Differentially Trendy? Understanding and Addressing Non-parallel Trend Bias in Smart Thermostat Impact Estimates

Peter Franzese, CPUC, San Francisco, CA
Ken Agnew, Lullit Getachew, DNV, Madison, WI

ABSTRACT

Consumption data analysis, traditionally referred to as billing analysis, is an important tool that provides empirically based impact estimates for residential program populations. However, many acknowledged challenges of the approach can affect the quality of those estimates. While these challenges exist for all billing analyses to some degree, they become particularly acute when attempting to estimate small savings from highly self-selected program populations. Smart thermostat programs have the potential to attract just such populations, and impact evaluations of these programs have, in recent years, highlighted the challenges of this kind of consumption data analysis. Consequently, the methodological challenges of consumption data analysis have been frequently cited as the reason for low savings estimates in smart thermostat evaluations, even though there is limited empirical evidence to confirm this claim.

This paper offers a technical response to this claim. Although no method is without limitations, the two approaches described in the technical response, together with their results, provide strong evidence that consumption data analysis can yield valuable insights even in challenging applications such as smart thermostats. More generally, this paper attempts to summarize the underlying challenges facing consumption data analysis in a way that is accessible to a broader audience. It seeks to illustrate that the challenge lies not in the statistical methods themselves, but in how we apply them, which is an issue that non-statisticians can grasp and help address.

Introduction

Consumption data analysis is a valuable tool for generating empirically based impact estimates for residential program populations using utility electric and gas consumption data of any granularity. However, the process faces many challenges that can affect the quality of these estimates. While such challenges are present in all consumption data analyses to some degree, they are particularly acute when estimating small savings from highly self-selected program populations.

Technically, this issue arises from a correlation between unobservable participant consumption characteristics and program outcomes. Conceptually, it reflects the fact that some consumption trends cannot be captured by the modeling approach, leading the analysis to conflate them with estimated savings. Although these unaccounted-for trends are generally small relative to overall consumption, they may be comparable in magnitude to the expected savings. Moreover, as self-selection causes the participant population to diverge further from the general population, these trends become both more likely and more influential.

These challenges have come into sharper focus in recent years through numerous impact evaluations of smart thermostats, which exemplify the conditions under which consumption data analysis is most vulnerable: Small expected savings and highly self-selected participants. Compounding these issues, the evaluations have shown substantial variability in smart thermostat impacts, with notable differences not only across brands but also within the same brand over time. This inconsistency has left open the broader question of whether, and under what conditions, smart thermostats can reliably deliver energy savings.

Since both utility energy efficiency program staff and thermostat manufacturers are invested in the success of smart thermostats, the methods used to analyze consumption data have become a central point of contention in the discussion around smart thermostat performance. Even in the absence of strong supporting evidence, these methodological challenges have often been cited as the reason for the modest savings observed in many evaluations.

In light of these concerns, this paper outlines a technical response to claims that consumption data analysis underestimates smart thermostat savings. While no model is perfectly accurate, the analysis presented here reinforces the value of consumption data analysis, even in difficult contexts. More broadly, it aims to make the core challenges and limitations of these methods accessible to a wider audience.

Billing Analysis and Its Shortcomings

Consumption data analysis, traditionally referred to as billing analysis, has been a core tool in the energy efficiency evaluation industry for at least four decades. Initially limited to the monthly consumption data used to calculate customer bills, it has since expanded to include more granular energy usage data. The fundamental premise of billing analysis is that comparing energy consumption before and after an energy-efficient installation provides an estimate of the savings achieved.

Seasonality and year-over-year weather variation complicate this comparison. As a result, controlling for weather effects is a key part of consumption data modeling. This modeling allows consumption from both periods to be adjusted to a common weather basis. Weather is a clear and relatively easy-to-address reason why pre-post consumption differences might fail to estimate savings accurately.

However, many other factors can also affect household energy consumption over time, and these can influence how accurately the pre-post difference reflects savings. These factors are essentially private household details, such as lifestyle changes or major purchases, that influence energy use. Analysts typically don't have access to information on these, so they cannot be directly incorporated into models or normalized like weather data. These non-program, non-weather factors are labeled exogenous effects.

