Behavioral Sciences Team
for a better choice

Comparing Matching Methods in Behavioral Programs

An Evaluation of Smartphone Energy Management Service App

Toshihiro Mukai, Ken-ichiro Nishio, Hidenori Komatsu, Toshiya Iwamatsu, Kim Hyunbae, Kazuyoshi Nakano, CRIEPI

Masanobu Sasaki, Takashi Ogawa, TEPCO Energy Partner

Satoko Otani, Chika Ito, Toppan Printing

Yoko Odate, Crossdoor

Wataru Maeki, Deloitte Tohmatsu Consulting

2019 IEPEC - Denver, CO

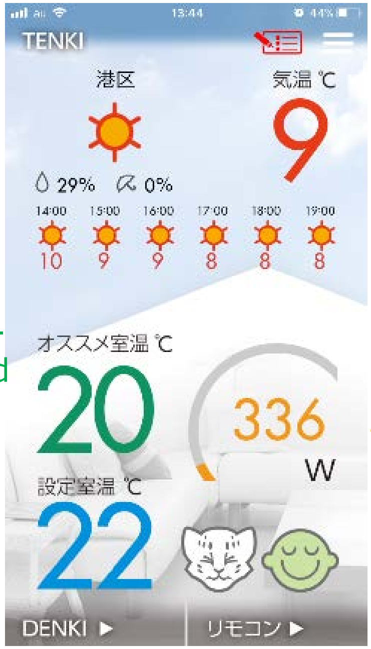
August 20, 2019

CRIEPI
Central Research Institute of Electric Power Industry

IEPEC

We launched behavior-based energy conservation programs in 2017

- Japan’s Ministry of the Environment has been conducting demonstration projects to facilitate low-carbon behavior change by using behavioral insights since 2017.



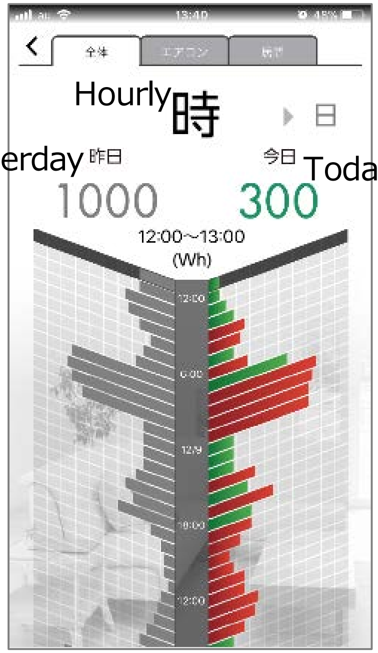
Outdoor temp.

