Novel Randomized M&V Examples

2024-05-21

Table of Contents

# 1 Objective

The purpose of this document is to provide open-source coding examples for building analysts who have a strong interest in applying the novel measurement and verification (M&V) procedure described in the paper to their own projects or datasets. By following the guidance below, users should be able to:

1. Replicate the novel M&V method proposed in the manuscript:
   * Design a randomized and balanced switchback control implementation schedule consisting of a baseline strategy and an intervention strategy,
   * Conduct sequential analysis to infer intervention effect on a target metric (e.g. energy consumption, carbon emissions, or thermal comfort).
2. Demonstrate the reliability of the novel randomized method and compare it with the conventional method under:
   * Scenario A: Discrete non-routine events
   * Scenario B: Continuous baseline change
   * Scenario C: Missing measurements

By replicating the method proposed in the manuscript, we aim to demonstrate that the novel M&V approach can reach a conclusion much faster than the standards requirements with sufficient uncertainty quantification. Additionally, by evaluating savings estimations under various interruption scenarios, we showcase the reliability of this new method in overcoming the limitations of conventional approaches (e.g. baseline becomes ‘stale’ and unable to account for unrelated changes). This example code follows the flowchart depicted in the manuscript but only provides a minimum for demonstration purposes. The building analyst should always make reasonable adjustments to specific M&V cases such as using more sophisticated modeling techniques or enforcing more reasonable stopping criteria and thresholds.

require(pacman)  
  
# load packages using pacman  
pacman::p\_load(tidyverse, lubridate, here, rmarkdown,   
 scales, patchwork, magrittr, janitor, qpcR, knitr, # general  
 ggpmisc, # linear regression  
 ggpubr,  
 sjstats, pwr, # anova results  
 nmecr, # m&v modeling package for TOWT  
 slider, # moving averages  
 sprtt, # sequential testing  
 base, blocksdesign, # blocking  
 rstatix, #pipe friendly stats  
 overlapping, # distribution overlapping percentage  
 effsize, bootES, dabestr) #effect size  
  
# turn off scientific notation  
options(scipen = 999, digits = 15)  
  
# set directory  
here::i\_am("Example.rmd")  
  
# set default theme for ggplot  
theme\_set(theme\_minimal())  
  
# define base ggplot theme  
theme\_update(plot.title = element\_text(size = 14, colour = "grey20", face = "bold", hjust = 0.5),  
 plot.subtitle = element\_text(size = 10, colour = "grey20", face = "italic", hjust = 0.5, margin = margin(b = 10)),  
 plot.caption = element\_text(size = 10, colour = "grey20", face = "italic", hjust = 0.5),  
 plot.background = element\_rect(fill = "white", colour = NA),  
 panel.grid.minor = element\_blank(),  
 panel.grid.major = element\_blank(),  
 axis.text = element\_text(size = 10),  
 strip.text = element\_text(size = 10, color = "grey20", face = "bold"),  
 strip.background = element\_blank())  
  
# define global color theme (from palatte "Set2")  
ls\_colours <- c("Baseline" = "#99d8c9",   
 "True baseline" = "#99d8c9",  
 "Biased baseline" = "#1b9e77",  
 "Intervention" = "#fdbb84",  
 "True interv" = "#fdbb84")  
  
# Get case study example data  
readfile\_path <- "./readfiles/"  
  
# Load defined functions  
function\_path <- "./functions/"

# 2 M&V use case

We apply the same use case described in the manuscript where a software-as-a-service control company offers a higher energy-efficient chiller control algorithm. A building analyst was hired to determine the amount of HVAC energy saving through this control intervention. The overall M&V protocol regarding sampling ratio, carryover effect, blocking design and stopping criteria remain the same as the manuscript. For demonstration purposes, the M&V procedure here does not include an annual saving estimation normalized on typical weather using fitted energy models. Instead, we use the average treatment effect (ATE) by calculating the mean difference between baseline measurements and intervention measurements as a proxy for intervention energy-saving effect and normalize on mean of the baseline measurements to obtain fractional savings.

# 3 Use case measurement dataset

For this demonstration, we simply create a hypothetical baseline and intervention measurement set by considering a linear function including outdoor temperature (), peak and off-peak hour indicator (, ) as independent variables:

The term accounts for random noise sampled from a normal distribution . The created measurement set for baseline and intervention controls is shown in figure 3.1. If the building analyst follows the conventional M&V method described in ASHRAE Guideline 14 (G14) or the International Performance Measurement and Verification Protocol (IPMVP), then only baseline measurements are measurable/observable during the pre-retrofit period and only intervention measurements are measurable during the post-retrofit period. However, if the analyst applies the novel M&V we proposed in the manuscript, both data points from both measurement sets can be sampled during both periods.

# load holidays  
list\_holidays <- read\_csv(paste0(readfile\_path, "us\_holidays\_2022.csv"),  
 col\_names = c("date", "weekday", "holiday")) %>%  
 mutate(date = mdy(date))  
  
# load tmy weather file  
df\_tmy <- read\_csv(paste0(readfile\_path, "USA\_IL\_Chicago.Midway.Intl.AP.725340\_TMY3.epw"),  
 skip = 8, col\_types = "ddddd-d---------------------------------",  
 col\_names = c("year", "month", "day", "hour", "min", "tmy")) %>%  
 mutate(year = 2022,  
 time = ymd\_h(paste(paste(year, month, day, sep = "-"), hour, sep = " ")),  
 temp = tmy) %>%  
 dplyr::select(time, temp)  
  
# load temperature from local weather station  
df\_weather <- list.files(path = readfile\_path, pattern = str\_glue("weather\_\*"), full.name = TRUE) %>%  
 map\_dfr(read\_rds) %>%  
 mutate(datetime = with\_tz(datetime\_UTC, tz = "America/Chicago")) %>%   
 select(c(datetime, t\_out)) %>%   
 mutate(across(t\_out, ~ zoo::na.approx(., na.rm = FALSE))) %>%  
 filter(datetime >= as.Date("2021-01-01")) %>%   
 filter(datetime < as.Date("2023-01-01")) %>%   
 unique() %>%   
 arrange(datetime)  
  
# Create a sequence of hourly timestamps for one year  
timestamps <- seq(from = as.POSIXct("2021-01-01 00:00:00"),  
 to = as.POSIXct("2022-12-31 23:00:00"),  
 by = "hour")  
  
# Define peak hours (e.g., 12 PM to 6 PM)  
peak\_hours <- as.integer(format(timestamps, "%H") %in% c("12", "13", "14", "15", "16", "17"))  
  
# Calculate non-peak hours  
non\_peak\_hours <- 1 - peak\_hours  
  
# Baseline model parameters  
intercept <- 45  
beta\_peak <- 20  
beta\_temp <- -0.2  
beta\_non\_peak <- 5  
  
# Compute the energy consumption  
energy\_consumption <- intercept +  
 beta\_temp \* df\_weather$t\_out +  
 beta\_peak \* peak\_hours +  
 beta\_non\_peak \* non\_peak\_hours +  
 rnorm(length(timestamps), mean = 0, sd = sqrt(100))  
  
# Create a data frame  
df\_base <- data.frame(datetime = timestamps,  
 t\_out = df\_weather$t\_out ,  
 power = energy\_consumption)   
  
# Output the first few rows of the dataset  
head(df\_base)

## datetime t\_out power  
## 1 2021-01-01 00:00:00 -1.66666666666667 44.9044973688139  
## 2 2021-01-01 01:00:00 -1.54000000000000 38.9526984693294  
## 3 2021-01-01 02:00:00 -1.41333333333333 47.2275512998179  
## 4 2021-01-01 03:00:00 -1.28666666666667 19.2429347877549  
## 5 2021-01-01 04:00:00 -1.16000000000000 39.0042852954496  
## 6 2021-01-01 05:00:00 -1.03333333333333 43.7710574174219

# Intervention model parameters  
intercept <- 40  
beta\_temp <- -0.7  
beta\_peak <- 12  
beta\_non\_peak <- 10  
  
# Compute the energy consumption  
energy\_consumption <- intercept +  
 beta\_temp \* df\_weather$t\_out +  
 beta\_peak \* peak\_hours +  
 beta\_non\_peak \* non\_peak\_hours +  
 rnorm(length(timestamps), mean = 0, sd = sqrt(100))  
  
# Create a data frame  
df\_interv <- data.frame(datetime = timestamps,  
 t\_out = df\_weather$t\_out,  
 power = energy\_consumption)

df\_base %>%   
 rename(Baseline = power,   
 base\_t\_out = t\_out) %>%   
 left\_join(df\_interv %>% rename(Intervention = power), by = "datetime") %>%   
 pivot\_longer(-c(datetime, base\_t\_out, t\_out), names\_to = "parameter", values\_to = "value") %>%   
 ggplot(aes(x = t\_out, y = value, color = parameter)) +  
 geom\_point(size = 0.1, alpha = 0.1) +  
 geom\_smooth() +  
 scale\_x\_continuous(expand = c(0, 0),   
 limits = c(-15, 35),  
 breaks = breaks\_pretty(n = 5),  
 labels = number\_format(suffix = " °C")) +  
 scale\_y\_continuous(expand = c(0, 0),   
 labels = number\_format(suffix = " kW")) +  
 scale\_color\_manual(values = ls\_colours) +  
 labs(x = NULL,   
 y = NULL,  
 color = NULL,   
 title = "Created baseline and intervention measurements") +  
 theme(panel.grid.major.y = element\_line(colour = "grey80", size = 0.25),  
 legend.direction = "horizontal",  
 legend.position = "bottom",  
 plot.margin = margin(t = 2, r = 7, b = 2, l = 2, unit = "mm"))