To a residential energy customer, exogenous effects might include a new baby, the arrival of an electric vehicle, a child leaving for college, or the removal of a large tree provding natural shade. Each of these can meaningfully impact consumption, independent of any program participation. To the energy analyst, they represent uncontrollable variability that, under certain conditions, can introduce directional error or bias.

Thus, while consumption data analysis is generally considered a statistical modeling exercise, the real challenges stem from the unavoidable limitations in our ability to model household consumption fully. If we had complete visibility into household-level changes and could incorporate those into models, many of these issues would disappear. Instead, we rely on the best feasible solution to this fundamental problem: the comparison group.

Constructing a suitable comparison group is one of the primary strategies used to address the limitations of consumption data analysis. In theory, if the comparison group is subject to the same unobserved trends as the participant group, any bias introduced by those trends can be mitigated.

Ideally, this group would be a perfect mirror of the participant group, identical in all respects except for not receiving the program intervention. In such a case, a standard difference-in-difference (DID) framework would yield an unbiased estimate of savings without the need for complex statistical adjustments. While randomized controlled trials (RCTs) come closest to this ideal, they are generally impractical in the context of rebate programs, where participants self-select rather than being randomly assigned.

Instead, analysts must construct comparison groups after the fact, using observable characteristics correlated with both participation likelihood and program outcomes. One practical approach is to use future or past participants as comparison candidates. These customers, like current participants, exhibit the key trait of program engagement, which may serve as a proxy for unobservable characteristics. For this reason, the Uniform Methods Project (UMP) recommends this strategy as the best available option for constructing comparison groups in billing analysis.[1]

However, this “best” solution underscores the limited data available for constructing comparison groups. We lack the information needed to identify matches likely to share similar consumption trends over time, just as we lack data to control for the exogenous effects driving those trends directly. Matched comparison groups are intended to address exogenous change through the assumption of parallel trends, but we have little reason to believe those trends will remain aligned.

Nevertheless, the generally accepted practice is to match on pre-participation consumption characteristics to construct comparison groups. Although using pre-participation energy consumption data is not ideal, because it offers limited insight into future energy use trends and is also used to measure outcomes, it remains the common approach. This approach helps create similar groups at the outset and reduces one key difference between the groups within the DID framework.

One way to think about DID is thus as a post-installation comparison between participant and comparison customers, adjusted for pre-installation differences. A matching strategy that minimizes the pre-installation differences between the two groups reduces the need for the post-installation adjustment. More importantly, it means the analysis relies entirely on the assumption that both groups will follow parallel trends over time.

If the participant and comparison groups are similar in the pre-installation period, the post-installation difference will only be unbiased if both groups’ consumption evolves in parallel (i.e., the parallel consumption trend assumption). Otherwise, trend differences, driven by different sets of exogenous effects, will be conflated with the savings estimate. If the comparison group’s consumption trend is greater than the participants’, savings will be overstated. If the participants’ trend is greater, savings will be understated.

Because comparison groups are typically selected based on static consumption characteristics and a handful of other variables with limited explanatory power, especially with respect to customer traits or long-term consumption trends, it is reasonable to question the validity of the parallel trends assumption.

Thermostats as a Billing Analysis Subject

Thermostats as an energy efficiency measure have a long history. Programmable thermostats were initially hailed as offering substantial savings. They were promoted as energy efficiency measures, despite offering only the novel capability of setting a schedule and leaving it. It took years to establish that baseline setpoint behaviors were often quite durable, regardless of the presence of a new programmable thermostat. Even when customers used the scheduling feature, they typically replicated their existing manual thermostat setpoints in the programmed schedule.

Indeed, for households that already practiced manual setbacks, programmed setpoints facilitated pre-conditioning, such as pre-warming a house before rising on a winter morning or pre-cooling before returning from work on a hot afternoon. Ultimately, the evaluation community did its job and sent a clear signal: programmable thermostats were not, on their own, the solution to optimizing HVAC consumption in residential homes. These findings were established through billing analysis and extensive survey analysis exploring customer usage behaviors.