Recom-mended temp. setting

Realtime feedback

Indoor temp. setting

Home screen



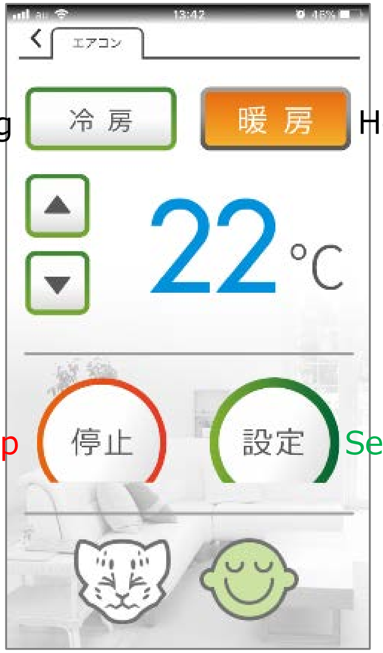
Hourly

Yesterday

Today

Feedback

✓ Whole house & AC
✓ Minutely, hourly,,, monthly



Cooling

Heating

Stop

Settings

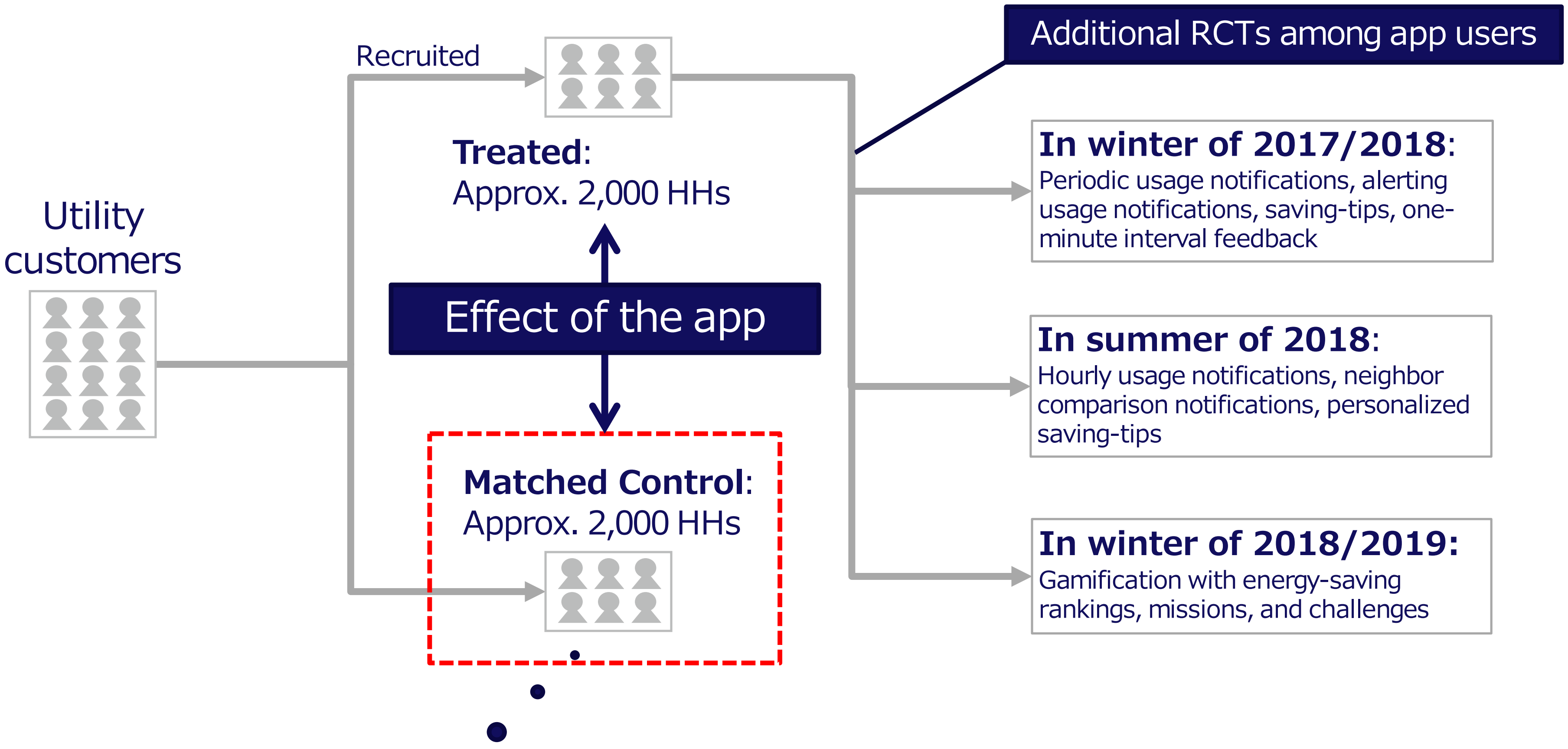
AC remote control

(Komatsu, et al. 2019)

Effect of the app in the winter of 2017 was *2.5%* (p<0.001)

Note: It was the effect in the first 82 days, from Dec. 11, 2017 to Mar. 2, 2018. The updated results of this project will be presented at BECC conference 2019 by Iwamatsu, et al. (2019)

Experimental design



Utility customers

Recruited

Treated: Approx. 2,000 HHs

Effect of the app

Matched Control: Approx. 2,000 HHs

Additional RCTs among app users

In winter of 2017/2018: Periodic usage notifications, alerting usage notifications, saving-tips, one-minute interval feedback

In summer of 2018: Hourly usage notifications, neighbor comparison notifications, personalized saving-tips

In winter of 2018/2019: Gamification with energy-saving rankings, missions, and challenges

Which matching methods perform better in terms of achieving balance?

What is “matching”?

Procedure	Examples of options
Step.1: Select attributes to be included in measuring distance	✓ Yearly/seasonal/**monthly**/**hourly usage** ✓ Fuel type, **climate area**, **dwelling type**, PV, EV, battery, family size, family income, etc.
Step.2: Select a distance metric	✓ Exact, **Mahalanobis distance**, **Propensity score**, Mahalanobis distance within the propensity score caliper, Decision tree, Genetic algorithm, etc.
Step.3: Select a matching method	✓ **Nearest neighbor matching** (pairwise or ratio matching), Optimal matching ✓ Subclassification, Full matching, Adjustments (weighting, **replacement**, etc.)
Step.4: Implement evaluation & diagnosis	✓ Common support assessment ✓ **Diagnosis of balance** of treated and matched control households in pre-treatment usages ✓ Analysis of the matching outcome, etc.

