Report: Empirical Project

On Evaluation of Regional Energy Retrofit Programs (2018-2023)

Duc-Anh Nguyen

Munich, February 9th, 2026

Abstract

In 2019 different regional governments implemented incentive programs to encourage energy efficiency retrofits (e.g., installing efficient heating systems) for homeowners whose properties are at least 10 years old. This analysis should evaluate the effectiveness of the programs. To address the challenges, a **Difference-in-Differences (DiD)** (see Baker et al. (2025)) framework enhanced by **Double Machine Learning (DML)** (see Chernozhukov et al. (2018)) are employed. This approach allows to control for a high-dimensional set of potential confounders—including building characteristics, local amenities, and commuting zones—without making strong functional form assumptions. An **Instrumental Variable (IV)** (see Wu et al. (2022)) strategy is also employed to recover the structural effect of the retrofits themselves, using random program assignment as an instrument for the decision to retrofit.

The analysis yields three main findings: First, financial incentives are highly effective, while information campaigns are not. Second, the mechanism driving these results is the adoption of retrofitting technology. Third, the policy is most effective for **older buildings** (pre-2000), where the marginal energy savings are approximately 20% higher than in newer housing stock. These findings suggest that future policy should **prioritize financial subsidies over information campaigns** and should be targeted specifically toward the **aging housing stock** to maximize the return on public investment.

Contents

1 Motivation, Research Questions and Data

1.1 Motivation and research questions

Improving residnetial energy efficiency is crucial for global climate policy. In 2019, different regional governments implemented incentive programs to encourage homeowners to retrofit their properties being more than 10 years old. They first announced and discussed about the programs among policy-makers and in the media in the same year, having no anticipation or public discussion before 2019. By the end of 2019, all program regions have had their programs fully operational.

There are two distinct policy experiments, which devided all program regions into 3 regions regarding to the policy (see Instructions Empirical Project (Resit) (2026)):

- Program "A" regions: Househholds from several regions received strong financial incentives, including generous subsidies and tax credits for households when implementing comprehensive energy retrofits to their homes.
- Program "B" regions: Households from these two regions received only information campaigns and small rebates for this implementation (due budget limitations).
- Control "C" regions: Households in these control regions may still undertake retrofits, but without subsidies.

Research Questions (see Instructions Empirical Project (Resit) (2026)): Evaluating such programs is empirically challenging. Households that participate in energy programs often differ systematically from those that do not—they may be wealthier, more environmentally conscious, or live in different types of buildings. A simple comparison of means would therefore yield biased estimates of the program's effectiveness. Furthermore, to design optimal future policies, policymakers need to know not just whether a program works, but why (the mechanism) and for whom (heterogeneity). The big question is then devided into four tasks, in each of them they wish to produce credible estimates of:
Task 1, The causal effect of each program (Program A and Program B) on energy consump tion and energy costs
Task 2, The causal effect of each program (Program A and Program B) on health
Task 3, The causal effect of retrofits themselves on energy consumption and energy costs
Task 4, Investigation: for which type of households the causal effect of the programs on energy consumption and energy costs was the smallest and largest (which type of households responded the most to the programs).

1.2 Data Preparation

The analysis utilizes a panel dataset (see Instructions Empirical Project (Resit) (2026)): a random sample of households from all regions, observed over three periods: a pre-treatment baseline ($t = 2018$) and two post-treatment follow-ups ($t = 2020$ and $t = 2023$). Only households who did not move during the time period are observed. All households are eligible for the Programs A and B, but not all households participate in the programs.

The dataset contains detailed information on energy usage, building characteristics, and socio-demographic factors for households across multiple regions. For the mechanism analysis, I defined a binary variable, *Retrofit Status* (D_i), which takes a value of 1 if the household implemented any type of energy efficiency upgrade (insulation, windows, or heating) between 2018 and 2023, and 0 otherwise.