Smart thermostats emerged not too many years after programmable thermostats largely disappeared from technical resource manuals. Smart thermostats are programmable but also offer additional functionality aimed at reducing HVAC-related energy consumption. This functionality may

[1] https://docs.nrel.gov/docs/fy17osti/68564.pdf

include geo-fencing, which detects when no one is home and adjusts HVAC usage accordingly, or more subtle optimization algorithms that make small setpoint adjustments to save energy over time.

Smart thermostats address some of the shortcomings of programmable thermostats. Unlike their predecessors, they can automatically adjust heating and cooling usage without requiring customer intervention. However, customers retain complete control and can override any optimization or reset setpoint schedules at any time. As a result, customer behavior still plays a critical role in determining how effective a smart thermostat will be in saving energy. Ultimately, smart thermostat savings depend on the device's algorithms making measurable and lasting changes to household heating and cooling patterns.

Estimating the energy efficiency impacts of smart thermostat programs has proven challenging. Results have been highly variable. In some instances, where results proved lower than expected, thermostat manufacturers have argued that standard billing analysis methods, commonly used in residential energy efficiency evaluations, could not accurately estimate smart thermostat savings.

The argument relies on well-known shortcomings of billing analysis, namely, that smart thermostat purchasers may exhibit rising energy use relative to the general population. Proponents suggest that this group often includes younger, well-off, tech-savvy consumers who are more likely to experience life changes, such as having children and otherwise accumulate energy-using gadgets around the time they adopt a smart thermostat. The inability of regression or matching methods to control for these specific participant characteristics means that the comparison group will look more like the general public with a lower consumption trend. Thus, the comparison group will not appropriately track the participants' natural consumption increase and, as a result, savings will be downwardly biased. This argument suggests a one-way, downward bias to thermostat savings, but there is limited hard data to back up any of its assumptions.

Possible Solutions

In this challenging environment, we conducted two evaluations of California's statewide smart thermostat programs. These studies addressed the issue of non-parallel trends in billing analysis broadly, and in smart thermostat evaluations specifically, using two complementary modeling approaches. The first applied a two-stage billing analysis with an adjustment for observed trend differences. The second formalized this adjustment within a panel model that directly estimated differential trends between participant and comparison groups. Together, these approaches provide a robust framework for evaluating smart thermostat impacts when traditional assumptions may not hold and support the credibility of the results.

Different Program Designs

The studies we pursued included two program types that provided smart thermostats under very different designs and served distinct customer populations. One was a standard opt-in rebate program that gave a relatively small rebate as an incentive to buy a smart thermostat. Even with the rebate, the price of a smart thermostat was substantially higher than competing thermostat technologies. Those customers who participated in the thermostat rebate program had some combination of interest in this new technology and a willingness to pay for it.

The other was a direct install initiative targeting low- to moderate-income customers, where thermostats were provided and installed at no cost. While these participants also opted into the program, they represent a very different demographic from the rebate program population.

The non-parallel trends hypothesis is specifically tailored to the rebate program population. In contrast, there is considerably less reason to believe that the direct install population would experience the same theoretical increase in consumption that could downwardly bias the results.

With the same measure and consistent evaluation methods applied across programs, if the non-parallel trends hypothesis is accurate, we would expect the direct install evaluation to provide results that were relatively less likely to reflect bias from a trend differential.

Component-Level Trends

To further explore the parallel trends issue, we considered whether certain components of energy consumption, particularly baseload versus heating and cooling, might be more prone to trend differences. Since smart thermostats are designed to regulate HVAC systems, it is reasonable to expect that their impacts would be limited to heating and cooling, and not appear in baseload consumption. At the same time, it seems reasonable to expect that a disproportionate share of consumption trend over time would be situated in baseload. Additions like EVs, electronics, and lighting primarily increase baseload consumption. In contrast, a household establishes its HVAC operations based on comfort preferences. Those may change gradually over time, as customers age, for example, but year over year, the relevant timeframe for an impact analysis, shifts in heating and cooling behavior are unlikely to drive substantial trend. Indeed, we would expect a change in occupancy to have a greater effect on baseload than heating or cooling. Hot water consumption in baseload is definitely correlated with the number of occupants, but occupancy-driven changes in heating or cooling would seem to be more limited unless the change in occupancy dramatically changes the existing thermostat setpoint schedule (e.g., different work hours).