Note: Highlighted options by red colors are used in energy program evaluation in literature.

Distance metrics

- Mahalanobis distance: $D_{ij} = (X_i - X_j)' \sum (X_i - X_j)^{-1}$

- Linear propensity score: $D_{ij} = |\mathrm{logit}(e_i) - \mathrm{logit}(e_j)|$

  where $e_i = Pr(T_i = 1|X_i)$

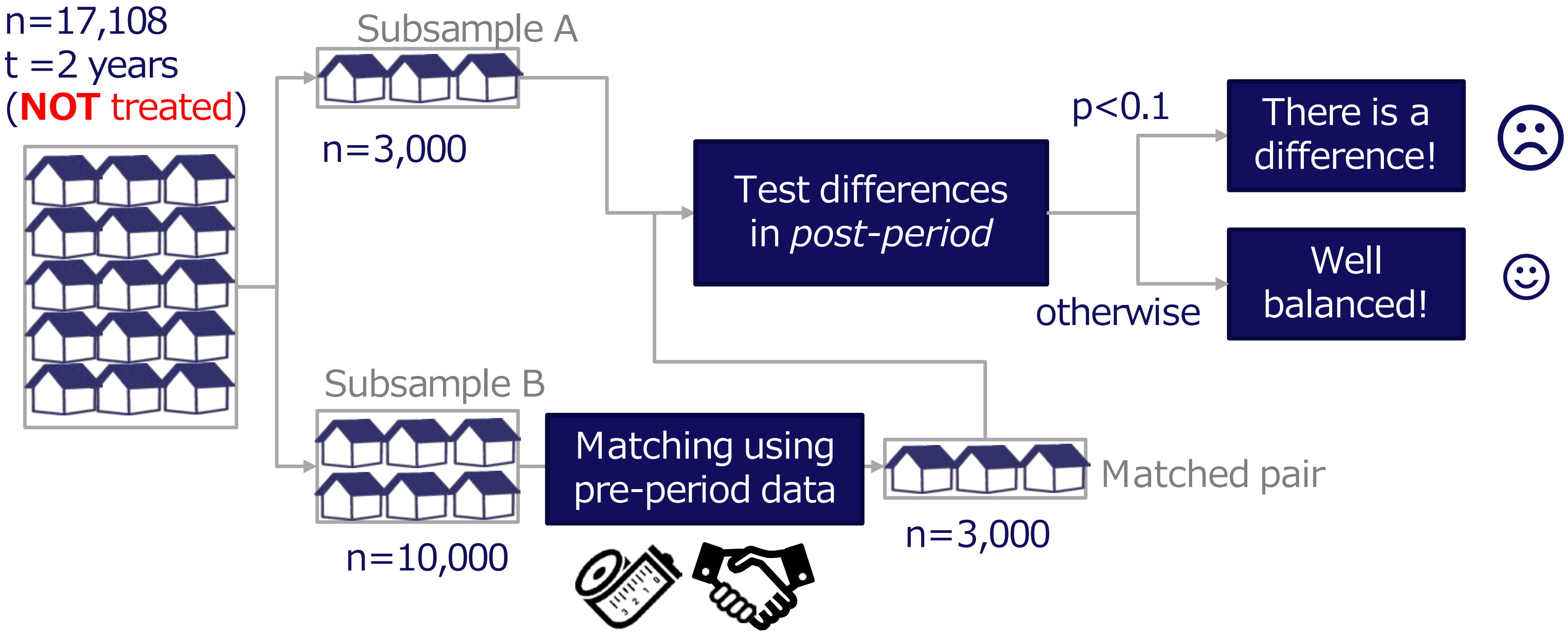
- Mahalanobis distance within propensity score caliper:

  If $|\mathrm{logit}(e_i) - \mathrm{logit}(e_j)| \leq c$ then $D_{ij} = (X_i - X_j)' \sum (X_i - X_j)^{-1}$

  otherwise $D_{ij} = \infty$

Matching method comparison

Performance comparison approach (1)


n=17,108
t =2 years
(**NOT** treated)

