README: Replication package for "Decomposing the Wedge Between Projected and Realized Returns in Energy Efficiency Programs"

Peter Christensen, Paul Francisco, Erica Myers, and Mateus Souza

The Review of Economics and Statistics

Overview

The code in this replication package generates the 5 figures and 4 tables in the paper. Code for all tables and figures from the Appendix are also provided, with the exception of the map from Figure A.1, which has been omitted to preserve the anonymity of subjects.

Data for this paper has been obtained under confidentiality agreements thus cannot be publicly shared. This paper has been exempted from the journal's data sharing requirement. To obtain permissions to access the data, replicators should submit a FOIA request to the Office of Community Assistance from the Illinois Department of Commerce and Economic Opportunity: https://www2.illinois.gov/dceo/CommunityServices/HomeWeatherization/Pages/default.aspx

Data Availability and Provenance Statements

Summary of Availability

- ☐ All data **are** publicly available.
- ☐ Some data **cannot be made** publicly available.
- ☒ **No data can be made** publicly available.

Details on each Data Source

- Data/ihwap_anonym.dta:

  Household-by-month energy consumption data merged with household-level program administrative data. Energy data provided by Ameren Illinois. Program administrative data provided by the Illinois Department of Commerce & Economic Opportunity (DCEO).

- Data/ihwap_state.dta:

  Household-level program administrative data provided by DCEO.

- Data/ihwap_degreedays.dta:

  Household-by-month weather data. Matched using data from the PRISM Climate Groups: http://prism.oregonstate.edu . The code for this matching process has been omitted to guarantee anonymity of program participants.

- Data/wx_weather.dta:

  Household-level climate data used in WeatherWorks, provided by DCEO.

- Data/furn_replace.dta:

  Household-level indicators for furnace replacements, and data on pre-treatment furnace size, provided by DCEO.

- Data/BBCorrection_UtilityData.dta and retail_final.dta:

  Auxiliary data for computing social cost of electricity, provided directly by Borenstein and Bushnell (2018): https://www.nber.org/papers/w24756 .

- Data/retail_gas.xls:

  Monthly Illinois Price of Natural Gas Delivered to Residential Consumers. Provided by EIA.

- Data/citygate_gas.xls:

  Monthly Natural Gas Citygate Price in Illinois. Provided by EIA.

- Data/electricity_prices.csv:

  Monthly average retail price of electricity in Illinois. Provided by EIA.

- Results/Model_Outputs/prism_gas.dta:

  Results from home-specific PRISM analyses. Results provided directly because this step is computationally demanding (could take over 200 hours). Replication code is provided.

- Results/Model_Outputs/predictpre_results_xgboost.csv, predictpre_results_rf.csv and predictpre_results_elastnet.csv:

  Results from machine learning hyperparameter tuning. Results provided directly because this step is computationally demanding (could take over 50 hours). Replication code is provided.

- Results/Model_Outputs/ihwap_wedge.dta:

  Main dataset for the study. Combines ihwap_anonym.dta with bootstrap results from the machine learning step, plus results from the home-specific PRISM analyses.

Computational Requirements

Software Requirements

- Stata MP, 64-bit, version 16.1
  - Required packages: `todate, winsor2, reghdfe, ftools, moremata, ds3, ivreghdfe, ivreg2, ranktest, rscript, labutil, distinct, psacalc, st0085_2, gr0059_1, st0043_2, st0035_1, rdrobust`
  - The replication package includes a folder (/Code/StataPackages) containing the specific versions of the required Stata packages.
  - The program "/Code/install_packages_Stata.do" will install all Stata packages locally, and should be run once.
  - `Rscript` allows the user to run R scripts from within Stata.
- R version 3.6.3 (2020-02-29), Nickname Holding the Windsock
  - Required packages: `"nnls", "SuperLearner", "dplyr", "arm", "onehot", "grf", "xgboost", "caret", "data.table", "knitr", "kableExtra", "magrittr", "kimisc", "data.table", "readstata13", "DescTools", "scales", "ggplot2"`
  - The replication package includes a folder (/Code/RPackages) containing the specific versions of the required R packages.
  - The program "/Code/install_packages_R.R" will install all R packages locally, and should be run once. It can be run from within Stata "/Code/main.do" line 60.

Memory and Runtime Requirements

The replication package was last run on a Dell Precision 3440 with the following configurations:

- Processor: Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz 2.90 GHz; 8-core
- RAM: 64,0 GB (63,7 GB usable)
- OS: Windows 10 Pro; Version 20H2; Build 19042.985; Windows Feature Experience Pack 120.2212.2020.0

With the above configurations, and after running the machine learning portions, the replicator should expect the code to run for about 20 hours. The machine learning portions demand significantly more compute time, as described below.