Partitioning impacts provides a way to test whether the parallel trends assumption holds in the context of smart thermostat evaluations. While the idea that most consumption trend over time resides in baseload is speculative, it is difficult to argue that heating and cooling consumption would exhibit greater trend. Removing baseload impacts offers a useful, testable hypothesis: if baseload trends are the main driver of year-over-year differences, then excluding them should reduce bias in the savings estimate. More specifically, if most of the upward consumption trend lies in the baseload, removing that component should reveal larger estimated savings in heating and cooling usage.

To operationalize this hypothesis, we applied two modeling approaches that separate energy consumption into heating, cooling, and baseload components, which we discuss in the following two sections.

Two Stage Partitioning

The first approach applies a two-stage billing analysis to estimate smart thermostat impacts across heating, cooling, and baseload components. This method allows for flexible modeling of individual household consumption patterns and provides a foundation for identifying where impacts are most likely to occur. By separately estimating pre- and post-installation consumption for each component, the two-stage approach offers a practical way to test whether observed savings are concentrated in heating and cooling, as expected, and whether baseload trends may be biasing overall results.

The first stage of the two-stage approach allocates each site's consumption to heating, cooling, and baseload. Site-level consumption modeling with a variable degree day base provides a breakout of heating, cooling, and baseload consumption based on a widely used structural model originally used in the PRISM model.[2] The underlying theory of the PRISM model holds that each site has optimal degree-day bases such that the model appropriately allocates consumption to heating and cooling when temperatures are colder or hotter than those respective bases. The chosen degree-day bases represent the outdoor temperature at which demand for heating and/or cooling becomes evident in the data, consistent with the best overall model fit. The heating and cooling parameters indicate the rate at which consumption increases with each additional degree day, above or below those bases, respectively. The

[2] Fels, M. F. (1986). PRISM: An Introduction. *Energy and Buildings*, 9(1), 5–18

final model provides a unique, optimal estimate of heating and cooling consumption for that house under any temperature regime.

Importantly, the pre- and post-installation site-level models for each site are estimated separately, and so are flexible to capture the expected shift in degree day base that would signify impacts of the sort a smart thermostat would produce. If the thermostats work as expected, heating demand will go down, the average heating degree day base will shift downwards (cooler) and heating consumption will be lower. As cooling demand goes down, the average cooling degree day base will shift higher (warmer) and cooling consumption will be lower.

The second stage model produces impact estimates by applying a DID structure to participant and matched non-participant, pre- and post-installation normalized consumption. Evaluators usually apply this approach at the household consumption level to get a single, overall savings estimate. Applying the DID model to the heating, cooling, and baseload components separately is less common, but it does offer insights into where the impacts are showing up. In this case, because thermostat impacts should only be present in the heating and/or cooling consumption, the parallel trends assumption would support a baseload impact result equal to zero.

Electric model results for the rebate program did, in fact, produce negative baseload impacts. Furthermore, those negative baseload impacts caused the overall impacts to also be negative. The preliminary rebate program results appeared to reflect the hypothesized outcome of non-parallel trends. The results also seemed to support the hypothesis that most of the upward trend would be found in the baseload. Positive, but small, heating and cooling savings indicated that non-parallel trends did not fully undermine savings.

In the interest of accounting for what appeared to be evidence of the non-parallel trend issue, we developed an ad hoc adjustment that would remove the estimated baseload trend from the savings estimates. Leaving the negative baseload savings out of the overall savings calculation was the obvious first step. This, however, adjusted for trend found in the baseload but not in heating and cooling. To adjust for a similar level of trend in heating and cooling, we calculated the negative baseload “impact” as a percentage of overall baseload consumption and applied that percentage to heating and cooling consumption as an upward adjustment to heating and cooling savings.