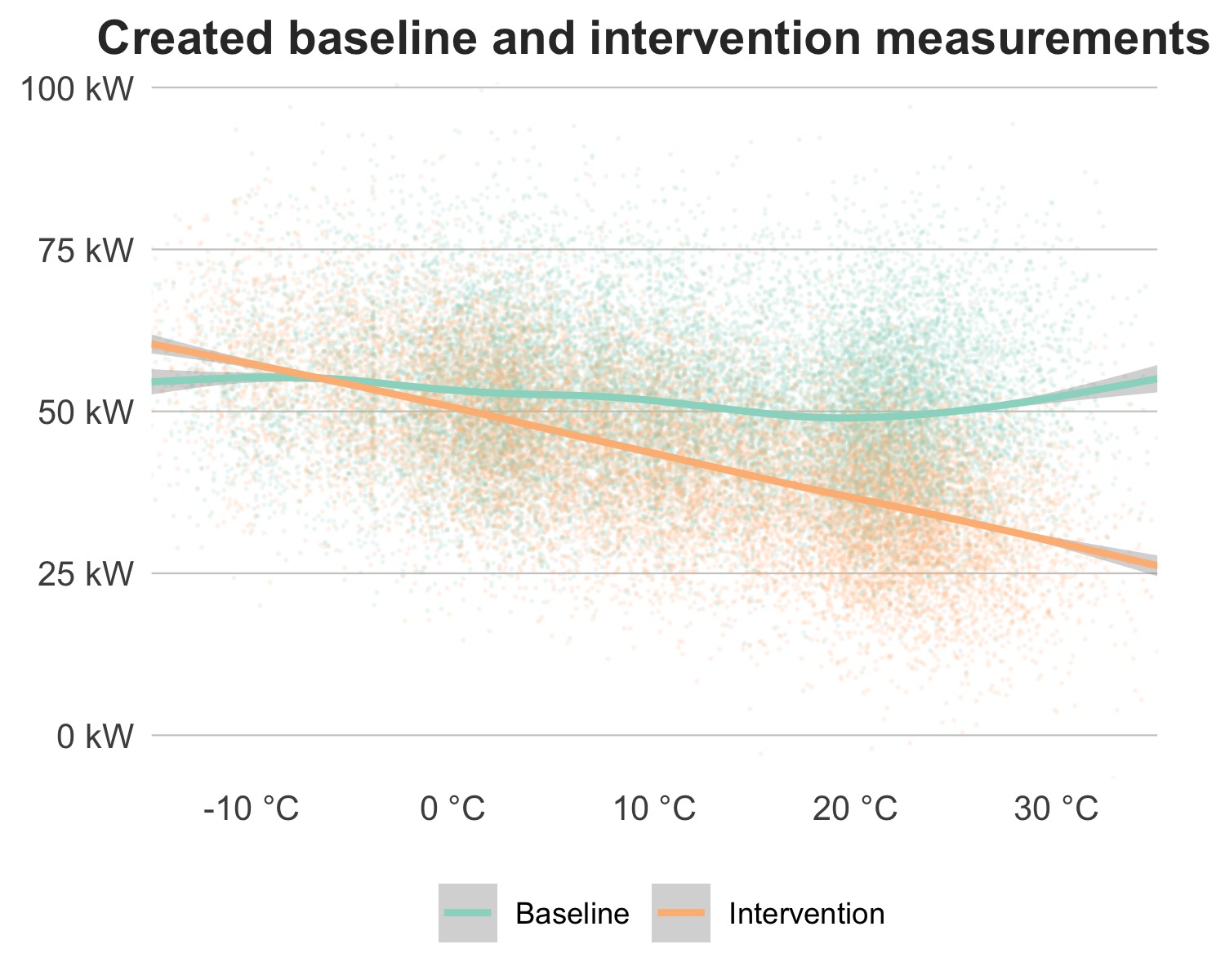


Figure 3.1: Generated hypothetical baseline and intervention measurements using a linear function (data points show hourly energy consumption and line fitted using the loess function).

# 4 Novel randomized M&V analysis

## 4.1 Switchback experimental design

As most control intervention strategies can be implemented interchangeably with the building baseline control, we argue in the paper that randomizing the implementation sequence between the intervention and the baseline can balance the impact of confounding variables when estimating the intervention effect. A simple way to get started is using a built-in function ’*sample()’* in R package *‘base’* to randomly select at each sampling interval which strategy to implement.

# Define strategies and number of days  
strategies <- c("Base", "Interv\_1", "Interv\_2")  
num\_weeks <- 15  
days\_per\_week <- 7  
total\_days <- num\_weeks \* days\_per\_week  
  
# Randomly assign strategies to each day  
assigned\_strategies <- sample(strategies, total\_days, replace = TRUE)  
  
# Convert the linear structure into a weekly structure for easier interpretation  
weekly\_assignment <- matrix(assigned\_strategies, nrow = num\_weeks, ncol = days\_per\_week)  
  
print(weekly\_assignment[1, ])

## [1] "Base" "Interv\_2" "Base" "Interv\_2" "Base" "Base" "Base"

In statistics, average treatment effect (ATE) estimation depends on the statistical power of unbiased randomization and balanced sample size. This means to accurately estimate the energy-saving effect of the intervention control strategy, we should design an unbiased randomization that has an equal amount of measurements for each level of confounding variables. Therefore, a simple complete randomization can still be biased, for instance, getting a poor randomization where a significant portion of the baseline is sampled on weekends. This can be avoided by designing a block randomization, which is a type of experimental design that first divides the testing units into blocks and then randomly assigns which strategy to implement for each unit within each block. Thus in practice, each block type corresponds to one confounding variable, and since strategies are sampled at the block level, this means each confounding variable is covered by all testing strategies. For example, building energy consumption is known to be affected by occupancy, time, and outdoor weather conditions. If the control service company wishes to bill the customer based on the savings solely from the control retrofit itself, the M&V analyst should decompose the measured energy savings and adjust for the influence from mild weather or decreased occupancy. As an example in the paper, we developed a block-in-block experiment shown in figure 4.1. We first blocked by season (i.e. 15 weeks), and within each 15-week block, we further blocked by day of the week. In other words, within each seasonal block, we would have 15 repetitions of Monday to Sunday as sub-blocks. And then we randomly assigned control testing strategies at each sub-block.

knitr::include\_graphics(paste0(readfile\_path, "block\_illustration.png"), dpi = 900)

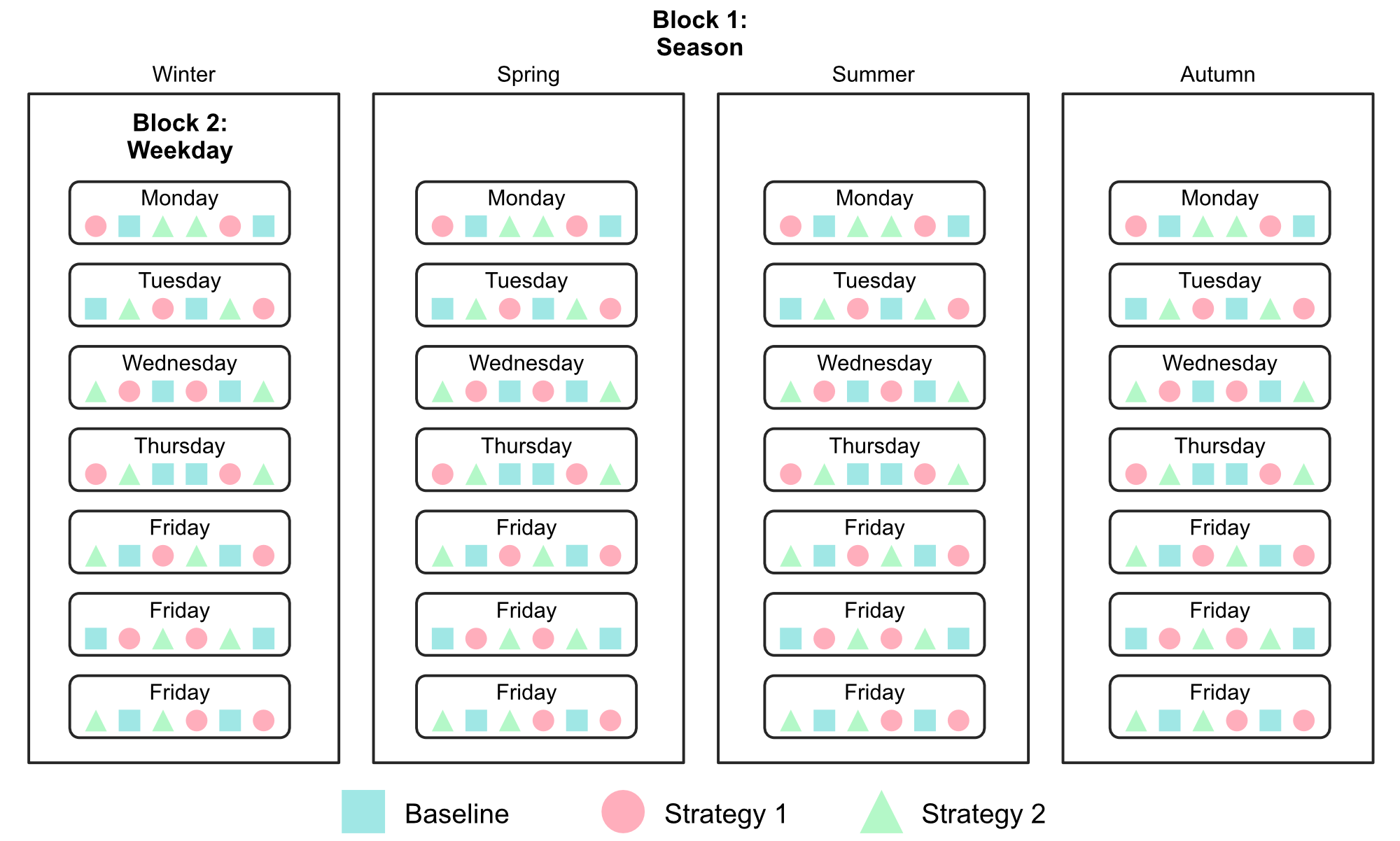


Figure 4.1: Illustration of block-in-block design that aims to balance day-of-week variable and seasonality

