Getting our Ducts in a Row: A Billing Analysis of Duct Sealing and Heat Pumps in the Northwest

Author: Jenny Yaillen - Evergreen Economics
Co-author: Carrie Cobb - Bonneville Power Administration

ABSTRACT

Northwest energy efficiency programs have included duct sealing and heat pumps, which provide significant savings potential, for more than a decade. The study described in this paper is the first impact evaluation of this program and, in part, is intended to measure the effectiveness of the program specifications. The evaluation included collecting customer billing data; developing a comparable control group; using a stakeholder advisory committee to assist in reviewing results; and using a number of different regression models to estimate savings.

This evaluation included collecting billing and program tracking data from 42 public utilities spanning 15,490 households in Oregon, Washington, Montana, and Idaho. The efforts associated with collecting and merging these billing data are particularly relevant as there is a push in many regions for statewide evaluations. Another element of the evaluation was the development of an appropriate control group. This paper discusses these challenges, lessons learned, and our successes.

Multiple regression models were developed including differences-in-differences, fixed-effects, and statistically adjusted engineering (SAE) model specifications. This paper describes why the fixed-effects specification was selected over the other models and how model selection was influenced by the dataset. Stakeholder review was also a valuable component of the evaluation, which informed model selection.

The methodology and findings presented in this paper will be of interest to a wide audience, given its use of large-scale data collection, a comparable control group, multiple regression modeling techniques, and the use of stakeholder input throughout the evaluation. The challenges and lessons learned in this evaluation are applicable to program evaluation in a variety of contexts and regions.

Introduction

The Bonneville Power Administration (BPA) is a federal nonprofit agency based in the Pacific Northwest. BPA is a wholesale power marketer with over 140 utility customers. Since the passing of the Pacific Northwest Electric Power Planning and Conservation Act ("the Act") was passed in 1980, BPA and its regional wholesale power customers have acquired cumulative electricity savings of more than 8,700 GWh (BPA 2012). Because BPA's customers are primarily utilities, rather than end-users of electricity, BPA's role is to support energy efficiency by offering efficiency measures and programs; technical tools and support; and financial resources to regional utility customers.

One of the programs offered by BPA is the Performance Tested Comfort Systems (PTCS) program. PTCS is a training, certification and quality assurance program to

deliver heat pumps and duct sealing. PTCS's specifications and standards are designed to ensure the optimal installation of quality duct sealing and heat pumps in order to maximize the energy savings. PTCS was developed by the Northwest Energy Efficiency Alliance (NEEA) in the late 1990s. BPA began offering PTCS in 2006 and between 2007 and 2011 the program tripled in participation. The program delivers the following measures:

- **Duct sealing.** This measure addresses ducts in both conditioned and unconditioned spaces that are unsealed or poorly sealed.
- **Air-Source heat pumps.** This measure includes upgrades of existing air-source heat pumps and conversions from other forms of electric heating, i.e., forced-air furnaces.
- **Ground-source heat pumps.** This measure includes upgrading existing air-source heat pumps or converting other forms of electric heating to a ground-source heat pump.
- **Commissioning and controls.** This measure is for commissioning and controls when a non-program-qualifying heat pump is installed.
- **Duct sealing, commissioning, and controls.** This measure name is used when both duct sealing and commissioning and controls are all conducted at the same time.
- **Variable speed heat pumps.** This measure is new in 2012, and so was not included in this evaluation

The PTCS program has been challenging to implement, both for BPA and for BPA's customer utilities. The specifications are complicated and often viewed as onerous. For example, for measures that include duct sealing, the protocol requires a duct-blaster test before the measure installation and a post-test after installation. To account for savings differences between climate zones, building types, and baseline conditions, there are hundreds of deemed measure permutations. A primary challenge of this evaluation was to estimate savings for as many of these measure permutations as possible.