Data Cleaning and Transformation

To prepare the data for the Difference-in-Differencesv(see Baker et al. (2025)) and Double Machine Learning (DML) (see Chernozhukov et al. (2018)) estimation, I performed the following transformations:

1. **First-Differencing:** To eliminate time-invariant household heterogeneity (fixed effects), I transformed all outcome variables into first differences relative to the baseline year ($t = 2018$). While calculating differences for both post-treatment periods (2020 and 2023), the analysis focuses primarily on the long-term effects observed in 2023 ($t = 2023$):

   $$\Delta Y_i = Y_{i,2023} - Y_{i,2018}$$

   Focusing on the 2023 endpoint allows sufficient time for the retrofits to be fully installed and for their energy-saving potential to be realized, providing a more robust estimate of program effectiveness. This ensures that our estimates capture the *change* in consumption attributed to the programs.

2. **Baseline Controls (X_i):** I constructed a high-dimensional matrix of control variables using 2018 values. This includes building age, floor area, household income, and over 100 binary indicators for local amenities and commuting zones. Using baseline values ensures that the controls are not ”bad controls” (outcomes of the treatment itself).

3. **Handling Missing Values:** Missing values in time-invariant building characteristics were imputed using the available 2018 data to maximise the effective sample size for the ML algorithms.

2 Analysis: Methodology

To estimate the causal effect of Program A (incentives) and Program B (information) on our outcomes of interest ($Y \in \{Energy, Cost, Health\}$), we face a high-dimensional confounding problem: the dataset contains over 100 control variables (X_i), including building characteristics, local amenities, and commuting zones. Standard OLS would overfit, while naive selection (e.g., single Lasso) would suffer from regularization bias. To address these challenges, according to the Frisch-Waugh-Lovell theorem (see Frisch and Waugh (1933)), we can regress the residuals of the outcome on the residuals of the treatment.

In a modern high-dimensional setting, this procedure is validated by the Double Machine Learning framework developed by Chernozhukov et al. (2018). I therefore employed a **Difference-in-Differences (DiD)** (see Baker et al. (2025)) framework enhanced by **Double Machine Learning (DML)** (see Chernozhukov et al. (2018)). This approach allows us to control for a high-dimensional set of potential confounders—including building characteristics, local amenities, and commuting zones—**without making strong functional form assumptions**. I defined the Intent-to-Treat (ITT) effect β using the partially linear model: $\Delta Y_i = \beta Z_i + g(X_i) + \varepsilon_i$ and $Z_i = m(X_i) + \nu_i$ where ΔY_i is the change in the outcome (2023 vs. 2018) and Z_i is the program assignment. I proceeded in two steps:

1. **Nuisance Parameter Estimation:** I used machine learning algorithms to learn the conditional expectations $\hat{E}[\Delta Y|X]$ and $\hat{E}[Z|X]$. For our main specification, I used **Lasso** with λ selected via 10-fold cross-validation. To assess robustness (Sensitivity Analysis), I also employed **Random Forests** to capture non-linearities.

2. **Orthogonalized Regression:** I regressed the residuals of the outcome ($\tilde{Y}$) on the residuals of the treatment ($\tilde{Z}$). This **partialling out** strategy satisfies *Neyman Orthogonality*, ensuring that our estimate of β is $\sqrt{N}$-consistent and unbiased even if the machine learning step makes small errors.

2.1 Task 3: The Causal Effect of Retrofits (IV Strategy)

The ITT estimates from Task 1 capture the effect of the porgrams. However, to understand the structural impact of the technology itself, we must estimate the effect of the actual retrofit (D_i). Since retrofitting is endogenous (self-selected), I used an **Instrumental Variable (IV)** (see Wu et al. (2022)) strategy within the DML (see Chernozhukov et al. (2018)) framework, to recover the structural effect of the retrofits themselves, using **random program assignment** as an instrument for the decision to retrofit, as I had an hypothesis that rich/motivated people self-select into retrofits.

I used the assignment to Program A (Z_i) as an instrument for the retrofit decision (D_i). Our identification relies on two assumptions:

1. **Relevance:** The financial incentives must significantly increase retrofit rates ($Cov(Z, D) \neq 0$). We verify this via the First Stage F-statistic.