To do block design, we referred to another built-in function in R called *‘blocks()’* from *‘blocksdesign’* package. We further bonded other functions and inputs to make the package more versatile and one example is shown below. Once successfully called, the function will return a list consisting of schedule output and schedule summaries. The detailed schedule can be viewed by calling the schedule key name.

source(paste0(function\_path, "blocking.R")) # block design schedule generation function  
schedule <- blocking(start\_date = "2021-01-01",  
 n\_weeks = 60,   
 n\_seasons = 4,   
 seed = 390,   
 searches = 20,   
 jumps = 20,   
 treatments = 2,   
 consec = 1)  
  
kable(head(schedule$schedule, n = 7))

| season | date | weekday | strategy |
| --- | --- | --- | --- |
| 1 | 2021-01-01 | Fri | 1 |
| 1 | 2021-01-02 | Sat | 2 |
| 1 | 2021-01-03 | Sun | 1 |
| 1 | 2021-01-04 | Mon | 1 |
| 1 | 2021-01-05 | Tue | 1 |
| 1 | 2021-01-06 | Wed | 2 |
| 1 | 2021-01-07 | Thu | 2 |

To check whether the blocking experimental design is balanced, we can call three built-in summary table outputs:

1. Weekday summary: calculates how many days are sampled for each season weekday and strategy combination.
2. Consecutive days summary: calculates how many consecutive days are sampled for each strategy

kable(head(schedule$weekday\_summary))

| season | weekday | strategy | n |
| --- | --- | --- | --- |
| 1 | 1 | 1 | 8 |
| 1 | 1 | 2 | 7 |
| 1 | 2 | 1 | 8 |
| 1 | 2 | 2 | 7 |
| 1 | 3 | 1 | 7 |
| 1 | 3 | 2 | 8 |

kable(head(schedule$consec\_summary))

| strategy | consecutive |
| --- | --- |
| 1 | 95 |
| 2 | 95 |

## 4.2 Sequential analysis of the intervention effect

This section replicates the sequential analysis described in the manuscript using the newly created measurement dataset. The results are shown in figure 4.2.

# source(paste0(function\_path, "saving\_pred.R")) # annual saving estimation function  
source(paste0(function\_path, "ol\_est.R")) # independent variable overlapping calculation  
  
# Combine random sampling  
df\_schedule <- data.frame(strategy = schedule$schedule$strategy,  
 datetime = schedule$schedule$date)  
df\_base\_sch <- df\_base %>%   
 left\_join(df\_schedule, by = "datetime") %>%   
 fill(strategy, .direction = "down") %>%   
 filter(strategy == 1)  
  
df\_interv\_sch <- df\_interv %>%   
 left\_join(df\_schedule, by = "datetime") %>%   
 fill(strategy, .direction = "down") %>%   
 filter(strategy == 2)  
  
df\_hourly <- bind\_rows(df\_base\_sch, df\_interv\_sch) %>%   
 arrange(datetime)  
  
# make daily totals  
df\_daily <- df\_hourly %>%  
 group\_by(datetime = floor\_date(datetime, unit = "day")) %>%  
 summarise(strategy = unique(strategy),  
 power\_ave = mean(power, na.rm = TRUE),  
 power\_peak = max(power, na.rm = TRUE),  
 t\_out = mean(t\_out, na.rm = TRUE)) %>%  
 ungroup()

# prepare sequential test dataset  
sprt\_hourly <- df\_hourly %>%  
 mutate(week = interval(min(datetime), datetime) %>% as.numeric('weeks') %>% floor())  
  
# set analysis parameters  
param\_sprt <- list(baseline = "Baseline",  
 strategy = "Intervention",  
 parameter = "power\_ave", # choice: power\_ave; mmoer; co2  
 label = "power", # choice: power; emissions  
 n\_weeks = 60)  
  
# prepare dataframe for analysis  
df\_sprt <- df\_daily %>%  
 mutate(strategy = as.factor(strategy),  
 strategy = recode\_factor(strategy, "1" = "Baseline", "2" = "Intervention")) %>%  
 filter(strategy %in% c(param\_sprt$baseline, param\_sprt$strategy)) %>%  
 pivot\_longer(cols = -c(datetime, strategy),   
 names\_to = "parameter",   
 values\_to = "value") %>%  
 mutate(week = interval(min(datetime), datetime) %>% as.numeric('weeks') %>% floor(),  
 value = value) %>%  
 filter(str\_detect(parameter, param\_sprt$parameter)) %>%  
 droplevels()  
  
  
# define list to store weekly means  
df\_means <- list()  
  
# calculate weekly means  
for (i in 1:param\_sprt$n\_weeks) {  
   
 # subset data by week  
 df\_means[[i]] <- df\_sprt %>%  
 filter(week <= i) %>%  
 group\_by(strategy,   
 parameter) %>%  
 summarise(week = i,  
 value\_ave = mean(value, na.rm = TRUE),  
 value\_sd = sd(value, na.rm = TRUE),   
 .groups = "keep") %>%  
 ungroup()  
  
}  
  
# combine list into df  
df\_means <- bind\_rows(df\_means)  
  
# define lists to store stopping criteria results  
## SPRT results  
sprt\_res <- list()  
  
## 80% independent variable  
overlap\_base <- list()  
overlap\_s2 <- list()  
quantile\_tmy <- quantile(df\_tmy$temp, na.rm = TRUE, c(0, 0.99))  
  
# do sample  
for (i in 2:param\_sprt$n\_weeks) {  
   
 # subset data by week  
 df\_seq <- df\_sprt %>%  
 filter(week <= i) %>%  
 droplevels()  
   
 # Calculate overlapping temperature range  
 overlap\_base[[i]] <- tibble("n\_weeks" = i,   
 overlap\_base = sprt\_hourly %>%  
 filter(week <= i & strategy == 1) %>%   
 ol\_est(., quantile\_tmy))  
   
 overlap\_s2[[i]] <- tibble("n\_weeks" = i,   
 overlap\_s2 = sprt\_hourly %>%  
 filter(week <= i & strategy == 2) %>%   
 ol\_est(., quantile\_tmy))  
  
   
 # do test  
 results\_seq <- seq\_ttest(x = value ~ strategy,   
 data = df\_seq,  
 d = 0.5,   
 power = 0.9,   
 alternative = "less", # greater  
 paired = FALSE,  
 verbose = TRUE)  
   
 # calculate effect size  
 results\_ci <- effsize::cohen.d(d = df\_seq$value,  
 f = df\_seq$strategy,  
 conf.level = 0.90,  
 paired = FALSE,  
 na.rm = TRUE)  
   
 # update user  
 # print(paste0("Calculating effect size for week ", i))  
   
 # bootstrap effect size  
 results\_bs <- bootES::bootES(data = df\_seq, R = 1000,   
 contrast = c(param\_sprt$baseline, param\_sprt$strategy),  
 data.col = "value", group.col = "strategy")  
   
 # extract test statistic  
 sprt\_res[[i]] <- tibble("n\_weeks" = i,   
 "threshold\_lower" = exp(results\_seq@B\_boundary\_log),  
 "threshold\_upper" = exp(results\_seq@A\_boundary\_log),  
 "statistic" = results\_seq@likelihood\_ratio,  
 "decision" = results\_seq@decision,  
 "cohens\_d" = round(results\_ci$estimate, digits = 2),  
 "ci\_low" = round(results\_ci$conf.int[[1]], digits = 2),  
 "ci\_high" = round(results\_ci$conf.int[[2]], digits = 2),  
 "ns\_stat" = round(results\_bs$t0, digits = 2),  
 "ns\_ci\_low" = round(results\_bs$bounds[[1]], digits = 2),  
 "ns\_ci\_high" = round(results\_bs$bounds[[2]], digits = 2))  
   
 }   
   
# work out when threshold is reached  
sprt\_res <- bind\_rows(sprt\_res)  
  
sprt\_res <- sprt\_res %>%  
 mutate(flag = ifelse(str\_detect(decision, "accept"), 1, 0))  
  
sprt\_overlap\_base <- bind\_rows(overlap\_base) %>%   
 mutate(flag = ifelse(overlap\_base >= 0.8, 1, 0),  
 flag = ifelse(flag == lag(flag, 1), 0, flag))  
  
sprt\_overlap\_s2 <- bind\_rows(overlap\_s2) %>%   
 mutate(flag = ifelse(overlap\_s2 >= 0.8, 1, 0),  
 flag = ifelse(flag == lag(flag, 1), 0, flag))