While BPA initiated the program in 2006, this is the first comprehensive impact evaluation and covers program years 2009 to 2011. The purpose of the evaluation is to measure program performance against calculated deemed savings values using a billing regression approach. The following sections present our data collection approach, data cleaning methods, billing model analysis, interim model results, and conclusions and lessons learned.

Data Collection Approach

BPA has requested billing data from utilities on a limited scale for other program evaluations, but the PTCS evaluation was unprecedented in terms of scope. A total of 69 utilities participated in this program between 2006 and 2011 with a total of 25,596 measures claimed during program years 2009, 2010, and 2011. Collecting that amount of billing data from a wide array of utilities presented a significant challenge and utilities

expressed concern well in advance of the data request. Specifically, there were concerns about producing faxed paper billing records, interpreting requested data fields, and missing key identifiers and linkages.

Extensive formal and informal outreach was undertaken to solicit utility assistance in organizing the data request. This led to restricting the evaluation period to the 2009 through 2011 program years because many utilities archive older billing records, making it difficult for utilities to access older data. In addition, many utilities told us they had switched billing systems and that linking between the new and legacy systems added an additional layer of effort we decided was not necessary for the utilities to undertake.

Feedback from utilities also informed our timeline for collecting data. We learned that some utilities were able to easily pull records and respond in as quickly as two weeks. Other utilities, however, required six to eight weeks to fulfill data requests. For this reason, we allowed three months in the project timeline for data collection. As expected, the bulk of utilities provided data within the first month. The next two months were spent working closely with the remaining utilities to obtain the requested data. After the data request was sent out, we conducted a webinar to explain the data templates, clarify expectations, and demonstrate the data submittal process. We also provided ad-hoc support to utilities if they had any problems or questions in fulfilling the data request. We were able to leverage an existing template from a similar evaluation for data collection to use for all utilities. This ensured clarity and consistency in the data for both the evaluator and utilities. The data request included a detailed data collection memo to make expectations clear and provide answers to anticipated utility questions. All of these factors contributed to a relatively smooth data collection process.

There were a total of 69 utilities with PTCS participating customers between 2009 and 2011. Many of these utilities were small and had relatively few participants and, because the associated burden on those utilities was excessive compared to the benefit of evaluating the small number of measures they claimed, we excluded utilities with 30 or fewer participants from the data request. This resulted in a total of 42 utilities ultimately included in the data request. A summary of data requested and received is shown in the Table 1 below.

Table 1: Summary of Participant Data Requested and Received

Participant Group	Number of Participants	Percent of Total
All PTCS participants between 2006 and 2011	24,305	100%
All PTCS participants between 2009 and 2011	18,195	75%
Participants selected for data request[1]	17,863	73%
Participants included in request to utilities[2]	17,858	73%
Participants received back from utilities	16,351	67%

Source: Evergreen analysis of tracking system data provided by BPA

Careful planning and the early, extensive outreach resulted in a smooth data collection effort. However, the initial data request accidentally omitted a few important measure categories, namely air source heat pump conversions and some duct sealing measures. The PTCS measures were complex, with over 400 individual measures and measure reference numbers, with no easy way to consistently identify and pull all 400 PTCS measures for this evaluation. While this has been fixed in BPA's internal tracking system to help ensure that measures do not get missed again, BPA is adding a layer of quality control in future evaluations by providing summaries of the data pulled to program staff who will be more likely to notice missed measures before the request goes to utilities.

Data Cleaning

Once all data was received from the utilities we began developing data screens to clean the billing data for analysis. It was important to remove any potentially erroneous billing data from the final modeling dataset so as to not to introduce excess noise into the model. Ultimately, the billing analysis was conducted on subset of data that had been subjected to a series of data screens. The screens used to produce this final dataset for modeling removed the following:

- Observations with monthly electricity consumption less than or equal to 0 kWh.

[1] This group includes participants between 2009 and 2011 from utilities that had over 30 participating customers between 2009 and 2011.

