KCC 2021 Oral session presentation

Causal Analysis of Observational Mobile Sensor Data: A Comparative Study

Yongshin Kim, Panyu Zhang, Gyuwon Jung, Heepyung Kim, Uichin Lee

KAIST

Contents

Contents

Mobile sensor data

- Smartphones and wearables have produced various data related to the user's behavior via their built-in sensors

- These mobile sensor data are collected using motion, position, and environment sensors, and we could provide a data-driven intervention at an opportune moment

- Previous studies have explored the relationship among variables, especially about the **"causality"**

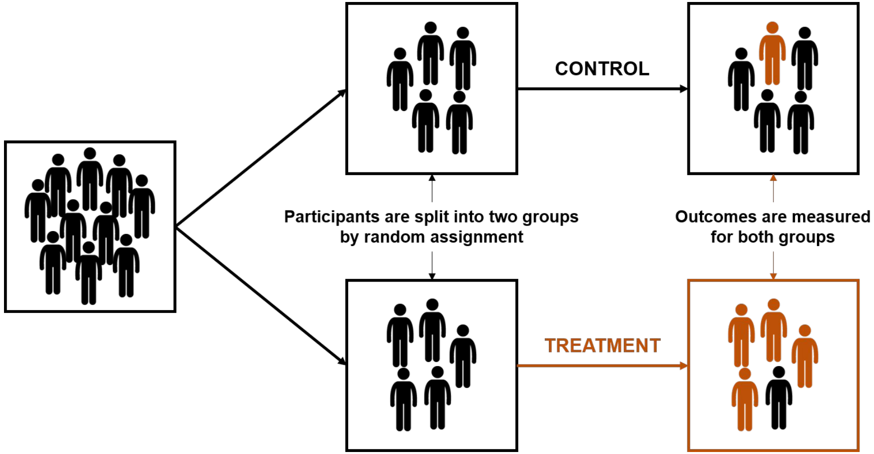
Overview

- We introduce how to analyze the causal relationship with observational mobile sensor data
  - Matching
  - Convergent Cross Mapping (CCM)

- As a case study, we describe how to implement these methods to show the existence of causality using a real-world sensor dataset (i.e., K-EmoPhone dataset)

Randomized Controlled Trial

- Most of the studies analyze the causal relationship among variables by conducting RCT
- Researchers **randomly assign** participants into two groups (control vs. treatment group)
- They examine the efficacy of treatment while **minimizing the effects of confounding variables**



Observational data

- Treatment variable is not under the control of the researcher
- Treatment assignment is no longer randomized
  - It could be biased by confounding variables

Why matching & CCM?

- Matching
  - Finding similar pairs by matching → split all pairs into two groups (control vs. treatment group)