Subsample A
n=3,000

Test differences in *post-period*

p<0.1 → There is a difference! ☹

otherwise → Well balanced! ☺

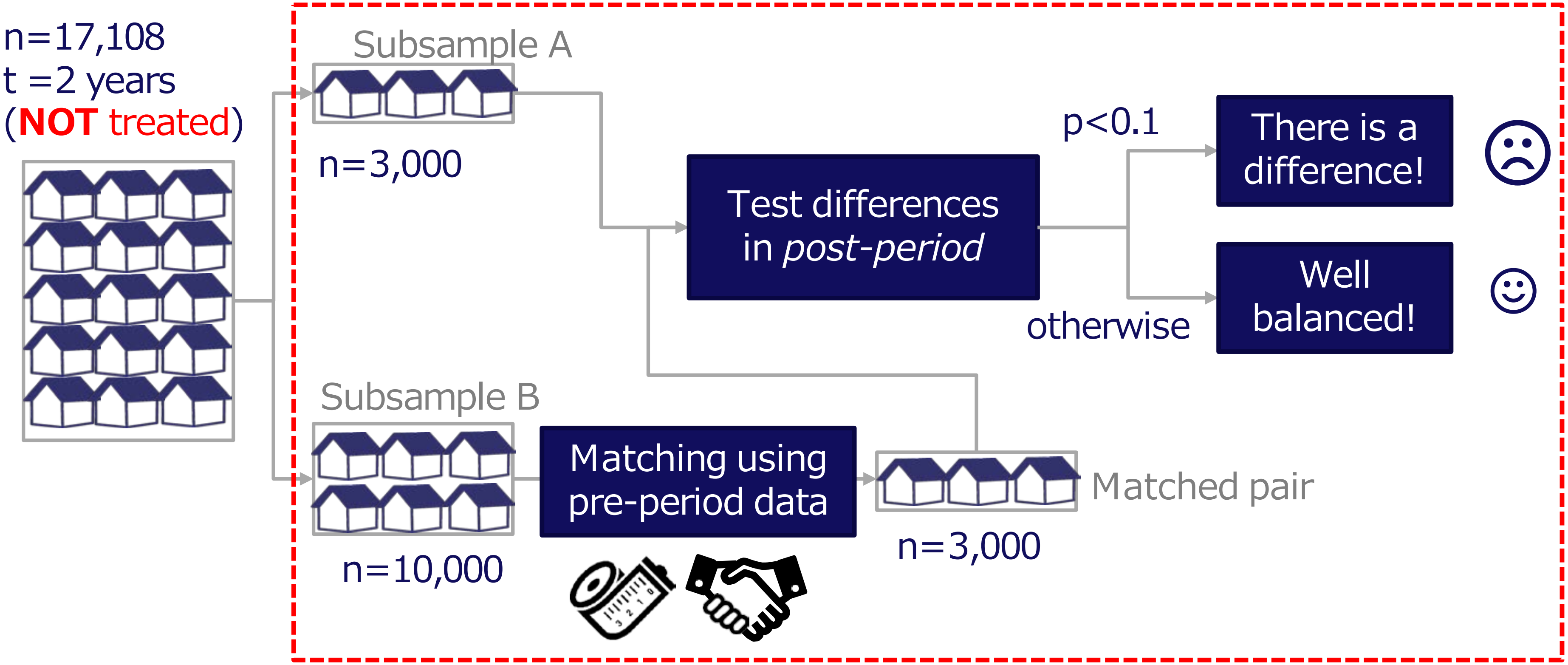
Subsample B
n=10,000

Matching using pre-period data

Matched pair
n=3,000

Bruhn and McKenzie, 2009, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," American Economic Journal: Applied Economics 2009, 1 (4), 200–232.

Performance comparison approach (2)



n=17,108
t =2 years
(**NOT** treated)