2. **Exclusion Restriction:** The program assignment Z_i affects energy consumption *only* through the retrofit D_i. Given that Program A consists of physical subsidies rather than behavioral nudges, I argued this assumption holds.

I estimated this using DML-IV, where I partialed out X from the instrument Z, the treatment D, and the outcome Y, and then perform Two-Stage Least Squares (2SLS) on the residuals.

2.2 Task 4: Heterogeneity (Who responds the most/least?)

To investigate which households responded most strongly to the measure, we move beyond average effects: I employed **Causal Forests** (Athey and Wager (2019)), a generalized random forest method designed to estimate the Conditional Average Treatment Effect (CATE):

$$\tau(x) = E[Y_i(1) - Y_i(0) \mid X_i = x]$$

Unlike standard interaction terms in OLS, Causal Forests allow the data to discover the relevant heterogeneity in a data-driven way. The algorithm modifies the splitting criterion of regression trees to maximize the variance of the estimated treatment effect across leaves, rather than minimizing prediction error.

After training the Causal Forest, I analysed the **Variable Importance** scores to identify the key drivers of heterogeneity (e.g., Building Age, Income) and computed the average treatment effects for specific subgroups defined by these drivers. Table 1 presents the estimated Average Treatment Effects (ATE) (see Sakaguchi (2020)) of the two energy efficiency programs. The two columns display the Intent-to-Treat (ITT) (see Gupta (2011)) estimates, representing the causal effect of being offered the program regardless of uptake.

3 Analysis results

3.1 Analysis results of task 1: on Energy Consumption and Costs

The results indicate a sharp divergence in the effectiveness of the two policy approaches. As shown in Panel A Table 1, households in **Program A** regions (financial incentives) reduced their monthly energy consumption by **99.86 kWh** on average ($\hat{\beta}_A = -99.86, SE = 7.14$)(a 13% reduction relative to the baseline). This result is statistically significant at the 1% level. Given the average baseline consumption in control regions, this corresponds to a **substantial reduction in energy usage**.

Financially, this translates to a reduction in monthly energy costs of **49.74 EUR** ($\hat{\beta}_A = -49.74, SE = 3.04$). It is worth noting the internal consistency of these estimates: the ratio of cost savings to energy savings ($\approx 49/100$) implies an effective electricity price of roughly **0.49 EUR per kWh**. This figure aligns with residential energy prices during the energy crisis period (2022-2023), lending credibility to the data quality and our estimation strategy. On an annualized basis, Program A saves the average household approximately **600 EUR per year**, a magnitude large enough to justify the transaction costs associated with retrofitting.
In contrast, **Program B** (information campaigns) failed to produce statistically distinguishable effects. The coefficient for Program B is positive but insignificant (14.41 kWh, $p > 0.10$), and the effect on costs is effectively zero. This suggests that information constraints were not the primary barrier to energy efficiency in this context; rather, financial constraints or the capital cost of retrofitting appear to be the binding factor.

	Intent-to-Treat (ITT) Estimates	
Outcome Variable ($\Delta Y_{2023-2018}$)	**Program A** (Incentives)	**Program B** (Information)
Panel A: Energy Outcomes (Monthly)		
Energy Consumption (kWh)	-99.86*** (7.14)	14.41 (9.66)
Energy Costs (EUR)	-49.74*** (3.04)	1.78 (3.74)
Panel B: Health Outcomes (Annual)		
Health Issues (Count)	-0.166** (0.064)	-0.053 (0.089)
Sample Information		
Observations (N)	2,856	2,270
Treated Clusters (Regions)	Multiple	2

Table 1: Estimated Average Treatment Effects (ATE) of Energy Programs

Notes: Estimates obtained using Double Machine Learning (DML) with Lasso for nuisance parameter estimation. Outcomes are first-differenced ($\Delta Y_{2023} - Y_{2018}$). Robust standard errors are reported in parentheses. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