[2] This number is slightly lower than the above number of "participants selected for data request," due to the fact that a small number of participants had no address listed in the data tracking system, and so could not
[2] This number is slightly lower than the above number of "participants selected for data request," due to the fact that a small number of participants had no address listed in the data tracking system, and so could not be requested from the utility.

- Observations with monthly electricity consumption greater than 10,000 kWh.
- Households with average monthly electricity consumption less than 200 kWh.

A summary of these data screens is shown in Table 2. A variety of data screens were tried on the models as a sensitivity test, but none altered the results or statistical significance of the results greatly, so we opted to use these data screens, which were recommended and approved by the stakeholder group.

Table 2: Summary of Data Screens

	All Data	Data Screened	Data Remaining	Data Screen (% of total)
Individual Observations	724,997	15,082	709,915	98%
Households	15,296	2,416[3]	12,880	84%

Source: Analysis by Evergreen Economics of data provided by BPA

Stakeholder Involvement and Review Process

The PTCS program and this impact evaluation have a wide array of stakeholders integral to their success and interested in the results. Program and planning staff comprised the majority of internal stakeholders while external stakeholders included Regional Technical Forum (RTF) members, Northwest Power and Conservation Council staff, and regional PTCS experts. Some stakeholder review was built-in from the start of the evaluation, but the project scope did not anticipate the full importance and value of stakeholder involvement. During the review process, stakeholders identified issues and asked thoughtful questions that directed our analytical efforts in new and more productive ways. Specifically, stakeholders influenced the evaluation in the following ways:

- Identified concerns with wood heating, which caused us to add a heating signature analysis to our scope of work
- Caught errors in how we defined measures
- Identified inconsistencies with measure baselines
- Provided valuable input that helped direct the data collection process, making that stage of the project much smoother than it would have been otherwise

Stakeholder review was an unexpectedly critical component of this evaluation, but could have been even more effective with better planning. Stakeholder involvement would have been improved by developing a communications plan at the beginning of the project that mapped out individual stakeholders and crucial communication points. While

[3] The number of households shown here is the number affected by the data screen, not the number of households completely removed by the data screen. For example, a household that had at least one monthly observation removed by the screen is counted as "data screened" in this table, but other observations for that household may remain in the dataset.

stakeholder feedback has been valuable, it could have been better leveraged through clear touch points and consistent protocols.

Selection of a Control Group

In order to identify the electricity savings attributable to the program and not due to other outside forces (such as economic conditions), a comparable control group should be used in the model. A comparable control group for this analysis would be limited to electrically heated homes with leaky ducts. A random sample of non-participating utility customers would not yield such a control group. In addition, we were only able to collect billing data for program participants. The solution was to use 2011 participants as a control group for 2009 and 2010 program participation. This would ensure that the households were electrically heated and had similar characteristics to the 2009 and 2010 participant households.

Billing Model Analysis

The impact evaluation of energy savings was conducted by developing and estimating a gross billing model. The billing model was used to identify the extent to which each of the PTCS measures can explain differences in the energy consumption of households before and after the installation of the measures. One of the advantages of using the billing regression model is that it allows us to consider confounding factors, such as home type, geographic location, and differences in the weather between the pre and post periods. It also allowed us to achieve the evaluation goal of estimating realized impacts by measure, heating zone, and home type for most measures.

At the outset of the evaluation, we planned to estimate two types of models as part of our analysis, a pooled fixed effects model and a difference-in-differences model. Ultimately, we decided on the fixed effects model to produce savings estimates for this evaluation. However, both of types of models—along with a statistically adjusted engineering (SAE) model—are described in more detail below along with the reasons for our decision to proceed with only the fixed effects model.

Model Selection Process

In the early stages of modeling, several regression models were developed including differences-in-differences, fixed-effects, and statistically adjusted engineering (SAE) model specifications. This section describes why the fixed-effects specification was selected over the other models and how model selection was influenced by the dataset.