These adjustments treated heating and cooling savings as if heating and cooling consumption had similar year over year trend differential as was evident in baseload. For both electric and gas rebate program savings estimates, savings estimates were positive but still dramatically below ex ante claimed savings. Our hypothesis was that year-over-year trend in heat and cooling consumption would be lower than baseload trend. Applying the same trend produced only modest savings. The only way that non-parallel trends could lead to even higher ultimate savings estimates would be that year-over-year trends in heating and cooling consumption were much higher than baseload trends.

Panel Model Approach to Partitioning

Building on the insights and limitations of the two-stage approach, we developed a panel modeling framework further to test the presence and effect of non-parallel trends. Unlike the earlier method, which applied adjustments after estimating impacts, this second approach was designed to formalize the ad hoc adjustments used in the two-stage modeling and to assess differential trends directly within a more integrated modeling framework.

A standard panel model billing analysis approach also produces separate estimates of heating, cooling, and baseload consumption as well as separate impact estimates for each component. However, the typical panel model applies a single PRISM-like structure to all customers at once. The whole population is characterized by a single heating degree day base and a single cooling degree day base. In contrast to the site-level approach, this implies that all sites’ demand for heating or cooling starts at the same temperature thresholds. This produces heating, cooling, and baseload impact estimates but does so without any flexibility to the variability of households' heating and cooling characteristics across the population. Also, across time, this simple panel model structure does not allow for changes in heating and cooling demand to be appropriately captured by a shifting degree day base. These structural limitations of the standard panel model would have made it a poor vehicle for improving on the prior year's ad hoc adjustment.

To address this issue, we developed a panel model that combines the best aspects of site-level and panel model approaches. In this panel model, each site's consumption is a function of its own degree days from its own optimal degree day bases. These were established from a separate baseline-period, site-level model run. In the post-installation period, the specification allows those degree day bases to vary, at the site-level, reflecting the expected impact of the thermostat algorithm on individual customer setpoints. In the panel model structure, this provides the flexibility for each customer's heating and cooling consumption to shift appropriately due to the thermostat, as was the case with two-stage approach.

Furthermore, the model explicitly estimates separate trend terms for participants and comparison group members. Those trends are effectively informed by trend in the baseload but are included in both the heating and cooling elements. These trends mimic the ad hoc adjustment but are consistently estimated in the presence of all other model effects by the panel specification. This means savings are estimated in the presence of separate estimated trends across the groups.

The panel model we estimated, which included terms to account for the effect of weather, trend, and reference temperature changes, has the following general specification:

$$\begin{aligned}
&Y_{tj} = \alpha_g\left(1+\lambda_g t\right) + \\
&\beta_g\left(1+\lambda_g t\right)\left[H_t\left(\tau_{Hj}\right)\left(1-P_t\right) + H_t\left(\tau_{Hj}+\delta_H+\ \delta_{pH}*T_j\right)P_t\right] + \\
&\gamma_g\left(1+\lambda_g t\right)\left[C_t\left(\tau_{Cj}\right)\left(1-P_t\right) + C_t\left(\tau_{Cj}+\delta_C+\delta_{pC}*T_j\right)P_t\right] + \\
&\varepsilon_{tj}
\end{aligned}$$

In this model:

t = time period index, starting at $t\ = 1$

j = customer index

g = group index, where $g\ =\ p$ for participants and $g\ =\ np$ for non-participants

Y_{tj} = energy consumption for customer j at time period t

α_g = group-specific intercept term that captures baseload consumption of participants, where $g\ =\ p$ for participants and $g\ =\ np$ for non-participants

λ = trend term that increments daily energy consumption, with $\lambda_{np}t$ capturing trend for non-participants and $\lambda_p t$ capturing trend for participants

T_j = 0/1 dummy for customer j, which equals 1 if a customer is in the participant group, 0 otherwise

P_t = 0/1 dummy for time t, which changes from 0 to 1 at $t =$ participation date for participants and their matches

β = heating use per heating degree-day (HDD)

γ = cooling use per cooling degree-day (CDD)