p2 <- ggplot(sprt\_res, aes(x = n\_weeks, y = ns\_stat)) +  
 geom\_ribbon(aes(ymin = ns\_ci\_low, ymax = ns\_ci\_high), alpha = 0.5, fill = "#D0E3F1") +  
 geom\_line(size = 1.25, colour = "grey20", alpha = 0.4) +  
 geom\_text(data = sprt\_res[2, ],   
 aes(x = n\_weeks, y = ns\_ci\_high, label = "Upper 90% CI"),  
 position = position\_nudge(x = 1),   
 colour = "grey20",   
 size = 2.5,   
 fontface = "italic") +  
 geom\_text(data = sprt\_res[2, ],   
 aes(x = n\_weeks, y = ns\_ci\_low, label = "Lower 90% CI"),  
 position = position\_nudge(x = 1),   
 colour = "grey20",   
 size = 2.5,   
 fontface = "italic") +  
 geom\_point(data = sprt\_res %>% filter(n\_weeks%%15 == 0),   
 size = 8,   
 shape = 16,   
 colour = "#347EB3",   
 alpha = 0.5) +  
 geom\_point(data = sprt\_res %>% filter(n\_weeks%%15 == 0),   
 size = 4,   
 shape = 16,   
 colour = "#347EB3",   
 alpha = 0.9) +  
 geom\_text(data = sprt\_res %>% filter(n\_weeks%%15 == 0),   
 aes(x = n\_weeks, y = ns\_stat, label = paste0(round(ns\_stat, 1L), " kW")),   
 position = position\_nudge(x = 2.5),   
 colour = "grey20",   
 size = 3.0,   
 fontface = "italic") +  
 geom\_errorbar(data = sprt\_res %>% filter(n\_weeks%%15 == 0),   
 aes(x = n\_weeks, ymin = ns\_ci\_low, ymax = ns\_ci\_high),   
 width = 1,   
 colour = "#347EB3",   
 alpha = 0.5,   
 size = 0.8) +  
 geom\_text(data = sprt\_res %>% filter(n\_weeks%%15 == 0),   
 aes(x = n\_weeks, y = ns\_ci\_high, label = paste0(round(ns\_ci\_high, 1L), " kW")),   
 position = position\_nudge(x = 2.5),   
 colour = "grey20",   
 size = 2.5,   
 fontface = "italic") +  
 geom\_text(data = sprt\_res %>% filter(n\_weeks%%15 == 0),   
 aes(x = n\_weeks, y = ns\_ci\_low, label = paste0(round(ns\_ci\_low, 1L), " kW")),   
 position = position\_nudge(x = 2.5),   
 colour = "grey20",   
 size = 2.5,   
 fontface = "italic") +  
 geom\_text(data = sprt\_res %>% filter(n\_weeks%%15 == 0 & flag == 1),  
 aes(x = n\_weeks, y = -15, label = "Difference\nfound"),   
 colour = "grey20",   
 size = 3.0,   
 fontface = "italic") +  
 geom\_text(data = slice\_tail(sprt\_res, n = 1),   
 aes(x = n\_weeks, y = ns\_stat, label = paste0(round(ns\_stat, 1L), " kW")),   
 position = position\_nudge(x = 0.5),   
 colour = "grey20",   
 size = 3.0,   
 check\_overlap = TRUE,   
 hjust = 0) +  
 geom\_text(data = slice\_tail(sprt\_res, n = 1),   
 aes(x = ifelse(is.na(first(flag)), 0, n\_weeks),   
 y = ns\_ci\_high,   
 label = ifelse(is.na(first(flag)), NULL, paste0(round(ns\_ci\_high, 1L), " kW"))),   
 position = position\_nudge(x = 0.5),   
 colour = "grey20",   
 size = 2.5,   
 check\_overlap = TRUE,   
 hjust = 0) +  
 geom\_text(data = slice\_tail(sprt\_res, n = 1),   
 aes(x = ifelse(is.na(first(flag)), 0, n\_weeks),   
 y = ns\_ci\_low,   
 label = ifelse(is.na(first(flag)), NULL, paste0(round(ns\_ci\_low, 1L), " kW"))),   
 position = position\_nudge(x = 0.5),   
 colour = "grey20",   
 size = 2.5,   
 check\_overlap = TRUE,   
 hjust = 0) +  
 geom\_vline(xintercept = seq(0, 50, by = 15),   
 linetype = "dashed",   
 color = "grey20",   
 alpha = 0.3,   
 size = 0.5) +  
 annotate(geom = "text",   
 x = seq(5, 50, by = 15),   
 y = 2.5,   
 label = paste0("15-week block"),   
 alpha = 0.5,   
 size = 3) +  
 scale\_x\_continuous(expand = c(0, 0),   
 limits = c(1, param\_sprt$n\_weeks + 0.5),  
 labels = number\_format(accuracy = 1L, suffix = "\nweeks")) +  
 scale\_y\_continuous(expand = c(0, 0),   
 limits = c(-15, 2.5),  
 breaks = pretty\_breaks(n = 3),  
 labels = number\_format(suffix = " kW")) +  
 labs(title = NULL,   
 subtitle = "SPRT results and estimated difference in power consumption without weather normalization",  
 x = NULL,   
 y = NULL) +  
 guides(alpha = "none") +  
 coord\_cartesian(clip = "off") +  
 theme(panel.grid.major.y = element\_line(colour = "grey80", size = 0.25),  
 axis.text.x = element\_blank(),  
 plot.margin = margin(3, 15, 3, 3, unit = "mm"))  
  
p4 <- ggplot(sprt\_overlap\_base) +  
 geom\_line(aes(x = n\_weeks, y = overlap\_base, color = "Baseline"),   
 size = 1.25) +  
 geom\_line(data = sprt\_overlap\_s2,   
 aes(x = n\_weeks, y = overlap\_s2, color = "Intervention"),  
 size = 1.25) +  
 geom\_text(data = sprt\_overlap\_base[nrow(sprt\_overlap\_base), ],  
 aes(x = n\_weeks, y = 0.8, label = "80% of\nTMY range\nthreshold"),  
 position = position\_nudge(x = 0.5),   
 color = "red",   
 size = 3.0,  
 check\_overlap = TRUE,  
 hjust = 0) +  
 geom\_hline(yintercept = 0.8, linetype = "dashed", color = "red") +  
 scale\_color\_manual(values = ls\_colours) +  
 scale\_x\_continuous(expand = c(0, 0),   
 limits = c(1, param\_sprt$n\_weeks + 0.5),  
 labels = number\_format(accuracy = 1L, suffix = "\nweeks")) +  
 scale\_y\_continuous(breaks = seq(0.2, 1.0, by = 0.2),   
 labels = c("20%", "40%", "60%", "80%", "100%")) +  
 labs(title = NULL,   
 subtitle = "Confounding variable (outdoor drybulb temperature) range coverage",  
 x = NULL,   
 y = NULL) +  
 coord\_cartesian(clip = "off") +  
 theme(panel.grid.major.y = element\_line(),  
 legend.position = "none",  
 plot.margin = margin(3, 15, 3, 3, unit = "mm"))  
  
# do running mean plot  
p1 <- ggplot(df\_means, aes(x = week, y = value\_ave)) +  
 geom\_line(aes(colour = strategy), size = 1.25) +  
 geom\_text(data = filter(df\_means,   
 week == param\_sprt$n\_weeks,   
 parameter == param\_sprt$parameter),   
 aes(x = param\_sprt$n\_weeks,   
 y = value\_ave,   
 colour = strategy,   
 label = strategy),   
 position = position\_nudge(x = 0.5),   
 size = 3.0,   
 check\_overlap = TRUE,   
 hjust = 0) +  
 scale\_x\_continuous(expand = c(0, 0),   
 limits = c(1, param\_sprt$n\_weeks + 0.5)) +  
 scale\_y\_continuous(expand = c(0, 0),   
 breaks = breaks\_pretty(n = 4),  
 labels = number\_format(suffix = " kW")) +  
 scale\_color\_manual(values = ls\_colours) +  
 labs(title = NULL,   
 subtitle = "Running average power consumption of the case study building",  
 x = NULL,   
 y = NULL) +  
 guides(alpha = "none", colour = "none") +  
 coord\_cartesian(clip = "off") +  
 theme(panel.grid.major.y = element\_line(colour = "grey80", size = 0.25),  
 axis.text.x = element\_blank(),  
 plot.margin = margin(3, 15, 3, 3, unit = "mm"))  
  
  
# Overall result  
p1 / p2 / p4 +  
 plot\_annotation(title = "Overall sequential evaluation results using pre-defined stopping criteria") +  
 plot\_annotation(tag\_levels = c('a'), tag\_suffix = ')') &  
 theme(plot.tag.position = c(0, 1),  
 plot.tag = element\_text(color="black"))



Figure 4.2: Comprehensive results summary of the proposed M&V method applied to the case study throughout a total of 60 weeks

As figure 4.2 indicates, the running average of power consumption shows more savings from the intervention strategy during warmer outdoor weather in general. This also corresponds to the SPRT results shown in the middle of the plot. The difference between baseline measurements and intervention measurements was detected early in week 15. However, since outdoor weather has not reached the 80% threshold range of the typical weather, we need to continue the M&V. Around week 25, the temperature measured in sampled intervention days covered 80% of its typical range and we report a 9 kW savings at the end of this blocking period (i.e. week 30). If this is a real project, and the building analyst follows the steps outlined in the standards, week 30 is around the mid-point of baseline collection assuming no interruptions or delays. If the analyst applies the novel randomized M&V and the building owner or the control service company wishes to keep sampling to week 60, the estimated energy savings can be detected with a further decreased uncertainty range indicated by the bandwidth of the blue ribbon in the SPRT plot.

# 5 Comparison between conventional M&V method

In this section, we demonstrate how random sampling between baseline and intervention strategy is more reliable than the conventional method. In the manuscript, we argue that since the conventional pre-/post-method requires a continuous and independent 12-month baseline and intervention measurement, the analyst can hardly know whether the baseline changes after 12 months. If those changes are not related to the control strategy, the savings estimation can be unreliable even if the modeling accuracy satisfies the minimum requirements from the standards. We simulated a change in building occupancy as a typical non-routine event in scenario A as a case study to illustrate the limitation of the conventional method. We then added other random interruptions/noises to the measurement set in scenarios B and C to further emphasize the better reliability of the novel M&V method. Those interruptions considered are only a subset of real-world issues but are representative.