We note that the sample for Program B is drawn from only two regions ($n = 355$), whereas Program A covers several regions ($n = 941$) (Note that here number of obervations $N = ProgramA(941) + Control(1,915) \approx \mathbf{2,856}$ and for program B $N = ProgramB(355) + Control(1,915) \approx \mathbf{2,270}$). A small number of treated clusters can sometimes lead to **downward-biased standard errors** if **intra-regional correlation is high**. However, since we find **no statistically significant effect** for Program B even with potentially understated standard errors, our conclusion that information campaigns were ineffective is **robust**. Furthermore, my DML specification includes detailed commuting zone and amenity controls, which absorb much of the region-specific variation that typically causes clustering bias.

3.2 Analysis results of task 2: on Health Outcomes

The results for health outcomes (Table 1, Panel B) indicate that households in Program A regions saw an average reduction of 0.166 cold-related health issues per year ($p < 0.05$). While small in absolute terms, this represents a meaningful improvement in public health outcomes. This finding suggests that the comprehensive retrofits incentivised by Program A (such as improved insulation and heating) improved the indoor living environment, leading to better public health outcomes. In contrast, Program B had no statistically significant impact on energy use or costs.

Estimator	**Main Specification** (DML-Lasso)	**Robustness Check** (DML-Random Forest)
Treatment Effect (kWh)	**-99.86*****	**-112.46*****
Standard Error	(7.14)	(7.71)
Model Features	Linear Selection	Non-Linear Interactions
Controls Included	Yes (> 100)	Yes (> 100)

Table 2: Sensitivity Analysis: Robustness to Functional Form (Program A)

Notes: Comparison of Average Treatment Effects (ATE) for Program A on monthly energy consumption. The main specification uses Lasso to partial out controls; the robustness check uses Random Forests to account for potential non-linearities and interactions between household characteristics. Significance levels: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Crucially, Program B showed no significant effect on health. This supports the interpretation that Program A improved health by physically altering the housing environment (e.g., better insulation leading to higher indoor temperatures and lower humidity), whereas Program B, which did not induce physical retrofits, yielded no such secondary benefits.

Sensitivity Analysis:
My main estimates relied on Lasso to partial out the high-dimensional controls. To test whether our findings are driven by the linearity assumptions inherent in the Lasso, I conducted a sensitivity analysis using **Random Forests** (see Breiman (2001)) for the nuisance parameter estimation ($E[Y|X]$ and $E[D|X]$). Random Forests allow for complex, non-linear interactions between building age, climate, and income.

The DML-Random Forest estimator yields a treatment effect for Program A of **-112.46 kWh** ($SE = 7.71$) (see Table 2). This estimate is statistically indistinguishable from our main Lasso estimate (-99.86 kWh). The robustness of the result across different machine learning learners suggests that the finding is not an artifact of functional form assumptions but reflects a genuine causal relationship.

3.3 Analysis results of task 3: IV Estimates: effects of retrofit adoption

Having established the average effectiveness of Program A, we now turn to understanding the mechanism driving these results and identifying which types of households benefited the most.

The Intent-to-Treat (ITT) (see Gupta (2011)) estimates in the previous subchapters capture the effect of the *policy offer*. However, from an engineering and structural perspective,

Estimator / Outcome	**Estimate**
First Stage (Compliance)	
Effect of Program A on Retrofit Probability	0.476*** (0.025)
F-Statistic (Instrument Strength)	*368.8*
Second Stage (LATE)	
Effect of Retrofit on **Energy Consumption** (kWh)	-208.30*** (14.91)
Effect of Retrofit on **Energy Costs** (EUR)	-103.50*** (6.36)
Observations (Program A vs Control)	2,856

Table 3: Mechanism: The Causal Effect of Retrofitting (IV Estimates)

Notes: IV estimates using assignment to Program A as an instrument for the decision to retrofit. The First Stage reports the compliance rate (increase in retrofit probability due to the subsidy). The Second Stage reports the Local Average Treatment Effect (LATE), interpreted as the monthly savings for a household induced to retrofit by the program. Standard errors in parentheses. $^{***}p < 0.01$.

policymakers need to know the energy savings generated by the actual *retrofit*.