Subsample A

n=3,000

Test differences in *post-period*

p<0.1

There is a difference! ☹

otherwise

Well balanced! ☺

Subsample B

n=10,000

Matching using pre-period data

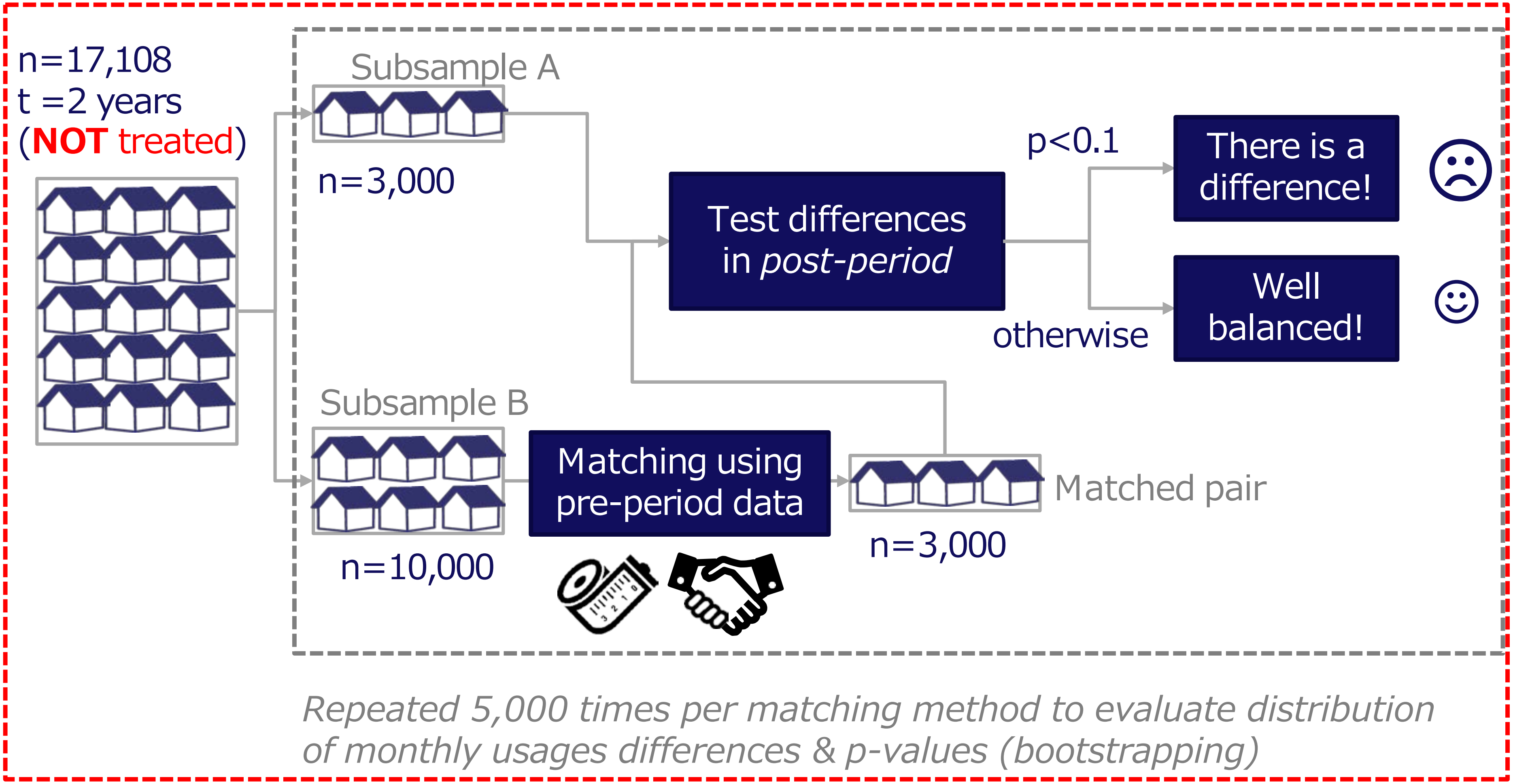
Matched pair

n=3,000

Repeated 5,000 times with replacement to evaluate distribution of monthly usages differences & p-values (bootstrapping)

Bruhn and McKenzie, 2009, “In Pursuit of Balance: Randomization in Practice in Development Field Experiments,” American Economic Journal: Applied Economics 2009, 1 (4), 200–232.