## 5.1 Scenario A: Discrete non-routine events

The most common and typical non-routine event in an energy-saving M&V project is occupancy fluctuations. This is because occupancy has a substantial influence on building energy consumption but is not related to the purpose of testing the intervention control. In reality, it is less common or convenient to have low-cost and accurate occupancy measurements compared to other independent variables such as outdoor temperature and therefore when a building analyst fits a regression model (e.g. TOWT), he/she normally assumes that occupancy remains unchanged throughout the 12-month baseline measurement and the subsequent full-range intervention measurement. In addition, the current G14 and IPMVP only require the building analyst to check for the coefficient of the variation of the root mean square error (CVRMSE) and normalized mean square error (NMSE) of a fitted baseline model. Thus, if occupancy changes after the retrofit implementation, the building analyst can still obtain an ‘accurate’ baseline model with qualified CV(RMSE) but biased due to such non-routine event being overlooked. The novel M&V method, on the other hand, can easily resolve the issue without adding either project cost or modeling complexity. We expect the impact of non-routine events on energy consumption to be balanced in sampled baseline days and intervention days if we sample randomly with equal probability. This means if occupancy changes over 4 weeks, the randomized M&V can ensure approximately 2 weeks are sampled with baseline and the rest 2 weeks are sampled with intervention. Therefore, the energy savings when estimated as the difference-in-means can still represent the intervention’s true effect. In this paper, our primary intention is not to advocate removing baseline modeling in pre-/post- analysis, but to encourage randomized switching between baseline and intervention.

In this scenario, we assume that the building analyst measured the 12-month baseline under a normal occupancy, and after the retrofit implementation, the energy consumption increased by 10 kW for each metered interval from April to September because of increased occupancy (e.g. temporary stay for visiting purpose) and the analyst was not informed by the change.

# Generate schedule for the post-intervention period  
schedule <- blocking(start\_date = "2022-01-01",  
 n\_weeks = 60,   
 n\_seasons = 4,   
 seed = 390,   
 searches = 20,   
 jumps = 20,   
 treatments = 2,   
 consec = 1)  
  
df\_schedule <- data.frame(strategy = schedule$schedule$strategy,  
 datetime = schedule$schedule$date)  
  
df\_base\_sch <- df\_base %>%   
 left\_join(df\_schedule, by = "datetime") %>%   
 fill(strategy, .direction = "down") %>%  
 filter(datetime >= as.Date("2022-01-01"))   
  
df\_interv\_sch <- df\_interv %>%   
 left\_join(df\_schedule, by = "datetime") %>%   
 fill(strategy, .direction = "down") %>%  
 filter(datetime >= as.Date("2022-01-01"))   
  
# Apply occupancy changes  
df\_base\_alt <- df\_base\_sch %>%  
 mutate(power = ifelse(month(datetime) >= 4 & month(datetime) < 10, power + 10, power))  
  
df\_interv\_alt <- df\_interv\_sch %>%   
 mutate(power = ifelse(month(datetime) >= 4 & month(datetime) < 10, power + 10, power))  
  
df\_hourly\_sch <- bind\_rows(df\_base\_alt %>% filter(strategy == 1), df\_interv\_alt %>% filter(strategy == 2)) %>%   
 arrange(datetime) %>%   
 mutate(power = ifelse(month(datetime) >= 4 & month(datetime) < 10, power + 10, power))

# Visualization  
ggplot() +  
 geom\_point(data = df\_base\_sch, aes(x = datetime, y = power, color = "Biased baseline"), size = 0.2, alpha = 0.1) +  
 geom\_smooth(data = df\_base\_sch, aes(x = datetime, y = power, color = "Biased baseline")) +  
 geom\_point(data = df\_base\_alt, aes(x = datetime, y = power, color = "True baseline"), size = 0.2, alpha = 0.1) +  
 geom\_smooth(data = df\_base\_alt, aes(x = datetime, y = power, color = "True baseline")) +  
 geom\_point(data = df\_interv\_alt, aes(x = datetime, y = power, color = "True interv"), size = 0.2, alpha = 0.1) +  
 geom\_smooth(data = df\_interv\_alt, aes(x = datetime, y = power, color = "True interv")) +  
 scale\_x\_datetime(date\_breaks = "2 months",  
 date\_labels = "%b") +  
 scale\_y\_continuous(expand = c(0, 0),   
 labels = number\_format(suffix = " kW")) +  
 scale\_color\_manual(values = ls\_colours) +  
 labs(x = NULL,   
 y = NULL,  
 color = NULL,   
 title = "Energy measurements in the post-retrofit period") +  
 theme(panel.grid.major.y = element\_line(colour = "grey80", size = 0.25),  
 legend.direction = "horizontal",  
 legend.position = "bottom",  
 plot.margin = margin(t = 2, r = 7, b = 2, l = 2, unit = "mm"))

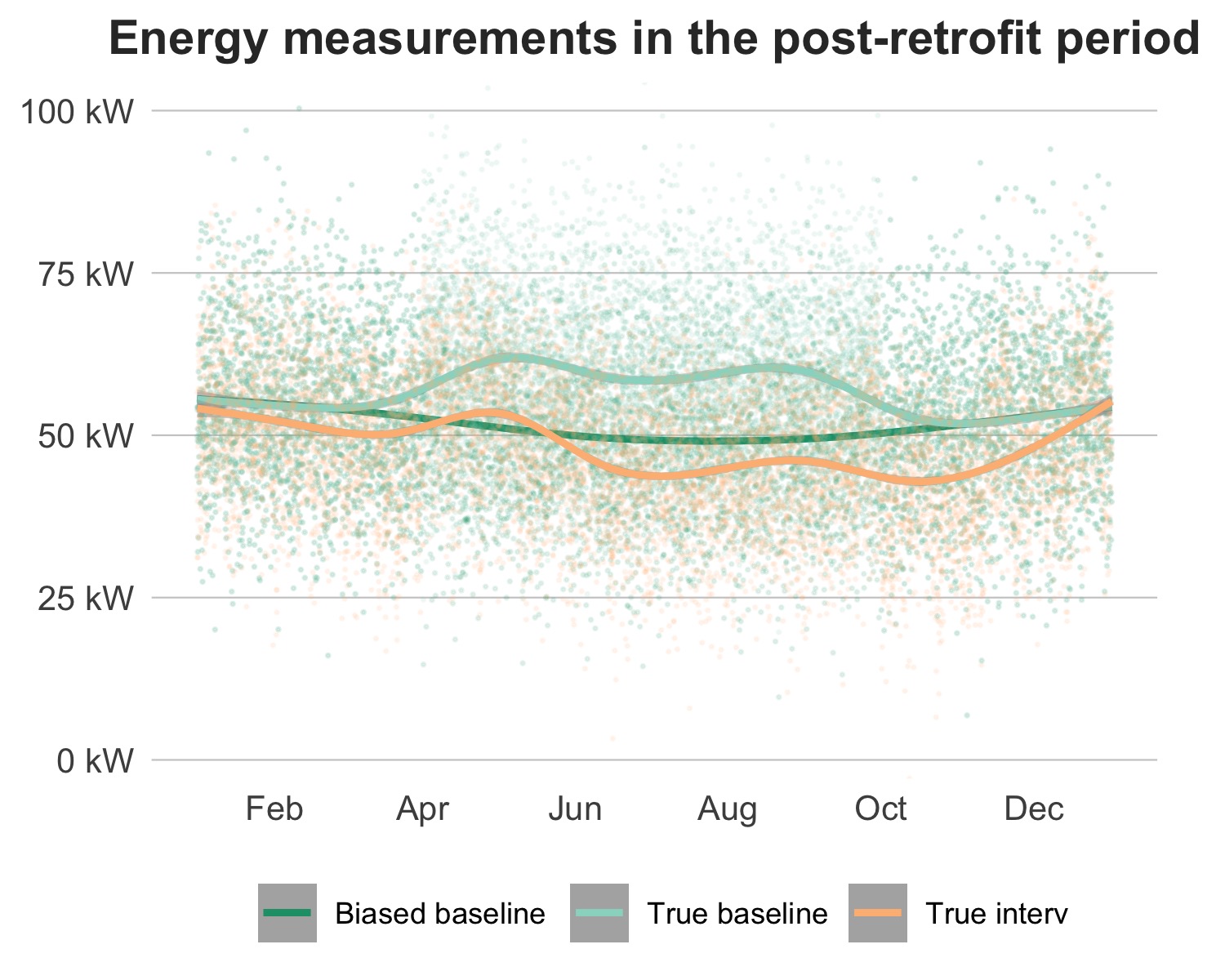


Figure 5.1: Simulated baseline and intervention energy under a discrete non-routine event (Biased baseline indicates analyst’s estimation without prior knowledge about the increase in occupancy. The true baseline and true intervention include the impact of occupancy increase).

If the analyst follows the steps described in G14 and IPMVP, he/she would start baseline measurements 12 months prior to the control retrofit implementation and fit an outdoor weather and time-dependent regression model using collected metered data as the baseline. Then the building manager enables the intervention control in the Building Automation System (BAS) and configures the existing baseline control as a backup when the intervention fails for safekeeping. Figure 5.1 shows what happens after implementing the intervention, there is a 10 kW increase in HVAC power output due to a higher occupancy in the case study building. The biased baseline represents the analyst’s estimation of the projected baseline in the post-retrofit period using modeling results. The figure shows from April to September, it deviates from the true baseline. Although the analyst has accurate intervention measurements, the savings calculated as the difference in mean values between baseline and intervention are underestimated by the conventional method.

As a comparison, if the analyst follows the steps of the novel randomized M&V described in the manuscript, he/she would configure both control strategies in the BAS and switch back and forth between them depending on the pre-determined schedule. Even though the analyst has no prior knowledge about the energy consumption increase due to a higher occupancy over that 6 months period, there is a decent chance that each control strategy has sampled 3 months with a balanced day-of-week category. Therefore, the savings calculated as the difference-in-mean are unbiased by the non-routine events. The results shown in figure 5.2 demonstrate that the conventional pre-/post- method is likely to provide unreliable effect estimation if the analyst can not adjust for non-routine events, but the novel randomized method is robust and requires no additional assumptions or modeling effort.

