Group 2: Bokashi/Compost Bioremediation

Marc Los Huertos

2025-04-27

Introduction

This document performs a Before-After-Control-Impact (BACI) analysis using simulated data. The goal is to detect whether an environmental impact causes a change relative to a control site.

BACI analysis, which stands for Before-After, Control-Impact, is a research design used to assess the effects of a treatment or intervention on a population or environment. In R, it can be performed using various statistical models and packages. The core principle is to compare a "control" group (impact group) with an "impact" group (intervention group) both before and after the treatment.

Here's a more detailed explanation of BACI analysis and how it can be implemented in R:

Understanding BACI Design:

1. **Before-After**: This refers to the time periods before and after the intervention or treatment. The "before" period serves as a baseline for comparison, while the "after" period allows researchers to observe any changes that may have occurred due to the intervention.
2. **Control-Impact**: The "control" group is a reference group that is not subjected to the intervention, while the "impact" group is exposed to the treatment or intervention. This allows researchers to assess whether any observed changes in the impact group can be attributed to the intervention rather than other external factors.

Statistical Modeling in R:

–t.test We can test the difference between the means of two groups (e.g., before and after) using a t-test. This is a simple approach but may not be suitable for all data types or distributions – but was will be testing the "delta" or change between treatments. Probably the simpliest to run!

– Simple Linear Models (See Hypothesis w/ Fake Data): For simple BACI analyses, a linear model can be used to assess the differences between groups and time periods. The model can include interaction terms to test for differences in slopes or intercepts between the control and impact groups.

– Generalized Linear Models (GLMs): For many BACI analyses, a GLM is a suitable model to fit the data. The model can incorporate factors like "before/after" and "control/impact" as independent variables, and the dependent variable will be the measured outcome of interest.

– Mixed-Effects Models: If there are hierarchical structures in your data (e.g., multiple sites, replicates, etc.), a mixed-effects model may be more appropriate. These models account for random effects, such as site-specific variability, that can affect the outcome.

Some Complex Approaches

```r
# Load necessary packages
library(lme4)  # For mixed models
```

Create experimental data of random values

```
## Loading required package: Matrix
```

```r
# Example data (replace with your own data)
data <- data.frame(
  site = factor(rep(c("Control", "Impact"), each = 20)),
  time = c(factor(rep(c("Before", "After"), each = 10)), factor(rep(c("Before", "After"), each = 10))),
  outcome = c(rnorm(10, mean = 6, sd = 2), rnorm(10, mean = 4, sd = 2),
              rnorm(10, mean = 5, sd = 2), rnorm(10, mean = 3, sd = 2))  # Example outcome variable
)
```

```r
# Fit a GLM (can be adjusted based on your data)
model <- glm(outcome ~ site * time, data = data)
summary(model)
```

Using GLM

```
##
## Call:
## glm(formula = outcome ~ site * time, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3505  -1.2596   0.0505   0.9889   3.3567
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.4172     0.5596   7.893 2.29e-09 ***
## siteImpact             -0.7732     0.7914  -0.977   0.3351
## timeBefore              1.7320     0.7914   2.188   0.0352 *
## siteImpact:timeBefore  -1.2252     1.1192  -1.095   0.2809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3.131754)
##
##     Null deviance: 148.23  on 39  degrees of freedom
## Residual deviance: 112.74  on 36  degrees of freedom
## AIC: 164.96
##
## Number of Fisher Scoring iterations: 2
```

```r
# Fit a mixed model (if you have nested data)
# Example data with site as a random effect
data$conc <- factor(rep(rep(c("Low", "High"), each = 5),4))
model_mixed <- lmer(outcome ~ site * time + (1 | conc), data = data)
```