$H_t(\tau_{Hj}) = HDD$ per day for customer j at heating reference temperature τ_{Hj}, at time period t

$C_t(\tau_{Cj}) = CDD$ per day for customer j at cooling reference temperature τ_{Cj}, at time period t

τ_{Hj} = heating reference temperature for customer j determined by site-level regression models

τ_{Cj} = cooling reference temperature for customer j determined by site-level regression models

δ_H = average shift in heating reference temperature for all customers in the post period

δ_C = average shift in cooling reference temperature for all customers in the post period

δ_{pH} = incremental shift in heating reference temperature for participants in the post period

δ_{pC} = incremental shift in cooling reference temperature for participants in the post period

The model includes terms that capture baseload consumption (alpha) for the participant and non-participant groups. These are interacted with trend terms (lambda) to capture possible differences in energy consumption trends between these two groups. The model also includes weather variables (CDD and HDD) to control for the effect of weather on energy consumption. The CDD and HDD variables were constructed using reference temperature estimates (τ_{Hj} and τ_{Cj}) derived from pre-period site-level models.[3] We also interacted these terms with the trend term interacted with baseload, which makes this the regression equivalent of the proportional adjustment we did in the two-stage analysis. The reference temperature shifts contend with a trend driver effectively driven by whatever trend difference exists in the baseload over time.

While the beta and gamma terms associated with heating and cooling terms capture the effect of weather on energy consumption, the model includes terms to estimate reference temperature changes (delta) of participants and non-participants. The delta terms for both groups capture reference temperature changes in the post-period. Additional reference temperature change terms for participants capture incremental changes following smart thermostat installations. The estimates of the incremental delta terms provide direct evidence of shifts in reference temperature that reflect the impact of smart thermostats on energy consumption.[4]

Results

Results from the panel model are presented in Figure 1 alongside estimates from the two-stage DID approach, both with and without trend adjustment. The left side of the top (electric savings) panel compares unadjusted and trend-adjusted results for the rebate program. Both unadjusted estimates are negative, suggesting the possible influence of non-parallel trends. Their similarity confirms consistency between the two modeling approaches, which rely on the same underlying data.

The two trend adjustments, one ad hoc based on the two-stage billing analysis results and the other from the panel model, represent distinct efforts to correct for these trends. Both adjusted estimates are positive, indicating that accounting for trend differentials shifts the savings modestly upward. Notably, the panel model yields a lower savings estimate, suggesting that the earlier ad hoc adjustment may have overstated savings.

The parallel rebate programs gas results on the left side of the bottom (gas savings) panel tell a different story. While the ad hoc adjustment shifts the unadjusted negative results to positive by

[3] Details of the weather normalization models are also provided in DNV's PY2020 impact evaluation report.

[4] Our analysis does not presume that $\tau_{.j}$ estimates setpoints. It assumes that shifts in setpoints result in shifts in $\tau_{.j}$ by the same amount, which is what basic PRISM theory states.

construction, the panel model does not produce a similar shift. Instead, despite correcting for a small trend differential, the panel model results become more negative. This divergence highlights that the panel model does not simply replicate the two-stage results. Overall, both electric and gas rebate program estimates suggest only modest savings, even under adjustments designed to account for hypothetical non-parallel trends.