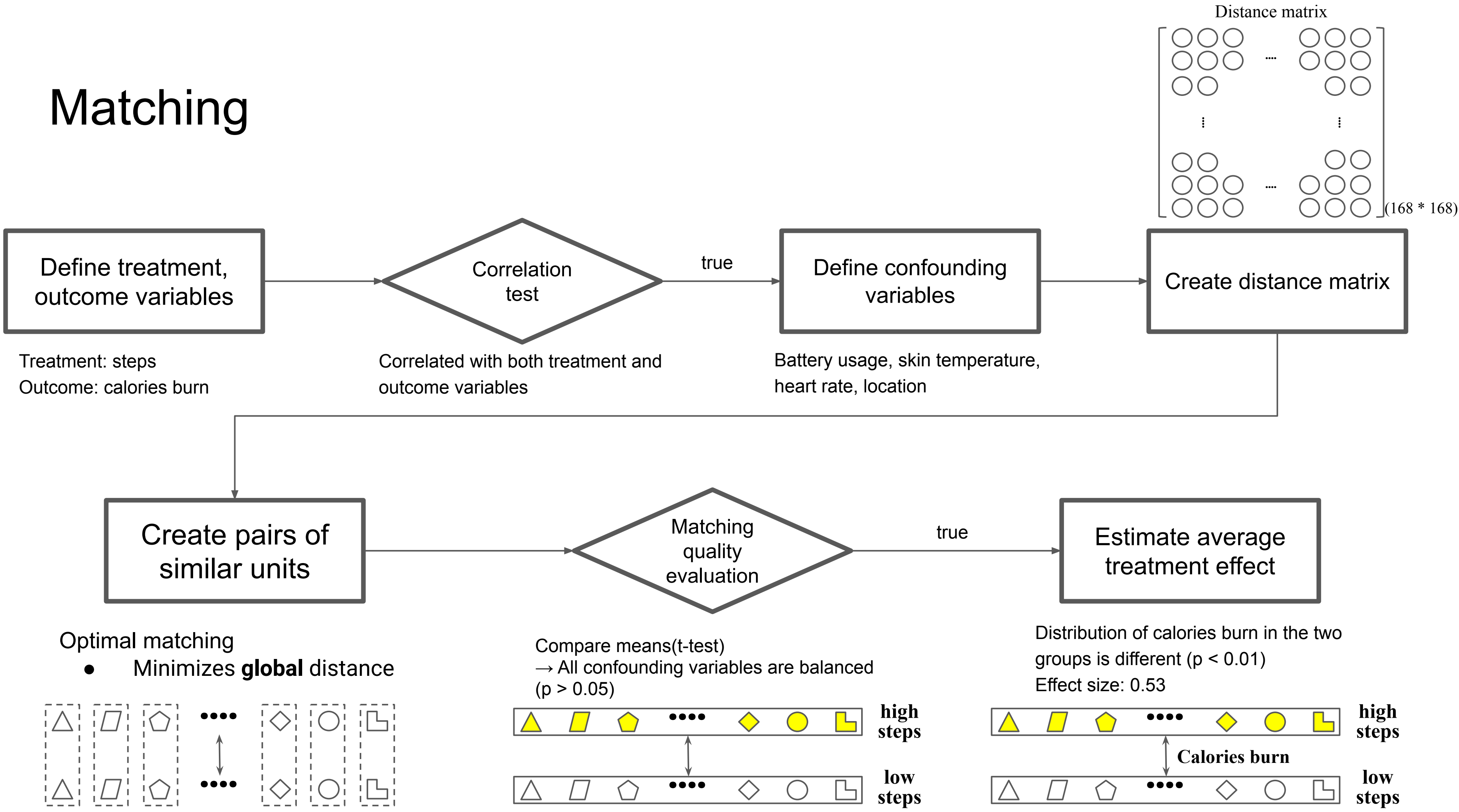
    → **treatment might be randomly assigned**
- Convergent Cross Mapping(CCM)
  - It's impossible to consider all confounding variables in reality
  - CCM allows us to estimate the causal relationship **without considering confounders**

Contents

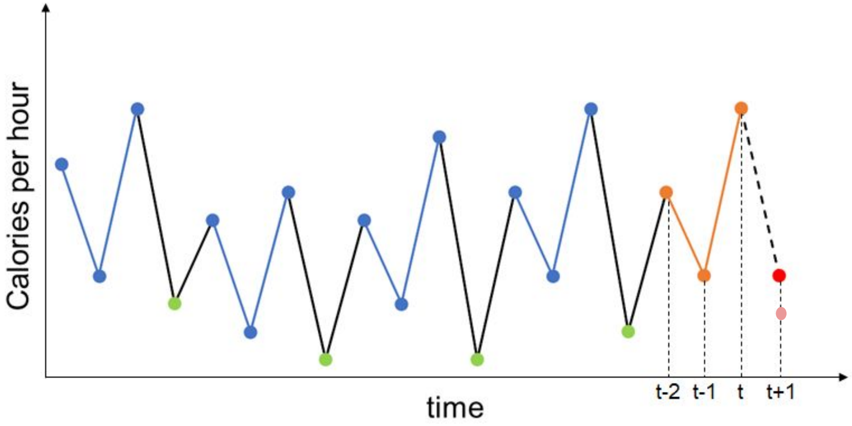
K-EmoPhone dataset

- Total subjects: 81 participants
- Period: one-week
- Device: Android smartphone, Microsoft band 2, Polar H10
- Data contents
  - **Objective sensor data**
    - Motion, Physiology, Environment, Network, Phone usage
  - Subjective information data
    - User Information, ESM data
- Preprocessing
  - Standard scaling for every variables
  - 1-hour time window for feature extraction
    - for one person, total units= 24hours * 7days= 168units
- Case study
  - N-of-1 trial(user id= 705)
  - Common scenario
    - More steps → more calories burn