Mixed-Effects Model

```
## boundary (singular) fit: see help('isSingular')
```

```r
null_model <- lmer(outcome ~ (1 | conc), data = data)
```

```
## boundary (singular) fit: see help('isSingular')
```

```r
summary(model_mixed)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: outcome ~ site * time + (1 | conc)
##    Data: data
## 
## REML criterion at convergence: 152.5
## 
## Scaled residuals: 
##      Min       1Q   Median       3Q      Max 
## -2.45835 -0.71175  0.02856  0.55879  1.89681 
## 
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  conc     (Intercept) 0.000    0.00    
##  Residual             3.132    1.77    
## Number of obs: 40, groups:  conc, 2
## 
## Fixed effects:
##                        Estimate Std. Error t value
## (Intercept)              4.4172     0.5596   7.893
## siteImpact              -0.7732     0.7914  -0.977
## timeBefore               1.7320     0.7914   2.188
## siteImpact:timeBefore   -1.2252     1.1192  -1.095
## 
## Correlation of Fixed Effects:
##             (Intr) stImpc timBfr
## siteImpact  -0.707              
## timeBefore  -0.707  0.500       
## stImpct:tmB  0.500 -0.707 -0.707
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

```r
anova(model_mixed, null_model)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Warning in optwrap(optimizer, devfun, x@theta, lower = x@lower, calc.derivs =
## TRUE, : convergence code 3 from bobyqa: bobyqa -- a trust region step failed to
## reduce q
```

```
## Data: data
## Models:
## null_model: outcome ~ (1 | conc)
## model_mixed: outcome ~ site * time + (1 | conc)
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)  
## null_model     3 171.91 176.98 -82.955   165.91                       
## model_mixed    6 166.96 177.10 -77.482   154.96 10.946  3    0.01202 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Group Hypotheses

Bokashi will be a more efficient method to reduce pesticide concentrations in decomposed green waste.

-The bokashi samples will see a greater decrease in absorption level compared to the control group and traditional composting group samples after 2 weeks of breakdown

NOTE: No one cares about absorbance, that's a proxy for concentration; please revise to test concentration…

Data Set

```
# Load the data
fakedata.csv = "/home/mwl04747/RTricks/00_Project_Group_Demos/Group2_FakeData.csv"
group1data = read.csv(fakedata.csv, header = TRUE)
str(group1data)
```

```
## 'data.frame':    54 obs. of  5 variables:
##  $ Sample.ID         : chr  "B1ZB" "B2ZB" "B3ZB" "B1FB" ...
##  $ Treatment         : chr  "Bokashi" "Bokashi" "Bokashi" "Bokashi" ...
##  $ Before.After      : chr  "Before" "Before" "Before" "Before" ...
##  $ Concentraion..ppm.: int  0 0 0 5 5 5 10 10 10 0 ...
##  $ Absorption        : num  0.045 0.021 0.034 0.241 0.223 0.234 0.461 0.452 0.443 0.056 ...
```

```
names(group1data)
```

```
## [1] "Sample.ID"          "Treatment"          "Before.After"
## [4] "Concentraion..ppm." "Absorption"
```

```
unique(group1data$Treatment)
```

```
## [1] "Bokashi"         "Compost"         "Water (control)"
```

```
unique(group1data$Before.After)
```

```
## [1] "Before" "After"
```

NOTE: Misspelled concentration… suggest you come up with one word column names and factors () (## Removing Water – since that is a correction, not a factor. I suggest this is Concentration_Initial and then you can have Concentration_Final as a column.

Really, intial will be use to "correct" final values; perhaps we want to test the difference between the two, but I think we want to test the difference between the bokashi and compost treatments.

```
# Remove water
group1data <- group1data %>%
  filter(Treatment != "Water (control)")
# Remove the "Water" treatment

# Check the data
unique(group1data$Treatment)
```

```
## [1] "Bokashi" "Compost"
```