To recover this parameter, we use the assignment to Program A as an instrumental variable (IV) for the decision to retrofit. The validity of this strategy relies on the strength of the first stage: the mechanism driving these results is the adoption of retrofitting technology. The IV estimates reveal the Local Average Treatment Effect (LATE) (see Martin and Kaspar (2019)) - the effect of retrofitting for those households induced to do so by the subsidy. We find that undertaking a retrofit reduces monthly energy consumption by **208.30 kWh** ($\hat{\beta}_{IV} = -208.30, SE = 14.91$) and monthly energy costs by **103.50 EUR**. The discrepancy between the program effect (ca. 100 kWh) and the retrofit effect (210 kWh) is explained by the compliance rate: only 47.6% of households in Program A regions were induced to retrofit, as reported in Table 3 (from code: $t = 19.05$, $F > 300$). This indicates a strong instrument, mitigating concerns about weak instrument bias.

Internal Validity Check: A crucial test of our model's internal consistency is the relationship between the ITT, LATE, and compliance rate. Theoretically, $ITT \approx LATE \times Compliance$. Our empirical results satisfy this identity almost perfectly:

$$-208.30(LATE) \times 0.476(Compliance) \approx -99.15kWh$$

This is virtually identical to our estimated ITT of -99.86 kWh. This consistency confirms that the reduction in energy consumption is driven entirely by the adoption of retrofits,

Subgroup (Split by Median)	**Mean Treatment Effect** (kWh Savings)	**% Diff from Average** (Baseline: -106.21)
Panel A: By Building Age		
Older Buildings (> Median)	**-115.67**	**+8.9%**
Newer Buildings (≤ Median)	-97.07	-8.6%
Panel B: By Household Income		
Lower Income (≤ Median)	-106.93	+0.7%
Higher Income (> Median)	-105.49	-0.7%

Table 4: Heterogeneity of Program A Effects (Causal Forest)

Notes: Estimates derived from Causal Forest predictions. Subgroups defined by the median of the respective variable. "Mean Treatment Effect" is the average estimated reduction in monthly energy consumption for that group. % Diff compares the group's savings to the global average savings (-106.21 kWh). Positive % indicates "better" (larger) savings than average.

rather than by other behavioral changes (e.g., households in Program A regions simply turning down thermostats to qualify for rebates).

3.4 Analysis results of task 4: Heterogeneity: Who benefits most?

I employed Causal Forests (see Athey and Wager (2019)) to test for heterogeneous treatment effects and identify the key drivers of this variation: the calibration test for the Causal Forest rejects the null hypothesis of constant treatment effects ($p < 0.001$), confirming that the program impacts some households significantly more than others. The variable importance analysis identifies **Building Age** as the single most important predictor of treatment effect magnitude, followed by **Income**.

As shown in Table 4 Panel A, the program is substantially more effective for older buildings. Households in the older half of the building stock save an average of **115.67 kWh** per month, compared to 97.07 kWh for newer buildings. This result is intuitive: older buildings likely suffer from poorer baseline insulation and outdated heating systems, meaning the marginal productivity of a retrofit is higher.

I also examined **heterogeneity by Income**: from Table 4 Panel B: while lower-income households save slightly more (106.93 kWh) than higher-income households (105.49 kWh), **the gradient is much flatter than for Building Age**. This suggests that while equity concerns are valid, targeting based on **physical** characteristics (**Building Age**) will yield larger aggregate energy savings than targeting based on *socio-economic* characteristics alone.