# saving calculation  
FS\_true <- (sum(df\_base\_alt %>% .$power) - sum(df\_interv\_alt %>% .$power)) / (sum(df\_base\_alt %>% .$power)) \* 100  
  
FS\_new <- (mean(df\_hourly\_sch %>% filter(strategy == 1) %>% .$power) - mean(df\_hourly\_sch %>% filter(strategy == 2) %>% .$power)) / (mean(df\_hourly\_sch %>% filter(strategy == 1) %>% .$power)) \* 100  
  
FS\_conv <- (sum(df\_base\_sch %>% .$power) - sum(df\_interv\_alt %>% .$power)) / (sum(df\_base\_sch %>% .$power)) \* 100  
  
df\_FS <- data.frame(  
 values = c(FS\_true, FS\_conv, FS\_new),  
 category = factor(c("True", "Conventional M&V", "Randomized M&V"), levels = c("True", "Conventional M&V", "Randomized M&V"))  
)  
  
df\_FS %>%   
 ggplot() +  
 geom\_col(aes(x = category, y = values),   
 alpha = 0.6,   
 position = "identity") +  
 geom\_text(aes(x = category, y = values, label = paste0(round(values, digits = 1), "%")),   
 position = position\_nudge(y = -0.5),  
 size = 4) +  
 scale\_y\_continuous(expand = c(0, 0),   
 label = number\_format(suffix = "%")) +  
 labs(title = "Estimated fractional savings\nby different methods",   
 x = NULL,   
 y = NULL) +  
 theme(legend.direction = "horizontal",  
 legend.position = "bottom",  
 plot.margin = margin(t = 2, r = 7, b = 2, l = 2, unit = "mm"))

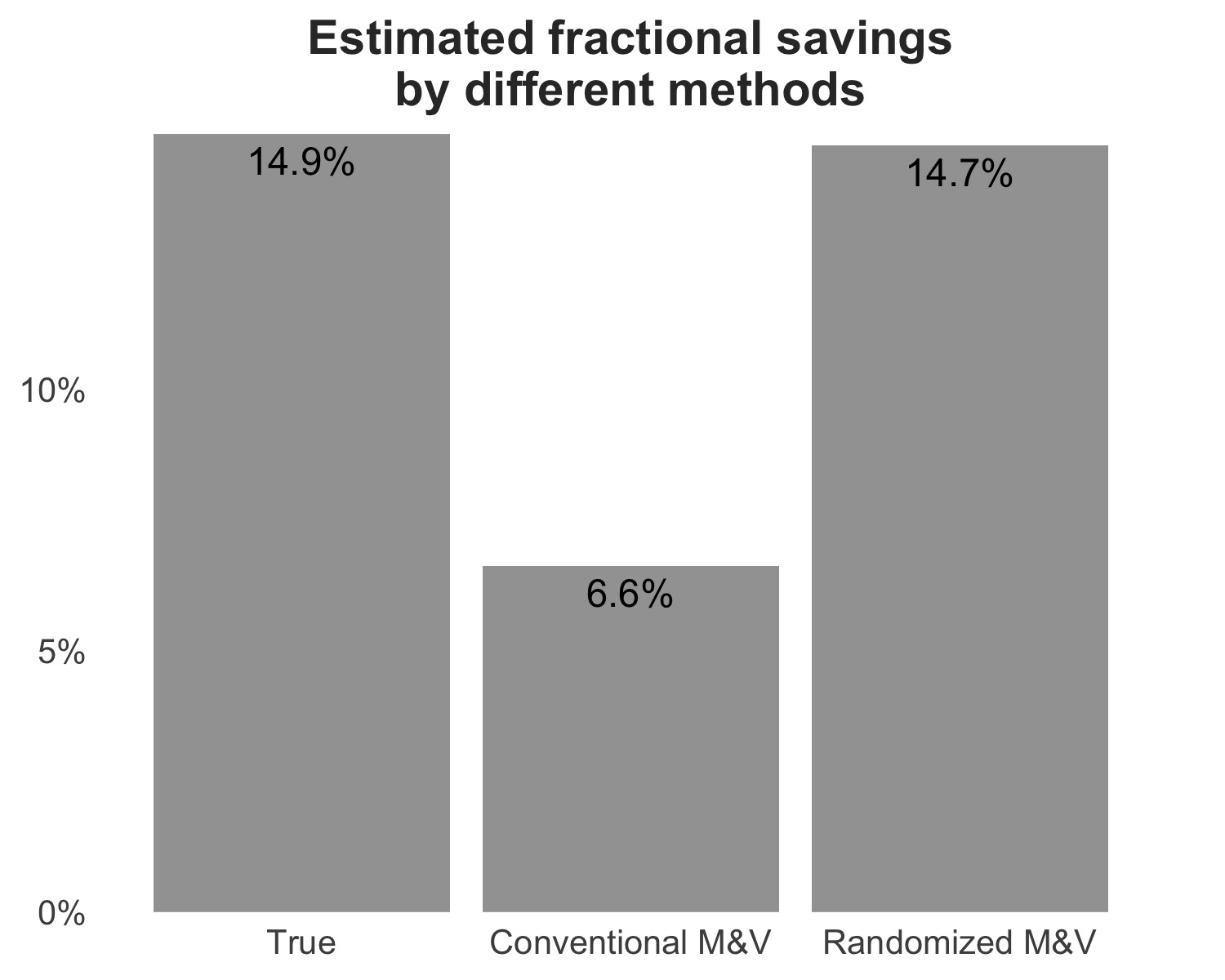


Figure 5.2: Comparison of savings estimation between conventional method and novel method proposed in the paper.

## 5.2 Scenario B: Continuous baseline change due to HVAC components malfunctioning

A baseline can also change continuously over time, which makes it even harder to detect. One example is filter clogging in the AHU with increased fan power due to increased static pressure to overcome the increased resistance of air passing through the filter. In this scenario, if the building manager changes the air filter before starting the M&V and the analyst chooses to use the conventional pre-/post-type analysis, the energy measured 12 months later in the post-retrofit period should correspond to a higher baseline due to increased fan energy as the particle accumulates in the filter. Therefore the conventional M&V underestimates the intervention energy-saving effect. If the analyst chooses to use the novel random sampling M&V, the change in baseline can be documented in sampled baseline measurements.

In this scenario, we simulated a situation where the fan power gradually increased due to accumulated particles in the inlet filter. We assume the fan power accounts for 25% of the HVAC electricity consumption and due to clogging, static pressure increases gradually by 0% to 100%. The relationship between fan power and static pressure can be roughly estimated as , which means the fan power would also increase from 0% to 100%. We demonstrate such increase is omitted by the conventional method in the post-retrofit period where only the control intervention is measured.

# Generate schedule for the post-intervention period  
schedule <- blocking(start\_date = "2022-01-01",  
 n\_weeks = 60,   
 n\_seasons = 4,   
 seed = 390,   
 searches = 20,   
 jumps = 20,   
 treatments = 2,   
 consec = 1)  
  
df\_schedule <- data.frame(strategy = schedule$schedule$strategy,  
 datetime = schedule$schedule$date)  
  
  
df\_base\_sch <- df\_base %>%   
 filter(datetime >= as.Date("2022-01-01"))  
  
df\_interv\_sch <- df\_interv %>%   
 filter(datetime >= as.Date("2022-01-01"))

gradi\_week <- 1 / 60  
df\_base\_alt <- df\_base\_sch %>%  
 mutate(week = interval(min(datetime), datetime) %>% as.numeric('weeks') %>% floor()) %>%  
 mutate(ratio = gradi\_week \* week) %>%  
 mutate(delta\_fan = power \* 0.25 \* ratio) %>%  
 mutate(power = power + delta\_fan)  
  
df\_interv\_alt <- df\_interv\_sch %>%  
 mutate(week = interval(min(datetime), datetime) %>% as.numeric('weeks') %>% floor()) %>%  
 mutate(ratio = gradi\_week \* week) %>%  
 mutate(delta\_fan = power \* 0.25 \* ratio) %>%  
 mutate(power = power + delta\_fan)  
  
ggplot() +  
 geom\_point(data = df\_base\_sch, aes(x = datetime, y = power, color = "Biased baseline"), size = 0.2, alpha = 0.1) +  
 geom\_smooth(data = df\_base\_sch, aes(x = datetime, y = power, color = "Biased baseline")) +  
 geom\_point(data = df\_base\_alt, aes(x = datetime, y = power, color = "True baseline"), size = 0.2, alpha = 0.1) +  
 geom\_smooth(data = df\_base\_alt, aes(x = datetime, y = power, color = "True baseline")) +  
 geom\_point(data = df\_interv\_alt, aes(x = datetime, y = power, color = "True interv"), size = 0.2, alpha = 0.1) +  
 geom\_smooth(data = df\_interv\_alt, aes(x = datetime, y = power, color = "True interv")) +  
 scale\_x\_datetime(date\_breaks = "2 months",  
 date\_labels = "%b") +  
 scale\_y\_continuous(expand = c(0, 0),   
 labels = number\_format(suffix = " kW")) +  
 scale\_color\_manual(values = ls\_colours) +  
 labs(x = NULL,   
 y = NULL,  
 color = NULL,   
 title = "Energy measurements in the post-retrofit period") +  
 theme(legend.direction = "horizontal",  
 legend.position = "bottom",  
 plot.margin = margin(t = 2, r = 7, b = 2, l = 2, unit = "mm"))