Pretty sure zero concentration (control) is also not a treatment but will be used as a correction factor… please correct fake data to preprocess that. Or we can use R if that would be something you'd like to do… probably take us an hour or so of working together to do that. But it excel it could take 10 min.

```
# Make sure site and time are factors
group1data <- group1data %>%
  mutate(Treatment = factor(Treatment),
         Before.After = factor(Before.After),
         Concentraion..ppm. = factor(Concentraion..ppm.))

# Check the data
str(group1data)
```

```
## 'data.frame':    36 obs. of  5 variables:
##  $ Sample.ID         : chr  "B1ZB" "B2ZB" "B3ZB" "B1FB" ...
##  $ Treatment         : Factor w/ 2 levels "Bokashi","Compost": 1 1 1 1 1 1 1 1 1 2 ...
##  $ Before.After      : Factor w/ 2 levels "After","Before": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Concentraion..ppm.: Factor w/ 3 levels "0","5","10": 1 1 1 2 2 2 3 3 3 1 ...
##  $ Absorption        : num  0.045 0.021 0.034 0.241 0.223 0.234 0.461 0.452 0.443 0.056 ...
```

```
group1data[sample(1:nrow(group1data), 8), ]
```

```
##    Sample.ID Treatment Before.After Concentraion..ppm. Absorption
## 30      C3ZA   Compost        After                  0      0.045
## 6       B3FB   Bokashi       Before                  5      0.234
## 8       B2TB   Bokashi       Before                 10      0.452
## 22      B1FA   Bokashi        After                  5      0.153
## 33      C3FA   Compost        After                  5      0.231
## 7       B1TB   Bokashi       Before                 10      0.461
## 16      C1TB   Compost       Before                 10      0.563
## 17      C2TB   Compost       Before                 10      0.554
```

Summary Stats

Fake data isn't really working – since you don't have any variance with ppm. We are not going reporting absorbance values, but I'll do this now, since it looks like there is some variance there.

```
# Summarize the data
summary_stats <- group1data %>%
  group_by(Treatment, Before.After, Concentraion..ppm.) %>%
  summarise(
    mean = mean(Absorption),
    sd = sd(Absorption),
    n = n()
  ) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Treatment', 'Before.After'. You can
## override using the `.groups` argument.
```

```
summary_stats
```

```
## # A tibble: 12 x 6
##    Treatment Before.After Concentraion..ppm.   mean     sd     n
##    <fct>     <fct>        <fct>               <dbl>  <dbl> <int>
##  1 Bokashi   After        0                   0.04   0.0135     3
##  2 Bokashi   After         5                  0.142  0.0263      3
##  3 Bokashi   After         10                 0.146  0.0108      3
##  4 Bokashi   Before        0                  0.0333 0.0120      3
##  5 Bokashi   Before        5                  0.233  0.00907     3
##  6 Bokashi   Before        10                 0.452  0.00900     3
##  7 Compost   After         0                  0.0693 0.0215      3
##  8 Compost   After         5                  0.249  0.0157      3
##  9 Compost   After         10                 0.329  0.0116      3
## 10 Compost   Before        0                  0.056  0.011       3
## 11 Compost   Before        5                  0.347  0.0286      3
## 12 Compost   Before        10                 0.552  0.0121      3
```