4 Conclusion/ Policy recommendation

This report evaluated the causal impact of two regional energy efficiency programs implemented in 2019 using a panel of households from 2018 to 2023. By combining a Difference-in-Differences framework with Double Machine Learning (DML) and Instrumental Variables (IV), we were able to isolate the causal effects of the policies while controlling for high-dimensional confounding factors. My analysis leads to **three principal conclusions**:
First: **the financial incentives are a necessary condition for behavioral change**: program A, which offered subsidies, significantly reduced energy consumption by approximately **100 kWh per month** and lowered energy costs by ca. **49 EUR per month**, which translates to an annual saving of nearly 600 EUR per household; in contrast, Program B, which relied solely on information campaigns, had no statistically significant impact on energy use or health outcomes. This suggests that information constraints were not the primary barrier to energy efficiency in this context.
Second, **the structural impact of retrofitting is profound**: using program assignment as an instrument, we estimate that the physical act of retrofitting reduces monthly energy consumption by **210 kWh**. The gap between the program effect (-100 kWh) and the retrofit effect (-210 kWh) is explained by the compliance rate: only 47.6% of households in Program A regions actually undertook the retrofits.
Third, **targeting is key to efficiency**: my heterogeneity analysis using Causal Forests reveals that the policy was not equally effective across the board: households in **older buildings** (pre-median age) achieved energy savings approximately **20% larger** than those in newer buildings (-115.67 kWh vs. -97.07 kWh).

Limitations:
While our results are robust to high-dimensional observed confounding, they rely on the assumption of **Selection on Observables**. If there are unobserved regional characteristics (e.g., unmeasured local political culture or non-program related infrastructure changes) that drive both program implementation and energy savings, our ITT estimates could remain biased. Furthermore, regarding Program B, we note that the absence of a statistically significant effect should be interpreted with caution given the small number of treated regions ($N = 2$), which may limit the generalizability of this specific null result. Finally, our analysis of the 2023 data captures the medium-term effect (3-4 years); the lifetime return on investment for retrofits may be even higher than estimated here.

Policy Recommendations:
Based on these findings, it is recommended that regional governments **discontinue the information-only approach** (Program B), as it yields no measurable benefits. Future resources should be consolidated into financial incentive schemes similar to Program A. However, to maximize the return on public investment, these subsidies should be targeted specifically toward the aging housing stock, where the marginal impact on energy efficiency is highest.

A Electronic appendix

My code can be found on my github repo: https://github.com/DucAnhValentinoNguyen/Assessing-causal-effect-of-program-on-energy-consumption-and-costs/tree/main

References

Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application.
URL: *https://arxiv.org/abs/1902.07409*

Baker, A., Callaway, B., Cunningham, S., Goodman-Bacon, A. and Sant'Anna, P. H. C. (2025). Difference-in-differences designs: A practitioner's guide.
URL: *https://arxiv.org/abs/2503.13323*

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.
URL: *https://link.springer.com/article/10.1023/A:1010933404324*

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters, *The Econometrics Journal* **21**(1): C1–C68.

Frisch, R. and Waugh, F. V. (1933). Statistical determinants of net regression coefficients, *Econometrica* **1**(4): 387–401.

Gupta, S. K. (2011). Intention-to-treat concept: A review, *Perspectives in Clinical Research* **2**(3): 109–112.
URL: *https://journals.lww.com/picp/fulltext/2011/02030/intention_to_treat_concept_a_review.9.aspx*

Instructions Empirical Project (Resit) (2026). Empirical project instructions (resit): Machine learning in econometrics, Ludwig-Maximilians-Universität München. Course Material, Moodle ID 40471.
URL: *https://moodle.lmu.de/course/view.php?id=40471*

Martin, H. and Kaspar, W. (2019). Local average and quantile treatment effects under endogeneity: A review, *Journal of Econometric Methods* **8**(1): 1–27.
URL: *https://ideas.repec.org/a/bpj/jecome/v8y2019i1p27n6.html*

Sakaguchi, S. (2020). Estimation of average treatment effects using panel data when treatment effect heterogeneity depends on unobserved fixed effects, *Journal of Applied Econometrics* **35**(3): 315–327.
URL: *https://ideas.repec.org/a/wly/japmet/v35y2020i3p315-327.html*

Wu, A., Kuang, K., Xiong, R. and Wu, F. (2022). Instrumental variables in causal inference and machine learning: A survey.
URL: *https://arxiv.org/abs/2212.05778*