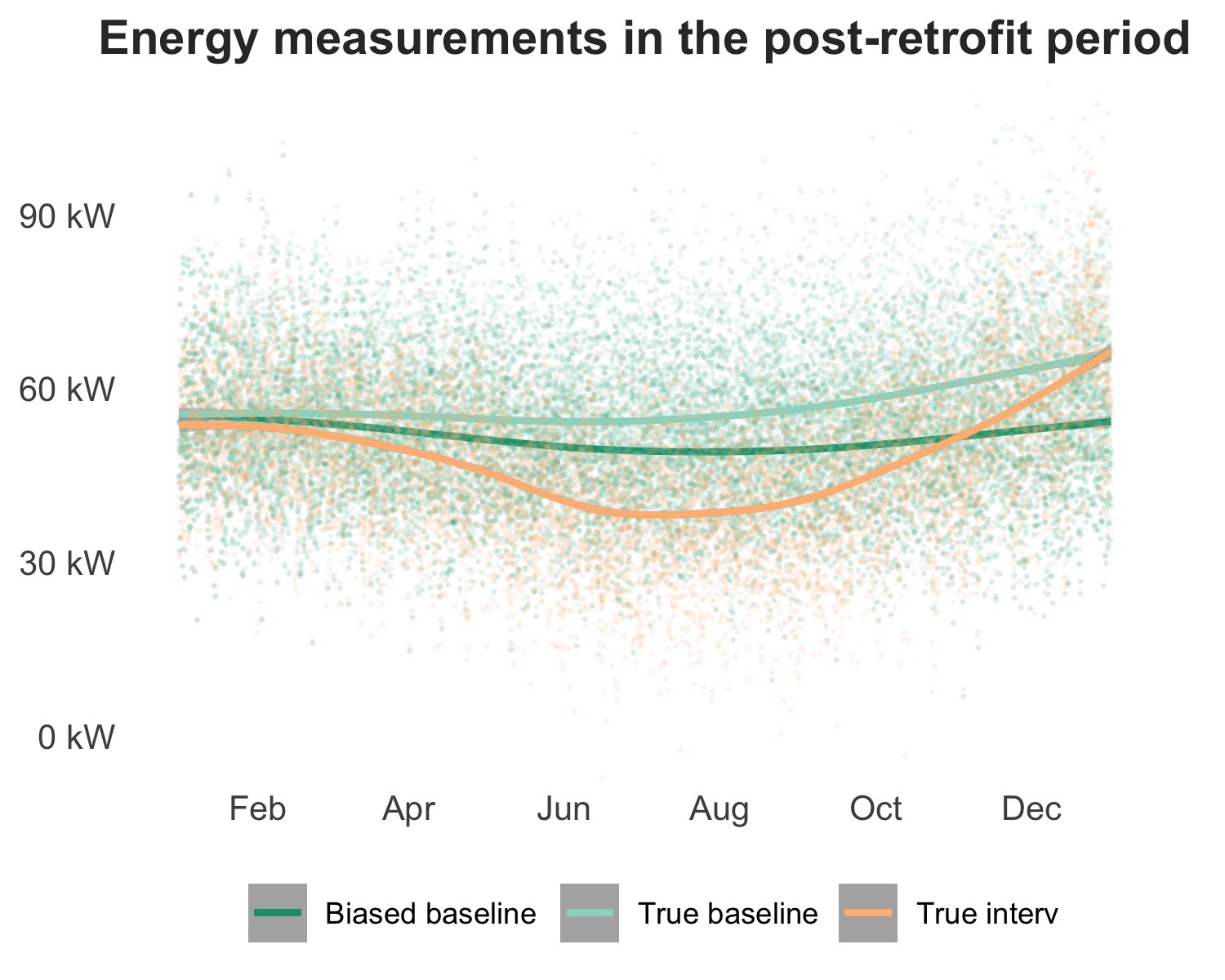


Figure 5.3: Simulated baseline and intervention energy under filter clogging scenario (Biased baseline indicates analyst’s estimation without prior knowledge about the fan power increase since the conventional method has no measure of baseline in the post-retrofit period).

The above figure 5.3 shows all types of measurements in the post-retrofit period in this scenario. Similarly, if the analyst follows the conventional method, he/she can measure the true intervention with increased fan energy. However, the analyst can only obtain the biased baseline without any knowledge or measure of how much fan energy has increased due to clogging filters. And the figure shows a significant amount of savings are underestimated by the conventional method.

# True M&V savings  
FS\_true <- (mean(df\_base\_alt %>% .$power) - mean(df\_interv\_alt %>% .$power)) / (mean(df\_base\_alt %>% .$power)) \* 100  
  
# Conventional M&V savings (biased)  
FS\_conv <- (mean(df\_base\_sch %>% .$power) - mean(df\_interv\_alt %>% .$power)) / (mean(df\_base\_sch %>% .$power)) \* 100  
  
# Noevl M&V savings   
df\_base\_sch <- df\_base %>%   
 left\_join(df\_schedule, by = "datetime") %>%   
 fill(strategy, .direction = "down") %>%   
 filter(datetime >= as.Date("2022-01-01")) %>%   
 filter(strategy == 1)  
  
df\_interv\_sch <- df\_interv %>%   
 left\_join(df\_schedule, by = "datetime") %>%   
 fill(strategy, .direction = "down") %>%   
 filter(datetime >= as.Date("2022-01-01")) %>%   
 filter(strategy == 2)  
  
FS\_new <- (mean(df\_base\_sch %>% .$power) - mean(df\_interv\_sch %>% .$power)) / (mean(df\_base\_sch %>% .$power)) \* 100  
  
df\_FS <- data.frame(  
 values = c(FS\_true, FS\_conv, FS\_new),  
 category = factor(c("True", "Conventional M&V", "Randomized M&V"), levels = c("True", "Conventional M&V", "Randomized M&V"))  
)  
  
df\_FS %>%   
 ggplot() +  
 geom\_col(aes(x = category, y = values),   
 alpha = 0.6,   
 position = "identity") +  
 geom\_text(aes(x = category, y = values, label = paste0(round(values, digits = 1), "%")),   
 position = position\_nudge(y = -0.5),  
 size = 4) +  
 scale\_y\_continuous(expand = c(0, 0),   
 label = number\_format(suffix = "%")) +  
 labs(title = "Estimated fractional savings\nby different methods",   
 x = NULL,   
 y = NULL) +  
 theme(legend.direction = "horizontal",  
 legend.position = "bottom",  
 plot.margin = margin(t = 2, r = 7, b = 2, l = 2, unit = "mm"))

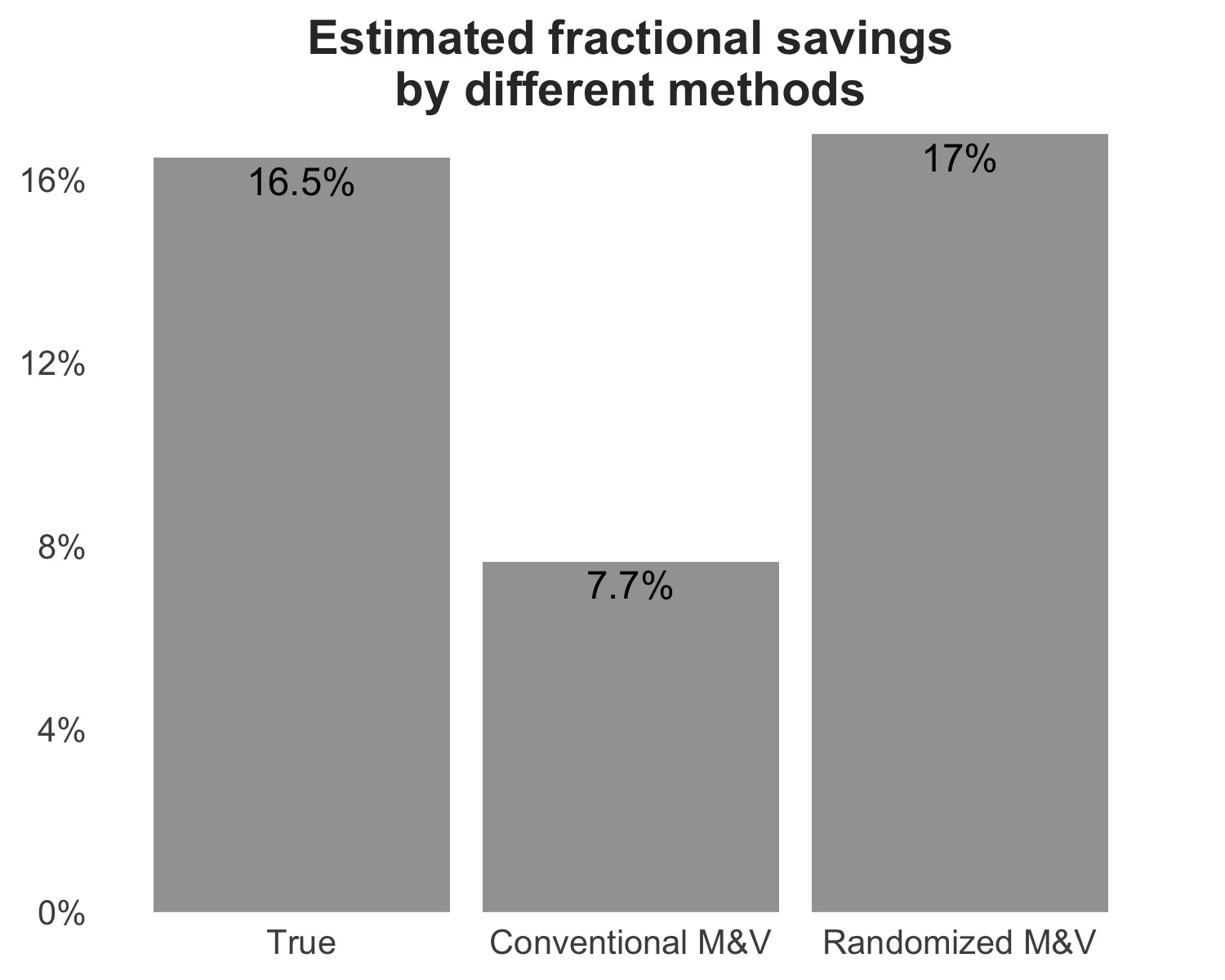


Figure 5.4: Saving estimation comparison between true savings and the two methods (conventional M&V method: no prior knowledge of filter clogging and does not measure baseline, and the novel randomized M&V method: no prior knowledge of filter clogging but measures updated baseline).