Performance comparison approach (3)



n=17,108
t =2 years
(**NOT** treated)

Subsample A

n=3,000

Test differences in *post-period*

p<0.1

There is a difference! ☹

otherwise

Well balanced! ☺

Subsample B

n=10,000

Matching using pre-period data

Matched pair

n=3,000

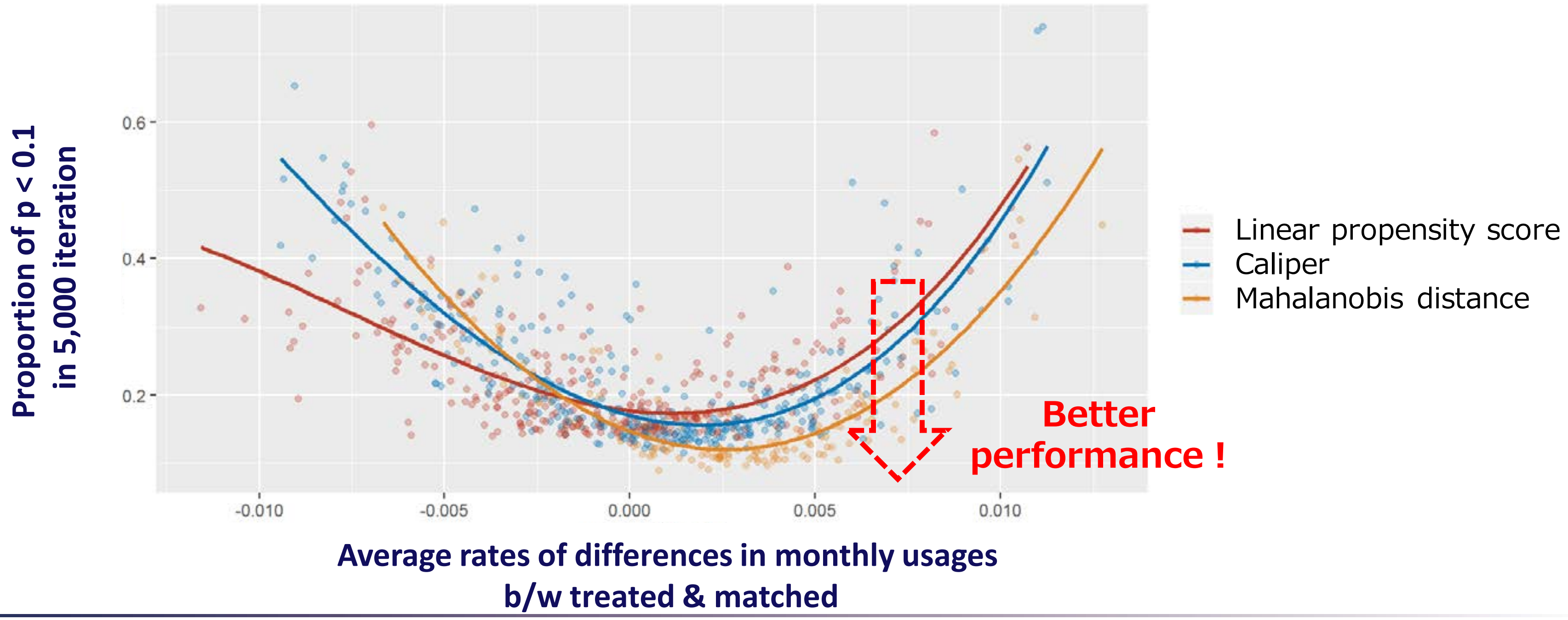
Repeated 5,000 times per matching method to evaluate distribution of monthly usages differences & p-values (bootstrapping)

Repeated 15 times to obtain externally valid results

Bruhn and McKenzie, 2009, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," American Economic Journal: Applied Economics 2009, 1 (4), 200–232.

Which matching methods perform better in terms of achieving balance?

- If
  - There are many treated units (e.g., more than 3,000 treated households)
  - Many more control pool units with a significant overlap in attributes
- Then, better distance metric & attributes to be included are ···
  - **Mahalanobis distance** > Caliper > Propensity score
  - **Three covariates** (Pre-period yearly, summer, and winter average usages) > 12 covariates (pre-period monthly usages)



Better performance !

Key findings

- Recommended metric & attributes:
  - If there are many treated units, **Mahalanobis distance with the three covariates** performed better
    - If there are relatively smaller size of treated units, the selection of the distance metric and the set of covariates should be carefully considered by comparing the results obtained using different matching procedures
- Other recommended specifications:
  - Allow matching **with replacement** as a default option
  - Use **stratification**, if there seem differences in key categorical attributes (e.g., fuel type, region, or dwelling type) between treated and control pool units.
    - Note that, however, that the use of too many attributes for the stratification can cause a deterioration in the balance.
  - If the complete smart meter data is unavailable in the pre-period, consider to **use monthly billing data** as an alternative dataset for calculating the three covariates (pre-period yearly, summer, and winter average usages)

Key findings (literature review)

- Lack of guideline and many matching options complicate recent matching applications procedures & reported information in literature

- We recommend evaluators to report:

  (1) **Methodological**

  - How did they implement matching?
  - Why did the evaluators use the matching procedures?