The replication package initially requires about 12 GB of hard-drive space. After replication of all the results, the complete folder will be of size 26 GB.

The machine learning portions of the analyses (hyperparameter tuning and bootstrapping), and the home-specific PRISM analyses were performed within a high-performance computing cluster. They may require over 200 hours of compute time, even within a high-performance environment. These were run within the Golub Cluster from the Illinois Campus Cluster Program: https://campuscluster.illinois.edu/about/system-info/hardware/#golub . Bash scripts for setting up the environment and parallelization within that cluster can be provided upon request.

Instructions to Replicators

The replicator must maintain the folder structure provided in this replication package.

All replication scripts are provided in the folder "/Code"

Main data files are stored in "/Data"

Auxiliary data files will be stored in "/Results/Model_Outputs"

All result figures will be stored in "/Results/Figures"

All result tables will be stored in "/Results/Tables"

To replicate all the tables and figures from the paper, run the "/Code/main.do" script. That script assumes that all software and packages described in the section above are installed locally.

Scripts should be run in the order presented in "/Code/main.do"

The replicator will need to change lines 25-27 of "/Code/main.do" to set the project's main directory, the path where R is installed, and the path where R packages are installed.

The replicator may choose to skip some portions of the code, such as Machine Learning, for example. However, it is essential to build the data file "/Results/Model_Outputs/ihwap_wedge.dta"

The replicator must run the preamble and then the data preparation scripts contained in lines 165-179 from "/Code/main.do" to build "ihwap_wedge.dta"

After building "ihwap_wedge.dta" the other portions of the code can be run independently. For example, if the replicator is interested in replicating results only from Section 4.1 of the paper, then they can run lines 248-266 from "/Code/main.do"

List of Tables, Figures and Scripts

The provided scripts reproduce all tables and figures in the paper. All tables and figures from the Appendix are also generated, with the exception of the map from Figure A.1, which has been omitted to preserve the anonymity of subjects. The table below provides details about which scripts produce each table or figure from the main body of the paper. More details about Appendix figures and tables are provided in "Code/main.do"

It is recommended that all scripts should be run from within "/Code/main.do"

Figure/Table #	Scripts	Output files
Table 1	/Code/decompose_wedge.do	simul_gap_results.tex; simul_gap_pctresults.tex
Table 2	/Code/contractor_quality.do	contqual_1ststage_save.tex; oster_combined_results.tex
Table 3	/Code/contractor_quality.do	simulations_contractors.tex; simulations_contractors_dropinter.tex
Table 4	/Code/behavioral_responses.do	new_reboundsave_results.tex; new_reboundsave_pctresults.tex; new_rebound_results.tex; new_rebound_pctresults.tex
Figure 1	/Code/descriptives.do	ratio_totalbins.pdf
Figure 2	/Code/average_effects.do	ML_trueVSpredict.pdf
Figure 3	/Code/wedgebymeasures_graphs.do	wedge_furnace_overlay.pdf; wedge_windows_overlay.pdf; wedge_airseal_overlay.pdf; wedge_attic_overlay.pdf; wedge_wallins_overlay.pdf
Figure 4	/Code/behavioral_responses.do	tavgbin_Gas_raw.pdf; hddbest_prepost_controls.pdf
Figure 5	/Code/cba_results.do	npb_homerank_12month_scc.pdf; npb_homerank_12month_retail.pdf

Data Dictionary

The following table describes the variables used in the study, contained in "/Results/Model_Outputs/ihwap_wedge.dta"

Variable	Level	Description	Unit
Household	household	Anonymized unique household ID	
treated	household-by-month	Treatment indicator	
max	household	Maximum R-squared from home-specific PRISM analyses	
HDDbest	household	Optimal heating degree days from home-specific PRISM analyses	
ProgramYear; py*	household	Program year in which household was treated	
ARRA	household	Indicator for program year during the American Recovery and Reinvestment Act	
ApplicationDate	household	Date in which household applied for the program	
AuditDate	household	Date from home energy audit	
FinalDate	household	Date from program final inspection	
CloseDate	household	Date in which all program operations are finalized within the system	
JobDays	household	Number of days elapsed between Audit and Final dates	
ExistingConsuption	household	WeatherWorks projection of pre-treatment energy consumption	Btu
ProjectedConsumption	household	WeatherWorks projection of post-treatment energy consumption	Btu
OverallSIR	household	Overall home savings-to-investment ratio according to WeatherWorks	
AirCoditioning	household	Indicator for presence of AC in home	
sqfeet	household	Floor area of home	Sqft
stories; nstories	household	Number of stories in home	
bedrooms; nbedrooms	household	Number of bedrooms in home	
occupants; noccupants	household	Number of home occupants	
buildingdate; builddate; tab_builddate*	household	Year range of construction of home	
liheapemergency	household	Indicator if household benefits from LIHEAP	
priority	household	Priority ranking of household	