Matching

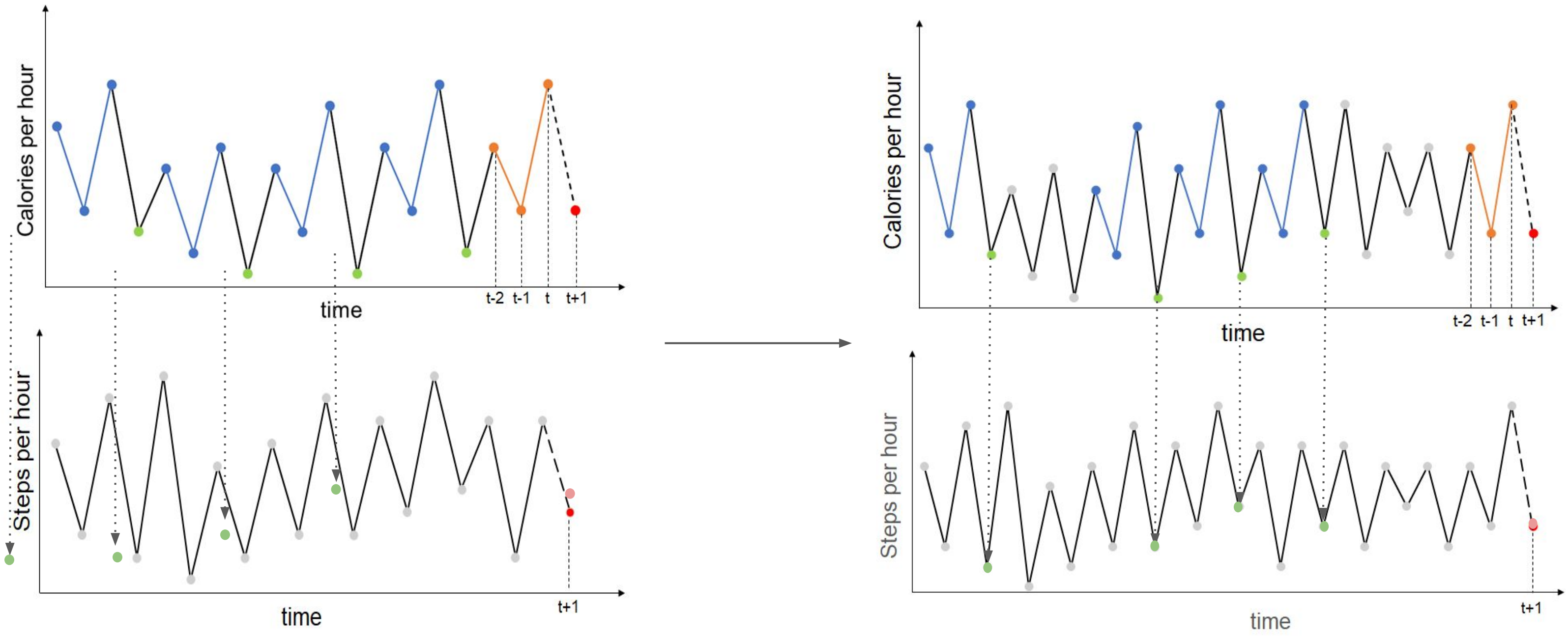

Distance matrix
(168 * 168)
Define treatment, outcome variables
Treatment: steps
Outcome: calories burn
Correlation test
Correlated with both treatment and outcome variables
true
Define confounding variables
Battery usage, skin temperature, heart rate, location
Create distance matrix
Create pairs of similar units
Optimal matching
• Minimizes **global** distance
Matching quality evaluation
Compare means(t-test)
→ All confounding variables are balanced
(p > 0.05)
high steps
low steps
true
Estimate average treatment effect
Distribution of calories burn in the two groups is different (p < 0.01)
Effect size: 0.53
high steps
Calories burn
low steps

Convergent Cross Mapping



- Deterministic dynamic systems
    - If initial condition was known, we will be able to predict the future state
    - Human behavior could also be deterministic
- Takens' idea
    - If one variable is deterministic, we may estimate its future value only using time lags of its own previous data by doing a knn in a delayed space

Convergent Cross Mapping

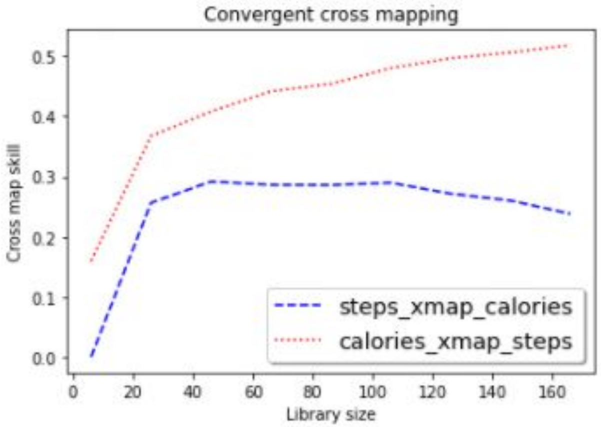


Step 1: The key question is to reconstruct or estimate steps using calories' information