  (2) **Data availability**

  - Are there sufficient overlaps among the treated, control pool, and matched control units for a credible implementation of matching? (e.g., report the summary statistics and graphical description of the important characteristics of the treated and *control pool units*)

Contact information

Toshihiro Mukai

Research Scientist

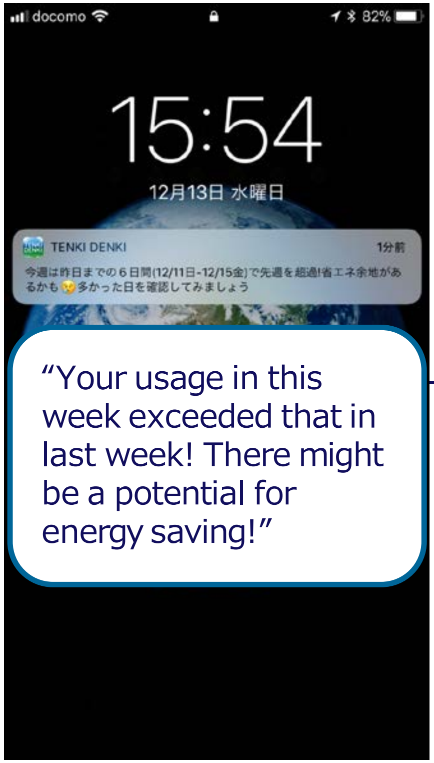
Central Research Institute of Electric Power Industry

mukai@criepi.denken.or.jp

Appendix

A Motion Example of Push Notification

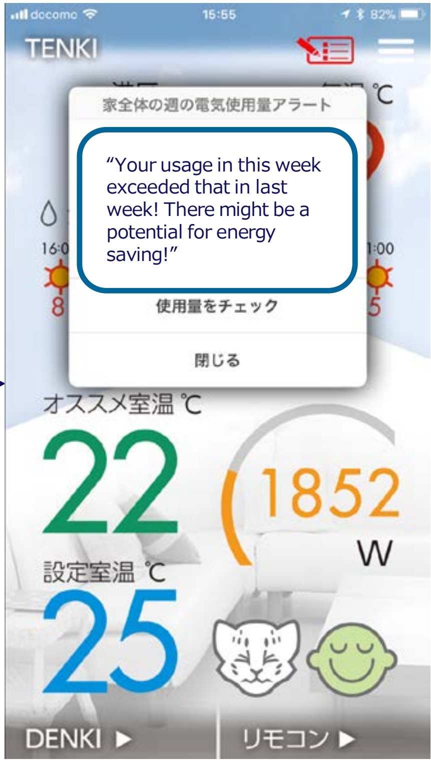
- By sending electricity usage notification in a timely manner, users are stimulated to continuously check the app



今週は昨日までの６日間(12/11日-12/15金)で先週を超過!省エネ余地があるかも😵多かった日を確認してみましょう

“Your usage in this week exceeded that in last week! There might be a potential for energy saving!”

Push notifications displayed on locked screen



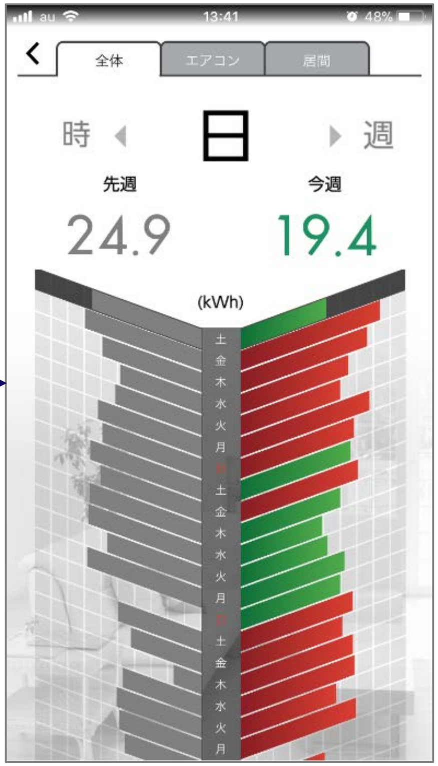
家全体の週の電気使用量アラート

“Your usage in this week exceeded that in last week! There might be a potential for energy saving!”

