates, and zip code for the site. The following data cleaning steps for project data are meant to ensure that the prepared project file meets these field requirements and uniqueness constraints.

Creating Work Start and Work End dates from raw project data

Accurately identifying baseline and reporting periods is important for reducing the modeling error associated with a savings calculation. However, we have observed considerable variation in database records identifying dates associated with project start and project completion.

In general, our guidance is to try to identify the fields in a project record that most closely match the actual work start and work completion dates. In the absence of either of these fields (that is, if a project record only contains one or the other set of dates), we recommend identifying an average time to completion.

For the Beta test data, we worked with the program implementer to find the best proxy for work start and work finish dates. An initial version of the project data required estimation of some dates, but an updated version contained a complete set of work start and stop dates for all of the projects. Specifically, The work start date and work end date fields for projects done before July 1st, 2016 may require using column mappings for proxy fields. For projects started after July 1, 2016, CalTRACK implementations should use official work start date and work end date fields provided by aggregators instead of the proxy fields.

Dealing with miscoded dates

- Implausible day values (>31) should be coded as the beginning of month if project start date and end of month if project end date so that the entire month is included in the intervention window
- Implausible month and year values should be flagged and that home not included in estimation.

Deduplicate project records

- If a home appears multiple times within a project database, and the project dates are the same the most complete record for that home should be used
- If a home appears multiple times within a project database and the project dates differ because there are multiple measures installed associated with the same incentive program, the start date of the intervention should be the earliest of the project start dates across projects and the end date for the intervention should be the latest of the project end dates.

2. Weather Data Preparation

For CalTRACK monthly billing analysis, the weather data requirements are detailed here. For monthly analysis, since the daily average temperature data from a nearby weather station is used to create values for the number of heating degree days (below 60F) and cooling degree days (above 70F) in each billing period, the primary consideration in preparing the data is how to deal with missing values.

Dealing with missing values

Weather data is notoriously incomplete, especially at the granular sub-daily level. Some weather stations generally fail to report data, other weather stations are simply inconsistent in reporting data. This becomes an issue when trying to match projects to their local weather conditions. If a nearby non-reporting weather station is selected, the savings model will fail. If the project is connected to a nearby intermittently reporting weather station, the model will suffer. Additionally, if a project is connected to a weather station that experiences a significantly different local micro-climate, the model will suffer.

- We recommend that for monthly billing analysis daily average temperature values that are missing not be imputed, but rather count against a site in meeting data sufficiency requirements (detailed at the end of the document).

3. Monthly electric and gas use data preparation

Dealing with missing values in monthly usage data

Usage data generally undergoes significant cleaning prior to release to program administrators or the general public. There are generally three types of missing usage data. First, monthly billing data will be populated with “estimated” reads, when the utility has imputed a likely consumption amount for the month for the purposes of billing, but has not actually recorded a meter reading. Second, there are gaps in AMI meter data, where there may have been a hardware failure or another similar type of infrastructure breakdown where the data was not recorded. Finally, there is the issue of data that goes missing in the process of transferring to program evaluators (in the CalTrack Beta test, two of the zip files holding monthly billing data were corrupted and unreadable). Each of these issues represents a unique challenge and must be dealt with independently.

- For the case of missing values where the cumulative value is in the following period (as in an estimated read), the cumulative number of days between the two periods will be used to generate the use per day for that period
- Missing usage values with no cumulative amount in the following period (such as missing AMI data) will be counted against data sufficiency requirements
- The working group does not offer firm guidance on auditing data sets for completeness, however, a data audit was conducted for the purposes of the Beta test and was found to have identified several missing data issues that would not have otherwise been identified. Thus, a data audit is generally recommended.
- If flags exist for estimated values, they are counted as missing and count against the site’s data sufficiency criteria detailed later in this guidance.

Dealing with extreme values in usage data

Occasionally, the project or consumption data may contain extreme values that are likely the result of a data error, but may also be an indicator of another factor (such as the presence of solar panels). We offer the following guidance:

- Negative values for monthly use should be treated as missing and count against sufficiency criterion. Negative values in monthly data may also be a valid sign of possible solar/net metering and should be flagged for verification.

4. Link project records and usage files

Once project, consumption, and weather data have met all of their respective requirements, the data must be matched in order for a savings estimation to be performed. We recommend using a key such as a utility account number that will clearly match a given project with a given meter. However, we also recognize that in certain cases, a project may encompass more than one meter or utility account. In these cases, we do not offer specific guidance.

- For the purposes of the Beta test, projects were matched to consumption files using a cross reference file supplied by the program administrator.

Unmatched data should be excluded from analysis.

- For the purposes of the Beta test, projects that were unmatched to usage data were listed in CalTRACK for data integrity reporting, but were not included in any estimation procedures and did not have estimated savings.

5. Linked project+use data preparation

Deduplicate records based on combined attributes

- If two duplicate records have identical consumption traces and date ranges, drop one at random
- If two duplicate records have identical consumption traces but different date ranges select the more complete record having more dates. If the dates are contiguous, or there are overlapping dates with the same usage values, combine the two traces into a single trace.
- If the records have the same date ranges, but different usage values, the project should be flagged and the record excluded from the sample.

Drop records not meeting data sufficiency requirements

Calculating energy efficiency savings requires a sufficient observation period of energy usage prior to and after an intervention. Generally, annualized models require at least 12 months of usage data on each side of an intervention in order to accurately calculate energy savings. Some models may be able to calculate energy savings with fewer than 12 months of data in the reporting period.

- 12 complete months pre-retrofit for monthly billing data to qualify for estimation or 24 months with up to 2 missing values from different, non-contiguous months
- Post retrofit data sufficiency for estimation will be dealt with in post-estimation model fit criterion
- Total annual savings estimates will require 12 months post-retrofit

Drop project records with unsupported characteristics

- Drop homes with known PV or EV added 12 months prior to or up to 12 months after the intervention. During the CalTRACK beta test, these homes were identified from the presence of reverse flow in the AMI data and/or indications of net metering in the cross reference tables. However, if you only have access to billing data, we recommend working with the utility to get flags for accounts that have net metering present so they can be excluded from the analysis.

6. Link weather data and project records

Weather station mapping requires locating the station nearest to the project. Each project file should contain a zip code that allows matching weather stations to projects

- For the purposes of the Beta test, weather station mapping was done using the 86 station standard mapping of zip code to CZ2010 weather files.

7. Final combined data sufficiency checks

- Billing periods (the period between bill start date and bill end date in the monthly usage data) with more than 10% missing days of weather data will be thrown out and count against the required number of billing period observations
- Any projects with fewer than 12 months pre and 12 months post are not included in the analysis