Fixed Effects Model

The benefit of a fixed effects model is that it controls for unique characteristics within each household, such as general levels of electricity use (i.e. a high usage or low usage household) and household occupancy, which would not otherwise be represented in the model. These sort of time-invariant characteristics are the "fixed" effects that the model controls for with a household-specific constant term. The general billing model using a fixed effects specification is given below. Variations on this model were explored

during the modeling processes, including a variety of interaction terms. The fixed effects model was estimated using the linear values of the dependent and independent variables.[4]

The model is specified as follows:

$$KWHNorm_{it} = \propto_i + \beta_1(MEAS * POST)_{it} + \beta_2 CDD_{it} + \beta_3(CDD * MEAS)_{it} + \beta_4(CDD * MEAS * POST)_{it} + \beta_5 HDD_{it} + \beta_6(HDD * MEAS)_{it} + \beta_7(HDD * MEAS * POST)_{it} + \epsilon_{it}$$

Where:

$KWHNorm$ = Household energy usage (KWH) normalized to a monthly value

$MEAS * POST$ = Interaction term between the indicator for post-installation observations and measure type indicator

CDD = Average cooling degree-days per month

$CDD * MEAS$ = Interaction between cooling degree-days and measure type indicator

$CDD * MEAS * POST$ = Interaction between cooling degree-days, measure type indicator and post-period indicator

HDD = Average heating degree-days per month

$HDD * MEAS$ = Interaction between heating degree-days and measure type indicator

$HDD * MEAS * POST$ = Interaction between heating degree-days, measure type indicator and post-period indicator

i = Index for household (i = 1,2, …, n)

t = Index for monthly time period (t = 1,2, …, T)

$\propto$ = Household-specific constant

$[\beta_1, \ldots, \beta_6]$ = Coefficients to be estimated in the model

ε = Random error term, assumed to be normally distributed

Difference-in-Differences Model

In addition to the fixed effects model described above, we also estimated a difference-in-differences model as an alternate method of estimating program savings using regression analysis. As mentioned earlier, a non-participant control group was not available, so 2011 participants were used as a proxy control group. In a difference-in-

[4] As opposed to the alternative of first transforming the dependent variable and/or the independent variables by the natural log function.

differences model, the estimated change in energy usage between the pre and post periods for the control group reveals the extent to which external factors that affect all households, such as economic forces, affect energy usage. The estimated change in energy usage between the pre and post periods for participants captures the effect of changes due to program measures as well as changes due to the external factors that all households face. Therefore, by comparing the pre and post period differences in energy use between the control group and the participant group we would difference out the savings attributable to the program (hence the "difference-in-differences" name). In theory, the results of this difference-in-differences model should provide a good comparison with the fixed effects model results.

In practice, the difference-in-differences approach works best when there is a true non-participant control group in order to isolate the effects of program measures on energy savings from changes in energy savings due to external factors. This was a major shortcoming of the difference-in-differences model for the purposes of this analysis. Another shortcoming of this modeling approach was that we were only able to use 2008 (pre-installation) and 2011 (post-installation) data, rather than using all years as we are able to do in a fixed effects model. Difference-in-differences requires a balanced and distinct pre-period and post-period for all households in the model, which required that we eliminated approximately one-half of the data points.[5] For these reasons we opted to use a fixed effects modeling approach for this analysis.

Statistically Adjusted Engineering (SAE) Model

An SAE model was estimated with the same specification of the fixed effects model above, except deemed savings values for each measure were used in place of measure installation indicators. The idea behind this type of model is that the coefficient estimates on the deemed savings variables will give a realization rate, or the percent of deemed savings that is actually being achieved by that measure. However, this requires that the deemed savings values be accurate and that baseline assumptions be in line with the type of equipment actually replaced by customers in the data. Additionally, the deemed savings values in the model may pick up some of the fixed effects that the household-specific constants are controlling for, thereby confounding the model results. Again, we decided to forgo the use of this alternative model in favor of the fixed effects model described above.