		regarding receiving weatherization	
CountyID	household	Anonymized ID for county where home is located	
InspectorID	household	Anonymized ID for home inspector	
ShieldingClass	household	Indicators for shielding provided by structures surrounding home	
BuildingType	household	Indicator for type of building	
Blower_Pre	household	Result from pre-treatment blower door test	CFM50
Blower_Post	household	Result from post-treatment blower door test	CFM50
Blower_Int	household	Result from intermediate blower door test, when performed	CFM50
NoBlower	household	Indicator for presence of blower door tests	
VermiculitePresent	household	Indicator for presence of vermiculite	
WindowCount	household	Home's number of windows	
WHoperational	household	Indicator if water heater was operational	
tbesetting	household	Indicator for typical setting of thermostat	
HeatTypeMain	household	Type of main heating equipment	
HeatTypeSecondary	household	Type of secondary heating equipment	
HeatTypeTertiary	household	Type of tertiary heating equipment	
MainHeatLoad	household	Load factor of main heating equipment	%
SecondaryHeatLoad	household	Load factor of secondary heating equipment	%
TertiaryHeatLoad	household	Load factor of tertiary heating equipment	%
MainHeatBTU	household	Size of main heating equipment	Btu
SecondaryHeatBTU	household	Size of secondary heating equipment	Btu
TertiaryHeatBTU	household	Size of tertiary heating equipment	Btu
MainHeatID	household	ID for main heating equipment	
SecondaryHeatID	household	ID for secondary heating equipment	
TertiaryHeatID	household	ID for tertiary heating equipment	
MainHeatFuelCost	household	Approximate cost of fuel for main heating equipment	Dollars /Btu
SecondaryHeatFuelCost	household	Approximate cost of fuel for secondary heating equipment	Dollars /Btu
TertiaryHeatFuelCost	household	Approximate cost of fuel for tertiary heating equipment	Dollars /Btu
MainHeatOperational	household	Indicator if main heating equipment was operational pre-treatment	

SecondaryHeatOperational	household	Indicator if secondary heating equipment was operational pre-treatment	
TertiaryHeatOperational	household	Indicator if tertiary heating equipment was operational pre-treatment	
MainHeatAge	household	Age of main heating equipment	
SecondaryHeatAge	household	Age of secondary heating equipment	
TertiaryHeatAge	household	Age of tertiary heating equipment	
MainHeatCondition	household	Indicator for condition of main heating equipment	
SecondaryHeatCondition	household	Indicator for condition of secondary heating equipment	
TertiaryHeatCondition	household	Indicator for condition of tertiary heating equipment	
MainHeatFuel	household	Fuel used by main heating equipment	
SecondaryHeatFuel	household	Fuel used by secondary heating equipment	
TertiaryHeatFuel	household	Fuel used by tertiary heating equipment	
MainHeatEfficPre	household	Pre-treatment efficiency of main heating equipment	
SecondaryHeatEfficPre	household	Pre-treatment efficiency of secondary heating equipment	
TertiaryHeatEfficPre	household	Pre-treatment efficiency of tertiary heating equipment	
MainHeatEfficPost	household	Post-treatment efficiency of main heating equipment	
SecondaryHeatEfficPost	household	Post-treatment efficiency of secondary heating equipment	
TertiaryHeatEfficPost	household	Post-treatment efficiency of tertiary heating equipment	
Attic_RValue	household	Pre-treatment level of attic insulation	R-value
Age	household	Age of household head	
Sex	household	Sex of household head	
Race	household	Race of household head	
Own	household	Indicator if dwelling is owned by residents	
Rent	household	Indicator if dwelling is rented by residents	
Wages	household	Household wages	Dollars
SSA	household	Another household income source	Dollars
Unemployment	household	Another household income source	Dollars
SSI	household	Another household income source	Dollars