The saving estimation shown in figure 5.4 implies that the conventional M&V significantly underestimates the intervention energy-saving effect due to a lack of measurements in the post-retrofit duration. We demonstrate that unless the analyst has prior knowledge about the filter condition and has sub-meters connected to the fan in the AHU, otherwise it is challenging to detect and quantify. The figure shows with only half of the baseline measured (a 50%/50% sampling ratio), the savings can be estimated much more accurately.

## 5.3 Scenario C: Missing data due to network transmission failure

The above scenarios both assume no system failure occurred, but in real projects, a common issue is missing data due to either the network system failing or disabling of the HVAC system for major repair. Under both cases, we expect a large chunk of data missing from the measurement set. If network transmission failure occurs during the cooling season, rendering measurements irrecoverable for two or three weeks, the conventional method would require the baseline to be re-calibrated until the next cooling season in the following year. Furthermore, this delay will also postpone the implementation of the planned intervention leading to a much longer M&V timeline.

In this scenario, we assume that there is a 1-month period of missing measurements in the base year. We demonstrate how the random sampling method can still provide accurate estimation without extension needed. Furthermore, as missing data occurs randomly in the real world, we randomly choose a two-week period and re-shuffle for 100 trials and compare the mean estimated savings with the true savings. Since we created the measurement set at the beginning, we can determine what the true saving is, and by measuring the deviation between the two, we can assess how reliable this novel M&V is.

# Define the number of hours in two weeks  
set.seed(1700)  
  
n\_weeks <- 4  
hours <- n\_weeks \* 7 \* 24  
  
# Define number of tests  
n\_tests <- 100  
FS\_new <- array(data = 0, dim = n\_tests)  
  
df\_hourly\_sch <- df\_hourly %>% filter(datetime < as.Date("2021-01-01") + weeks(60))  
df\_base\_sch <- df\_base %>% filter(datetime < as.Date("2021-01-01") + weeks(60))  
df\_interv\_sch <- df\_interv %>% filter(datetime < as.Date("2021-01-01") + weeks(60))  
  
for (i in 1:n\_tests){  
 random\_start <- sample(1:(8760 - hours), 1)  
  
 # Generate end index  
 random\_end <- random\_start + hours - 1  
   
 df\_hourly\_shuff <- df\_hourly\_sch  
   
 # Randomly erase energy consumption  
 df\_hourly\_shuff$power[random\_start:random\_end] <- NA  
   
 FS\_new[i] <- (mean(df\_hourly\_shuff %>% filter(strategy == 1) %>% .$power, na.rm = T) - mean(df\_hourly\_shuff %>% filter(strategy == 2) %>% .$power, na.rm = T)) / (mean(df\_hourly\_shuff %>% filter(strategy == 1) %>% .$power, na.rm = T)) \* 100  
   
}  
  
# True savings:  
FS\_true <- (sum(df\_base\_sch %>% .$power, na.rm = T) - sum(df\_interv\_sch %>% .$power, na.rm = T)) / (sum(df\_base\_sch %>% .$power, na.rm = T)) \* 100  
  
# Conventional method  
# FS\_conv <- (sum(df\_base\_alt %>% .$power, na.rm = T) - sum(df\_interv\_alt %>% .$power, na.rm = T)) / (sum(df\_base\_alt %>% .$power, na.rm = T)) \* 100  
  
  
FS\_df <- data.frame(FS = FS\_new)

FS\_df %>%   
 ggplot(aes(x = FS)) +  
 geom\_histogram(bins = 10, alpha = 0.5) +  
 scale\_x\_continuous(expand = c(0, 0),   
 labels = number\_format(suffix = "%")) +  
 geom\_vline(aes(xintercept = FS\_true, color = "True savings"), lty = "dashed") +  
 geom\_vline(aes(xintercept = mean(FS\_new), color = "Mean estimated savings"), lty = "dashed") +  
 labs(color = NULL,   
 title = "Repeated analysis of average intervention effect\n(with randomly selected missing days)",  
 x = "Estimated fractional savings",  
 y = "Frequency") +  
 theme(legend.direction = "horizontal",  
 legend.position = "bottom",  
 plot.margin = margin(t = 2, r = 7, b = 2, l = 2, unit = "mm"))

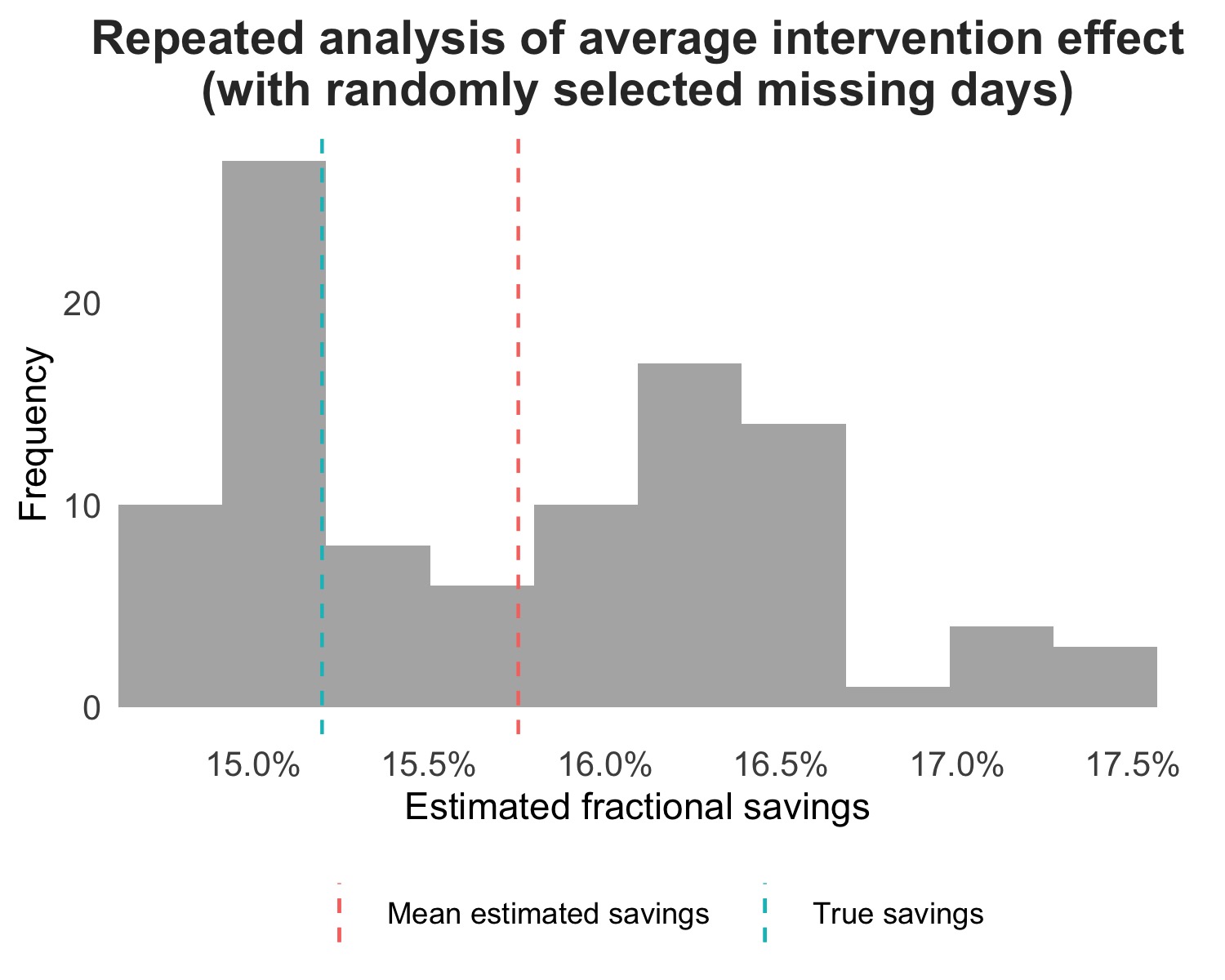


Figure 5.5: Saving estimation comparison between true savings and the mean estimated savings after 100 trials with random 2-month data drop.

The above figure 5.5 shows that the deviation between the mean estimated saving and the true saving is less than 1%. Although the variance of the estimation is slightly larger, and this is because the intervention effect varies with the time of the year and so the saving estimation can be higher/lower compared to the true savings depending on which period of time missing data occurs. But in general, the random sampling method is robust.

# 6 Conclusion

This demonstration code first replicates the novel M&V method described in the manuscript on a hypothetical case study example where we artificially created the baseline and intervention datasets. As a result, the novel randomized M&V method obtained an answer much quicker compared to the timeline outlined by G14 and IPMVP. We then designed three scenarios with different interruptions to assess how reliable the novel method is compared to the conventional pre-\post method in savings estimation. All three scenarios are chosen to be representative of the real-world situations that an M&V analyst would encounter, which include occupancy change as a discrete non-routine event, a gradual change of baseline energy consumption due to filter clogging and missing a period of measurement data due to system failure. We illustrated that it is challenging to detect and address the impact of those interruptions using the conventional method and proved that the deviation between the estimated savings and the true savings is much smaller when using the randomized method.