Modeling Considerations

Measure Baselines

After discussing draft model results with stakeholders and BPA program staff, we discovered that baseline assumptions for the deemed savings were different than the actual baseline (i.e. replaced equipment) for air source heat pump upgrades. This being the case, comparing our model results to the deemed values was not appropriate.

[5] Since the analysis was conducted on 2009 and 2010 participants, measures were being installed continuously over those two years. The solution was to use 2008 as the pre-period, before any 2009/2010 participants had started installing measures, and use 2011 as the post-period, after any 2009/2010 participants had finished installing measures.

Ultimately, the decision was made to drop air source heat pump upgrades from the models due to inconsistent baselines with the deemed savings value.

The commissioning and controls measure was also excluded from the analysis for a similar reason. In early model runs the coefficient for this measure was picking up additional savings due to a non-PTCS heat pump being installed along with the PTCS commissioning and controls. We did not have enough information to control for the non-PTCS heat pump in the model, which caused the estimated savings for commissioning and controls to be much higher than it should have been. Since we could not accurately estimate savings for this measure it was removed from the analysis.

Another area where we found that baseline assumptions may not be consistent with reality was for any measure that includes duct sealing. Deemed savings assumptions established by the Regional Technical Forum about the leakiness of ducts may be different than the actual leakiness of ducts found in the participant data. As this factor arose late in the evaluation, we are now conducting analysis to compare average leakage flow test results from the pre-period and the post-period for duct sealing jobs in the program to the deemed measure assumptions.

Confounding Factors

In addition to the measures discussed above for which savings could not be properly estimated due to inconsistent baselines, we theorize that the use of wood heat may be confounding our results. In some of the regions included in this study, where winters are very cold, the use of wood heat is a relatively common practice. This confuses the model since a participant who had received a heat pump through the program may have been primarily using a wood stove to heat their home prior to program participation. After the new heat pump is installed, they may begin to use more electricity than before participation because they now rely less on the wood stove to heat their home. This is likely a common issue in the heating zone covering areas of Idaho and Montana, and could contradict the energy savings assumptions for these measures. This geographical area is also a less populated region and had fewer participants than the other heating zones, making it difficult to get statistically significant savings estimates for this region. To test the theory of wood heat use, we are about to begin an analysis of the heating signatures of the homes in order to investigate if this explains the lack of precision and the low realization rates we see in some of the heating zones.

Interim Model Results

This section presents the results of our interim regression models, which produced savings estimates only for the duct sealing and duct sealing, commissioning, and controls measures. All other program measures were excluded from these interim models due to inconsistent baselines (air source heat pump upgrades and commissioning and controls) or insufficient data (ground source heat pump upgrades and air source heat pump conversions). In the second round of modeling we will use data on air source heat pump conversions and plan to estimate savings for that measure as well. This second round of models will be estimated this summer, with final models and results published in the evaluation report at the end of August. If timing allows, we will present an overview of final model results as part of our conference presentation.

Models were estimated separately by home type (single family or mobile home) and heating zone, as well as a model with all home types and heating zones together. Table 3 and Table 4 below show the results of the duct sealing and the duct sealing, commissioning, and controls models that were estimated for all home types and heating zones combined. For these interim models we estimated separate models for these two measure types, but plan to try a combined model for all measure types in the final modeling round.

Table 3: Interim Model Results – Duct Sealing

Variable Name	Coefficient Estimate	Standard Error	b/Std.Er.	p-value
DuctSeal_post	-95.07	4.85	-19.61	<1%
CDD	2.01	0.04	49.87	<1%
DuctSeal_CDD	0.06	0.06	1.03	30%
DuctSeal_post_CDD	1.01	0.07	14.14	<1%
HDD	1.82	0.01	272.73	<1%
DuctSeal_HDD	0.11	0.01	12.68	<1%
DuctSeal_post_HDD	0.04	0.01	5.29	<1%