AARD	household	Another household income source	Dollars
TANF	household	Another household income source	Dollars
GovTrans	household	Another household income source	Dollars
Other	household	Another household income source	Dollars
Total	household	Total household income	Dollars
Disabled	household	Indicator for presence of a disabled household member	
Minors; hasminor	household	Indicator for minors present in household	
Elderly; haselderly	household	Indicator for presence of an elderly household member	
NativeAm	household	Indicator for presence of a native american household member	
tot_est… AirCon AirSeal Attic Baseload Door Foundation Furnace General HealSfty WallIns Window WtHtr	household	Pre-treatment quote of amount to be spent on the corresponding upgrade categories: Air Conditioning; Air Sealing; Attic; Baseload; Doors; Foundation; Furnace; General; Health and Safety; Wall Insulation; Windows; Water Heater	Dollars
tot_act…	household	Actual amount spent in the corresponding upgrade categories	Dollars
est_lab…	household	Pre-treatment quote for labor costs for the corresponding upgrade categories	Dollars
est_mat…	household	Pre-treatment quote for cost of materials for the corresponding upgrade categories	Dollars
FirmCount	household	Number of firms that worked on the home	
N_...	household	Number of measures performed in a given upgrade category	
SIR_...	household	Average SIR, according to WeatherWorks, for the measures in a given upgrade category	
BlowerPreBins	household	Binned version (categorical) of BlowerPre	
Blower_Pre1000	household	Transformed version of BlowerPre	CFM50 /1000
BlowerReduc	household	Reductions in the blower door test:	CFM50

		-1*(BlowerPost - BlowerPre)	
lnBlowerReduc	household	Natural logarithm of BlowerReduc	
BlowerReduc1000	household	Transformed version of BlowerReduc	CFM50 /1000
blowertarget	household	Target post-treatment blower door reading	CFM50
tab_race1-6	household	Separate indicators for race of household head	
roundage	household	Age of household head, rounded to nearest ten	
Real_income	household	Inflation-adjusted total household income	Dollars in 2017
Real_income1000	household	Transformed version of Real_income, diving by 1000	
roundIncome	household	Household income rounded to the nearest 5000	
nelderly	household	Number of elderly occupants	
AgencyID	household	Anonymized ID for agency that served the home	
inspectmonth	household	Month of final inspection	
auditmonth	household	Month of energy audit	
Blower_performance	household	Index for air sealing performance = Blower_Post/blowertarget	
above_target	household	Indicator if air sealing reductions went beyond target	
CFMPost_minus_Target	household	Blower_Post - blowertarget	CFM50
BonusReduc	household	Air sealing reductions beyond target	CFM50
target_reduc	household	Target air sealing reductions	CFM50
chicago	household	Indicator if home located in Chicago	
total_cost	household	Total costs of all upgrades performed	Dollars
oth_cost	household	Summation of the following cost categories: Air Conditioning; Baseload; Doors; Foundation; General; Health and Safety; Windows; Water Heater	Dollars
nonHS_cost	household	Total costs of upgrades, excluding Health and Safety	Dollars
Real_*	household	Inflation-adjusted costs for the corresponding categories of spending	Dollars in 2017
Real_*_1000	household	Transformed version of Real_*, dividing by 1000	
engSave	household	Percent energy savings according to the WeatherWorks model = (ProjectedConsumption/ExistingCo	%

		nsumption) – 1	
engSave_level	household	Energy savings, in levels, according to WeatherWorks model	MMBtu
roundAtticR	household	Attic R-value rounded to the nearest 5	R-value
binBlowerReduc	household	Binned version of BlowerReduc	
arch_contID	household	Anonymized ID for architectural contractor serving the home	
bin*	household	Binned versions of spending in corresponding upgrade categories	
rowid	household-by-month	Unique ID of row in dataset	
meterend	household-by-month	End date for the corresponding energy bill cycle	
meterstart	household-by-month	Start date for the corresponding energy bill cycle	
weight_boots*	household	Weights for the corresponding bootstrap sample	
pred_boots*	household-by-month	Counterfactual ML energy usage prediction with the corresponding bootstrap sample	MMBtu
resids_boots*	household-by-month	Residuals (true minus predicted energy usage) from ML with the corresponding bootstrap sample	MMBtu
error_boots*	household-by-month	Prediction errors (residuals divided by true energy usage) from ML with the corresponding bootstrap sample	%
total_mmbtu	household-by-month	Total energy usage (combined gas plus electric) in bill cycle	MMBtu
kfold	household	Random grouping of homes into different folds	
kfold2	household-by-month	Random grouping of observations into different folds	
pred_full	household-by-month	Counterfactual ML energy usage prediction with using the full sample and with the best-performing algorithm	MMBtu
cv_preds	household-by-month	“Out-of-sample” ML Energy usage predictions for validation folds, using the best-performing model	MMBtu
cv_resids	household-by-month	Residuals from cv_preds	MMBtu
cv_errors	household-by-month	Prediction errors from cv_preds	%
inresids_full	household-by-month	Residuals from pred_full	MMBtu