使用量をチェック

閉じる

Dialog
Displayed in the app



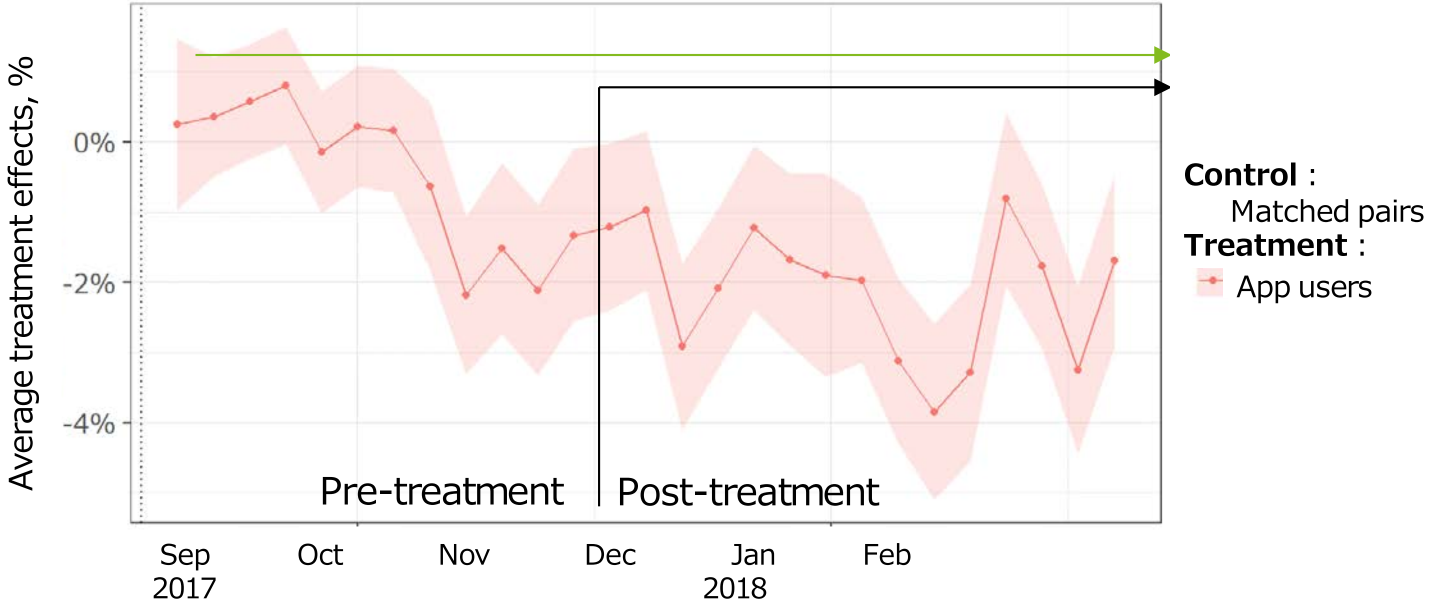
Move to the feedback graph after user taps

Effect of the app

◆ Effect of the app in the winter of 2017/2018 was ***2.5%*** (p<0.001)※1

Monthly average treatment effects

The band reflects a 95% confidence interval



Notes:

- The figure shows the weekly average treatment effects in the first 82 days of the program, from December 11, 2017 to March 2, 2018. The estimation result by panel regression analysis using household-level daily electricity use data. Household-level electricity use from September 2016 to August 2017 were controlled by using post-only model. Matched control households were extracted from the database (HER non-mailed households) by using matching method.

Reference:

※**1** Komatsu, et al. 2019, "Empirical Experiments for a Smartphone App Energy Conservation Service Targeting Residential Sectors: Energy Conservation Effects in Winter 2017," *Energy and Resources* (in Japanese), 40 (3).

Findings from literature review

- Application field is growing in energy programs
  - e.g.) behavior change, dynamic pricing, audit tools
- There isn’t an agreement on ***which metric is appropriate***
  - Braithwait et al. (2017), Olig et al. (2017) uses Mahalanobis distance
  - Smith and Schellenberg (2015); Baylis et al. (2016); DNV-GL (2017) use Propensity score
  - No examination regarding other metrics
- ***Selection of the variable to be included*** in measuring the metric shows both similarity and originality
  - Similarity – many evaluators included monthly usage
  - Originality - Hourly usages on weekdays; Climate zone; Dwelling type Estimated base load, heating and cooling demands

Lacking reliable guideline and too many methodological options, existing application procedures varys