Table 4: Interim Model Results - Duct Sealing, Commissioning, and Controls

Variable Name	Coefficient Estimate	Standard Error	b/Std.Er.	p-value
DuctSealCC_post	-68.72	26.71	-2.57	1%
CDD	2.01	0.04	49.87	<1%
DuctSealCC_CDD	-0.06	0.17	-0.35	73%
DuctSealCC_post_CDD	0.24	0.25	0.96	33%
HDD	1.82	0.01	272.73	<1%
DuctSealCC_HDD	-0.09	0.04	-2.43	2%
DuctSealCC_post_HDD	-0.07	0.05	-1.48	14%

Post-model calculations were done to arrive at an annual savings estimate for each measure. These calculations scale the estimated monthly savings up to an annual value while taking into account weather during the modeled period. The savings estimates, associated confidence intervals, and realization rates resulting from these calculations are presented below in Table 5.

Table 5: Savings Estimates and Confidence Intervals

Measure Name	Annual Savings Estimate (kWh/year)	Lower Bound of 95% Confidence Interval	Upper Bound of 95% Confidence Interval	Average Deemed Savings	Realization Rate
Duct Sealing	672	622	722	877	77%
Duct Sealing, Commissioning, and Controls	1,179	886	1,472	1,675	70%
Overall					**76%**

In these interim models, duct sealing was found to have a realization rate of 77 percent, while the duct sealing, commissioning, and controls measure was found to have a realization rate of 70 percent. To arrive at these realization rates, estimated savings were compared to deemed savings values approved by the RTF. It is possible that the duct leakage assumptions used in calculation of the deemed savings value are different than the actual amount of duct leakage found in participant homes, which may partially explain the realization rates being less than 100 percent. We are currently looking into how the actual duct leakage for these households compares to the deemed measure assumptions, and findings will be included in the final report this summer.

Conclusions and Lessons Learned

Despite the many challenges faced in conducting this impact evaluation, we had many successes due to careful planning and coordination with utilities, RTF staff, and other PTCS stakeholders. We found that when conducting impact evaluations across multiple utility territories, logistics and communications are as important as the technical model specifications. We did a fair amount of coordination to get utility input up front, but could have done more to incorporate stakeholder knowledge from the very beginning. Clarifying at the front-end of the evaluation the stakeholders and their roles in the evaluation will help ensure that the evaluation can benefit from the insight and wisdom of experts.

One of the major successes of this project was a smooth data collection process that included over 40 utilities. The logistics that made this task successful included gathering feedback from a small number of utilities on what was reasonable to request, leveraging an existing data template to ensure clarity and consistency in the data, providing a detailed data collection memo to make expectations clear, and allowing sufficient time for all utilities to meet the request. Using a clear, simple data collection template helped reduce confusion for the evaluators and for the utilities, as the data requested were clear for both parties. In addition, it is important to expect that data

collection will take a long time; when working with multiple parties allowing a cushion of time will ensure that inevitable snags won't slow down the final deliverables.

The major lesson learned in this evaluation is to conduct quality control of the program data that is being used to inform the utility billing data request. Unfortunately, some measures were missed in the first request and we are currently in the process of collecting a second round of data from many of the same utilities. This has delayed deliverables and put additional burden on utilities to fulfill the request again. When requesting data from multiple utilities, instituting a thorough review process before sending out the request is a valuable step that could catch errors. When working with a single utility, this step is nice to have, but when working with multiple parties this step will avoid requiring a lengthy follow-up request if data are not included in the initial request.

On the technical side, modeling efforts will be greatly improved if the evaluators have a complete understanding of the intricacies of the program. In this case, there were issues pertaining to deemed measure baselines and possible use of wood heat that we were not aware of until late in the modeling process. Being aware of potential issues like these up front can help inform the modeling approach and make the results more valuable within the context of the program. It can be difficult to know the right questions to ask, but getting stakeholders involved upfront can reveal some of this more intricate information.