inerrors_full	household-by-month	Prediction errs from pred_full	%
tmax	household-by-month	Average maximum outdoor temperature during billing cycle	Celsius
tmin	household-by-month	Average minimum outdoor temperature during billing cycle	Celsius
precip	household-by-month	Precipitation during billing cycle	Inches
ContractorID*	household	Anonymized ID for contractors serving the homes	
meterdays	household-by-month	Number of days in billing cycle	
HDD60	household-by-month	Heating degree days (base 60) in bill cycle	
HDD65	household-by-month	Heating degree days (base 65) in bill cycle	
CDD75	household-by-month	Cooling degree days (base 75) in bill cycle	
cyclediff	household-by-month	Number of days elapsed between end of previous bill cycle and beginning of current	
electric_mmbtu	household-by-month	Electricity usage	MMBtu
gas_mmbtu	household-by-month	Natural gas usage	MMBtu
gas_share	household-by-month	Share of gas usage relative to total energy consumption in bill cycle	%
electric_share	household-by-month	Share of electricity usage relative to total energy consumption in bill cycle	%
elec_prices	monthly	Residential electricity prices in Illinois	Cents/kWh
gas_prices	monthly	Residential gas prices in Illinois	Cents/therm
is_estimated_gas	household-by-month	Indicator if gas meter was read or if usage was estimated	
is_estimated_electric	household-by-month	Indicator if electric meter was read or if usage was estimated	
ceda	household	Indicator if home was served by CEDA (agency)	
start_day, start_month, start_year	household-by-month	Split version of meterstart	
end_day, end_month, end_year	household-by-month	Split version of meterend	
month_year	household-	Month of sample	

	by-month		
audit_day, audit_month, audit_year	household	Split version of AuditDate	
log_pred_boots*	household-by-month	Natural logarithm of pred_boots*	
log_total_mmbtu	household-by-month	Natural logarithm of energy usage	
log_electric_mmbtu	household-by-month	Natural logarithm of electricity usage	
log_gas_mmbtu	household-by-month	Natural logarithm of natural gas usage	
twoyears_prepost	household-by-month	Indicator for observations within a window of two years before and after treatment	
months_since_treat	household-by-month	Indicators for months since/from treatment period	
tau_ml	household-by-month	Monthly energy savings according to best-performing ML model: (total_mmbtu - pred_full)	MMBtu
tau*	household-by-month	Monthly energy savings according to ML model, for corresponding bootstrap sample	MMBtu
pct_tau_ml	household-by-month	Percent monthly energy savings according to best-performing ML model: (total_mmbtu - pred_full)/pred_full	%
pct_tau*	household-by-month	Percent monthly energy savings according to ML model, for corresponding bootstrap sample	%
regsample	household-by-month	Indicator to keep sample stable across regressions	
job_obs	household	Number utility of observations available for a given home	
ww_treat	household	Indicator to restrict sample to two observations per home (one pre and one post-treatment)	
projected_savings	household	Percent energy savings according to WeatherWorks: (ProjectedConsumption - ExistingConsumption)/ExistingConsumption	%
realized_savings	household-by-month	Percent energy savings according to best-performing model: tau_ml/pred_full	%
savings_gap	household-by-month	Energy savings performance wedge: realized_savings - projected_savings	%

realized_savings*	household-by-month	Percent energy savings for corresponding bootstrap sample	%
savings_gap*	household-by-month	Savings performance wedge for corresponding bootstrap sample	%
income_level	household	Indicator bins for household income level	
age_level	household	Indicator bins for age of household head	
hhsize	household	Indicator bins for household size (number of occupants)	
AirSeal_cat	household	Indicator bins for amount spent on Air Sealing	
Attic_cat	household	Indicator bins for amount spent on Attic	
Furnace_cat	household	Indicator bins for amount spent on Furnace	
WallIns_cat	household	Indicator bins for amount spent on Wall Insulation	
spending_cat	household	Indicator bins defining interactions between AirSeal_cat, Attic_cat, Furnace_cat, and WallIns_cat	
gas_utility_ID	household	Anonymized ID for gas utility serving home	
electric_utility_ID	household	Anonymized ID for electric utility serving home	