Step 2: if we have more and more data of calories, we can better estimate the steps

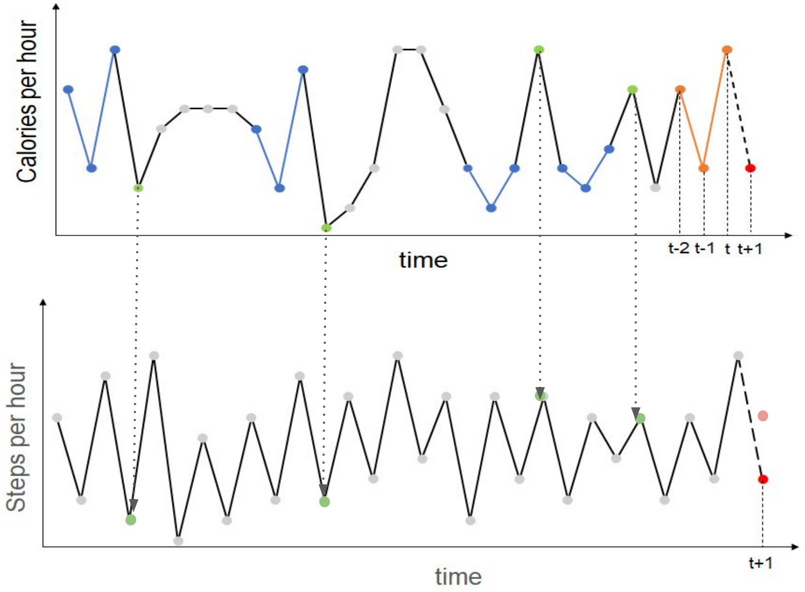
- The estimation performance will converge with more and more data which is a sign of causation

Convergent Cross Mapping



- As in Fig (a), only estimation performance of “calories estimate steps” is monotonically increasing when increasing the length of time series. The other direction is not monotonically increasing.

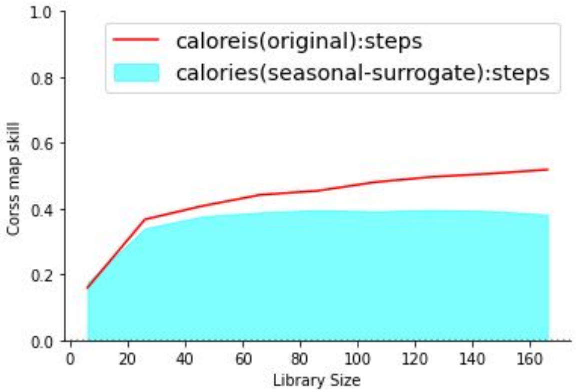
Convergent Cross Mapping



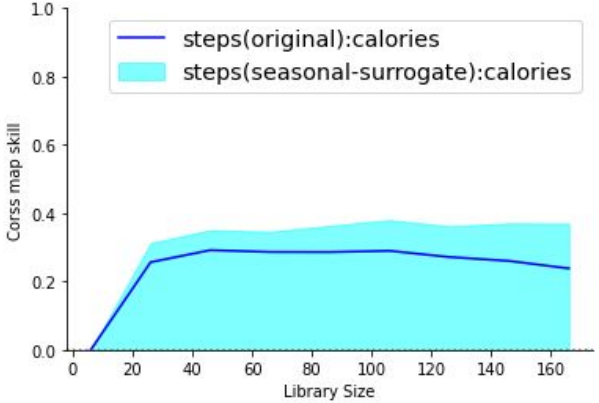
Step 3: if we randomly shuffle the order of calories, the prediction performance should be lower than original data

- If the requirement in step 2 and 3 are met, we can conclude that steps cause calories

Convergent Cross Mapping



(a)



(b)

- As in Fig (a) and (b), only estimation performance of “calories estimate steps” is higher than that of randomly shuffled time series
- In summary, steps cause calories but not vice versa

Contents

Discussion and Conclusion

- In this study, we reviewed how to perform causal analyses on observational mobile sensor data by implementing the two causal inference techniques; Matching and CCM

- These techniques could be leveraged to measuring the therapeutic efficacy of digital health interventions or optimizing user interface design

- However, when applying these causal inference techniques, note that:
  - Determining the appropriate time window size during the data preprocessing process
    - Results may vary depending on windows size
      - Overall consideration required, such as type of data, variables, etc.
      - Refer to previous domain knowledge
      - Set the optimal window size through iterative analysis
  - Distribution of confounding variables between treatment and control groups may not be equal
    - Need to tune hyperparameters to get the balance

Q & A