Eric Ziemba, Seth Wayland

Cleaning Up the Mess of Billing Data: An Investigation of Differences in Billing Analysis Results Caused by Data Cleaning Methodologies

In this paper we investigate how the level of rigor used to clean billing data affects modeled energy savings. Are there significant benefits to the rigorous cleaning of billing data, or can a cursory and more time efficient level of meticulousness be used. In other words, does a cleaner dataset lead to more accurate results? Utilizing a dataset that combines participant tracking and billing data from four home energy audits programs, we clean data at increasing levels of rigor. We then compare savings estimates to investigate how data cleaning affects the results of billing analyses. Preliminary results suggest that raw data may provide savings estimates that are with the 90% confidence interval of estimates from fully cleaned data.

Billing analyses often begin with rigorous cleaning of program tracking and customer billing data. Data from utility billing systems often include missing or misleading observations. Common issues include missing or negative energy usage, and billing periods that overlap each other or have gaps between them. Data often include estimated records, or meter reads that are meant to be adjustments to other billing periods. The time required to fully address these data issues varies, but in rarely inconsequential. Is this always time well spent, or can efficiencies be gained through more streamlined data cleaning?

Even with fully cleaned data, variations in when a customer’s meter is read can cause billing periods from the same month to represent different time periods. To address this additional source of potential error, we test two methodologies for aligning the billing periods. The first method assigns the period to the month in which the most days of bill occurs. The second, referred to as calendarization, takes the average daily energy consumption from each bill, assigns that value to each day, and constructs new billing periods based on calendar months. We look to quantify any differences in savings provided by each method, and show the advantages and pitfalls to each method. The methodologies are tested using case studies from prior evaluations of home energy audits, conducted by our firm.

Differences in billing analysis results based on these data cleaning and billing period methods, are cause for concern. Identifying how data cleaning methods effect final savings estimates will lead to improved accuracy in billing analyses for evaluating energy efficiency programs. Understanding these effects will become increasingly important as our industry continues to move into more data intensive evaluations. Full results from this work are expected the available at the end of 2016.