Hypothesis Tests w/Fake Data

```r
# Fit a linear model
model <- lm(Absorption ~ Treatment + Concentraion..ppm. * Before.After, data = group1data)
# Summarize the model
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Absorption
##                                   Df  Sum Sq  Mean Sq F value    Pr(>F)
## Treatment                          1 0.07719 0.077191  67.876 4.396e-09 ***
## Concentraion..ppm.                 2 0.62341 0.311705 274.088 < 2.2e-16 ***
## Before.After                       1 0.12192 0.121917 107.204 2.980e-11 ***
## Concentraion..ppm.:Before.After    2 0.11515 0.057574  50.626 3.470e-10 ***
## Residuals                         29 0.03298 0.001137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

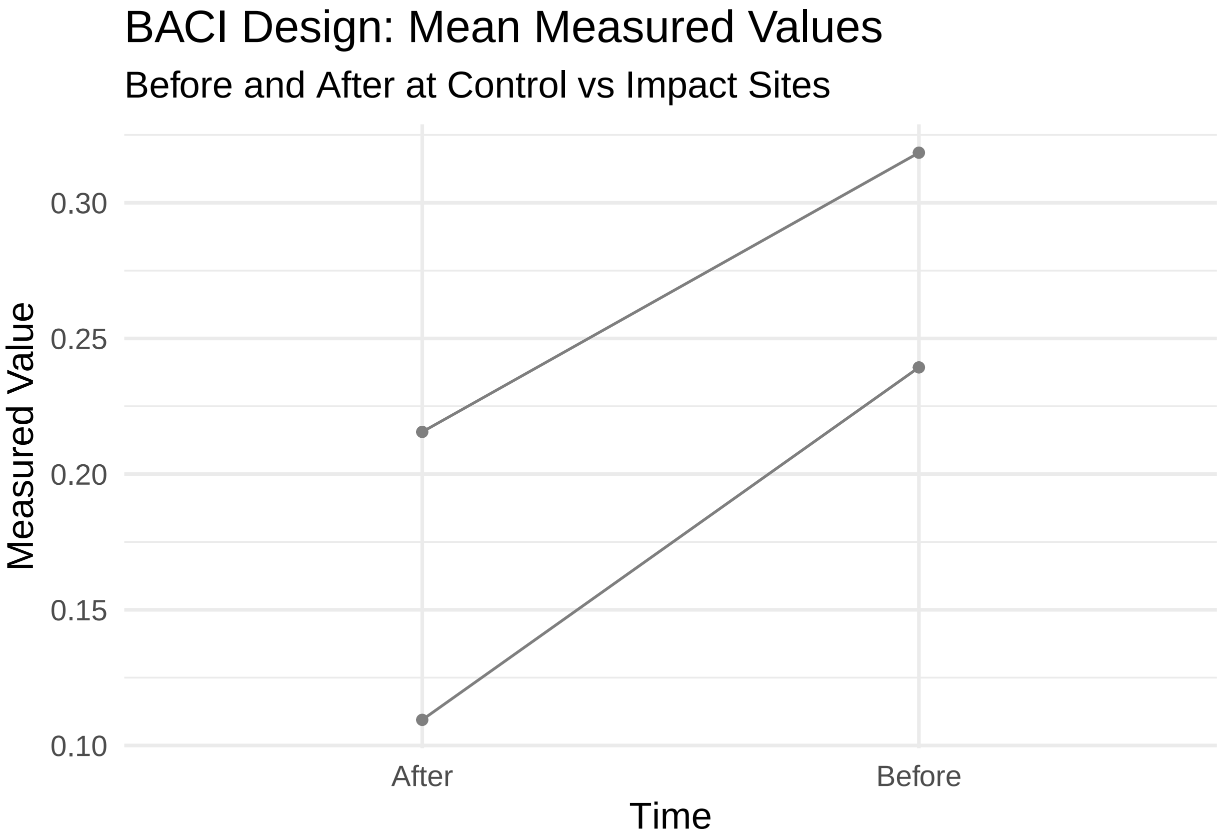
```r
model_summary <- tidy(model)
model_summary
```

```
## # A tibble: 7 x 5
##   term                                    estimate std.error statistic  p.value
##   <chr>                                      <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                              0.00836    0.0149     0.562 5.78e- 1
## 2 TreatmentCompost                         0.0926     0.0112     8.24  4.40e- 9
## 3 Concentraion..ppm.5                      0.141      0.0195     7.22  5.91e- 8
## 4 Concentraion..ppm.10                     0.183      0.0195     9.39  2.69e-10
## 5 Before.AfterBefore                      -0.0100     0.0195    -0.514 6.11e- 1
## 6 Concentraion..ppm.5:Before.AfterBefore   0.105      0.0275     3.80  6.84e- 4
## 7 Concentraion..ppm.10:Before.AfterBefore  0.274      0.0275     9.97  7.08e-11
```

```r
# Key interaction term:
key_interaction <- coef(model)["Treatmentimpact:Before.Afterafter"]
```

Plots

```r
ggplot(group1data, aes(x = Before.After, y = Absorption, color = Treatment, group = Treatment)) +
  stat_summary(fun = mean, geom = "line") +
  stat_summary(fun = mean, geom = "point") +
  labs(
    title = "BACI Design: Mean Measured Values",
    subtitle = "Before and After at Control vs Impact Sites",
    y = "Measured Value",
    x = "Time",
    color = "Treatment"
  ) +
  scale_color_manual(values = c("control" = "#1f77b4", "impact" = "#d62728")) +
  theme_minimal(base_size = 14) +
  theme(legend.position = "bottom")
```



5. Conclusion The analysis suggests:

There was a decrease in the measured value at the impact site after the event.

The significant site × time interaction supports a likely effect of the environmental disturbance.

6. Appendix (Optional) You can improve the model by considering:

Mixed models (lmer) if you have random effects (e.g., multiple sites)

Repeated measures ANOVA

Adding covariates (e.g., weather, seasonality)