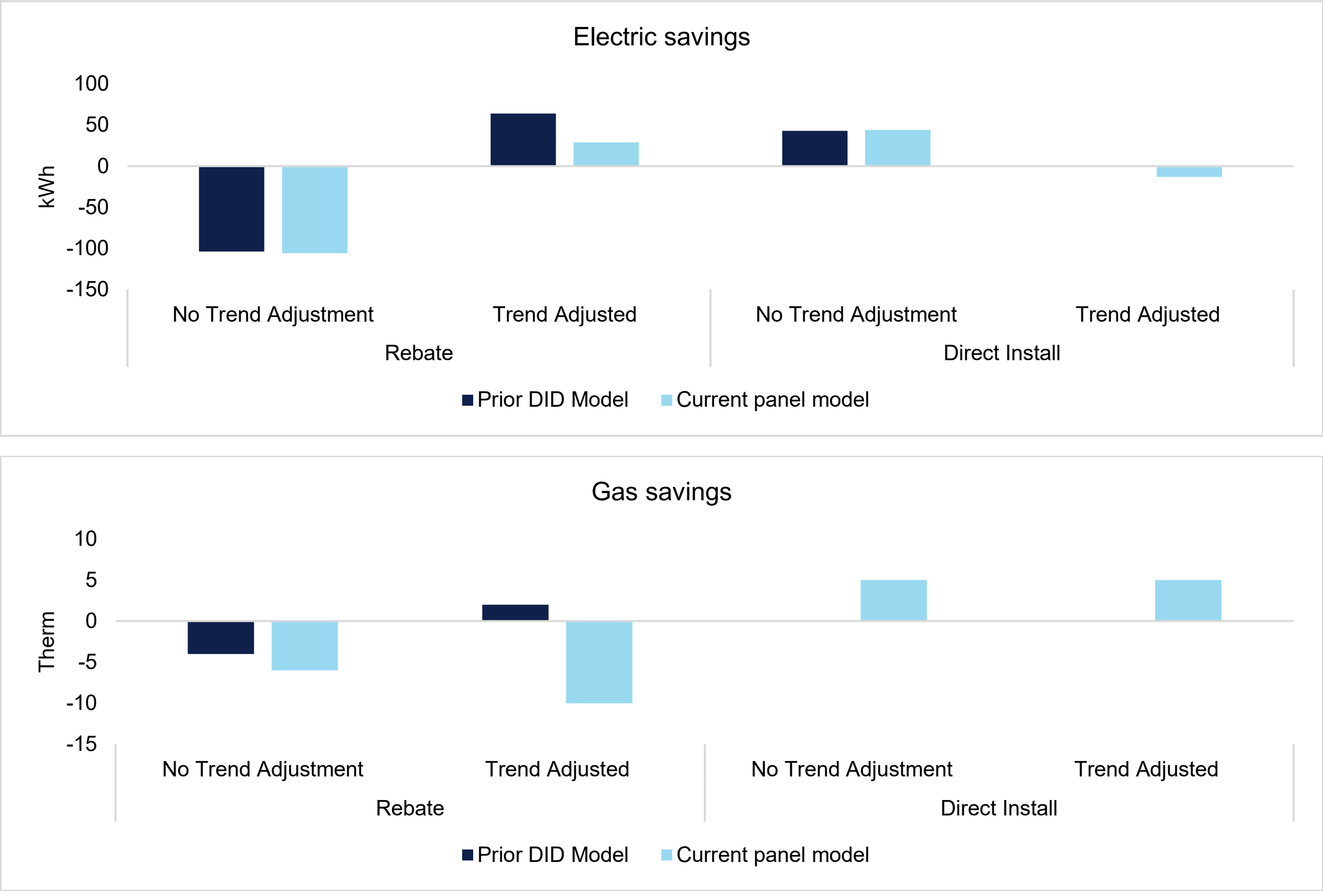

Figure 1. Current panel and prior model estimates of first-year savings per household

The direct install results are more nuanced but ultimately lead to a similar conclusion. Again, the electric savings estimates across the no-trend-adjustment versions of both models present in the right panel are almost identical. The savings are modest. The ad hoc adjustment was not applied for direct install customers due to the absence of a differential baseload trend. The panel model does adjust the savings, but like the gas rebate results, the adjustment is downward. Those results are slightly negative, effectively zero. The gas results without trend adjustment are zero and a modest five therms for the prior year DID model and the panel model, respectively. As with the electric direct install results, the ad hoc adjustment was not applied. The panel model trend-adjusted results are indistinguishable from the unadjusted results.

These results ultimately provide support for modest non-parallel trends for the rebate program while also undermining the argument that those differential trends are the reason smart thermostat savings are so low. The electric rebate program results correct for apparent bias across both modeling approaches. However, the magnitude of that correction produces only modest positive savings. These adjustments both reflect the generous assumption that heating and cooling trend differentials are as big as baseload trend differentials. The only way the adjustments could produce higher savings would be if heating and cooling trend differentials outstripped baseload trend differentials. We can think of no scenario that supports this, nor does it appear that even such supercharged adjustments would move savings toward the levels claimed by programs.

The direct install results provide modest savings estimates in the context of a population for which there is little reason to expect a trend differential. Unlike the rebate program, the direct install participant population does not fit the hypothetical profile of young, tech-savvy consumers early in their energy-consuming lives. The panel model results bear this out with no upward adjustment, although the model is flexible to differential trends in the data that could lead to that result.

The panel model indicates that the ad hoc adjustment approach may overstate thermostat savings. However, as a less technical approach, the ad hoc approach offers a simple way to reasonably acknowledge the challenges faced by quasi-experimental design approaches while still also demonstrating that smart thermostat savings are modest.

Conclusion

This paper discusses the ongoing challenge of recognized shortcomings in widely used billing analysis methods. The concern for bias due to non-parallel trends is not new, but has been pushed to the forefront in our industry in recent years with the evaluation of smart thermostats. Supporters of smart thermostats, knowledgeable of concerns related to non-parallel trends, saw those shortcomings as an explanation for what they considered the too-modest savings showing up in smart thermostat impact evaluations. Though smart thermostat rebate program participants could be characterized in ways that made the possibility of non-parallel trends appear plausible, there was never solid evidence that this hypothesis had any particular credibility. Despite the lack of well-supported alternative results, the non-parallel trends argument was used to question evaluators' efforts to develop valid empirical estimates of thermostat savings.

This paper attempts to shed light on this challenge in a way that is accessible to a broader audience, including non-technical analysts. The nature of the non-parallel trend challenge does not require statistics to understand. Indeed, it is rooted in household characteristics and lifestyle changes that are profoundly relatable, and it is precisely the accessibility of the non-technical narratives seemingly supporting a problematic trend differential that made this an issue in the first place. The hypothesis of a young, tech-savvy population having more babies has a ring of feasibility even if there is little concrete evidence to support it.

The results presented here for this smart thermostat evaluation provide a strong, if necessarily incomplete, response to the claim of non-parallel trends bias being the driver of low thermostat savings. Through two reasonably constructed adjustments and a comparison with a very different, direct-install program, the modest thermostat savings remained consistent. There was, in fact, some evidence of non-parallel trends, but because of the unique nature of thermostat savings, which are limited to heating and cooling consumption, we were able to demonstrate with some confidence that the bias was not enough to rescue hoped-for savings. While these methods are primarily applicable to smart thermostats, the considerations regarding non-parallel trends have broader implications. It is through understanding the shortcomings of our methods that we can provide the most accurate results.

References

Agnew, Ken and Mimi Goldberg. 2017. "Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol." *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. Golden, CO: National Renewable Energy Laboratory. NREL/SR-7A40-68564. https://docs.nrel.gov/docs/fy17osti/68564.pdf

California Public Utilities Commission. 2023. *Forward-Looking Smart Thermostat Study. CPUC Group A Impact Evaluation – Final Report*. CALMAC ID: CPU0367.01. https://www.calmac.org/publications/CPUC_GroupA_Fwdlooking_Tstat_FinalReport.pdf

California Public Utilities Commission. 2021. *CPUC Group A Residential PY2019 Smart Thermostat Impact Evaluation – Final Report*. CALMAC ID: CPU0354.01. https://www.calmac.org/publications/CPUC_Group_A_Residential_PY2019_SCT_Final_Report_CALMAC.pdf

Fels, M. F. 1986. "PRISM: An Introduction." *Energy and Buildings*, 9(1), 5–18.

Stewart, J. I., Olig, C., Shahinfard, S., Agnew, K., Wayland, S., Horvath, Z., Lai, J., & Kurnik, C. 2023. "Smart Thermostat Evaluation Protocol*:* December 2016–May 2023" *NREL Technical Report No. NREL/SR-5R00-86175. National Renewable Energy Laboratory*. https://doi.org/10.2172/1988025