Imputation and Allocation of CE Data

Clayton Knappenberger
Economist
Division of Consumer Expenditure Surveys
2018 CE Microdata Users' Workshop
July 20

Outline

1. Process Overview
2. Imputation
3. Allocation
4. Edit Rates and Conclusion

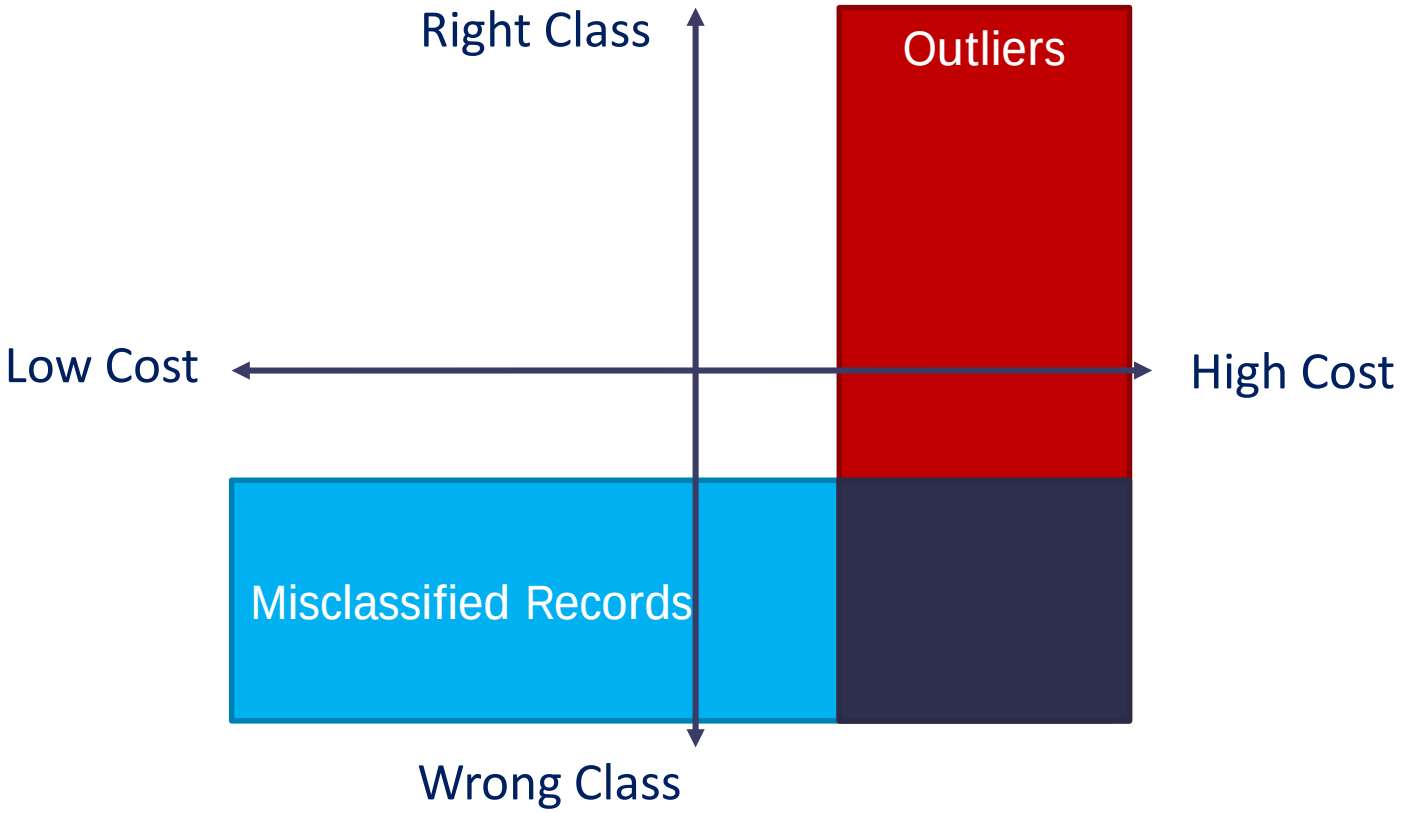
Process Overview

- CE's goal is to map expenditures
  - As monthly amounts
  - To specific Universal Classification Codes (UCCs)
  - In a specific month and year
- However, collected data are often insufficient
  - Collected information has mistakes
  - Respondent does not know or refuses to provide

Process Overview

1. **Data Screening** – check data for errors
2. **Impute** missing values
3. **Allocate** combined expenditures to components for mapping.
4. **Mapping** expenditures to months and UCCs (as well as higher level aggregations)

Data Screening



Misclassifications

- Specific keyword lookups for “hard to classify” items
  - iPad/iPhone/iPod
  - “Glasses”/”Cable”/”Nails”
- General text analysis of item descriptions
- Updates are made based on the reported item description and any interviewer notes

Outlier Review

- Three methods:
  1. Largest Gap – biggest gap between records above the mean
  2. P-Index – value divided by gap minimum
  3. Z-Score – value divided by IQR
- Updates are made by:
  1. Correcting values with available information
  2. Flag the expenditure for imputation

Imputation

1. Weighted Mean Imputation
2. Hot Deck Imputation
   - Use valid records with similar characteristics to replace missing values
3. Percent Distribution Imputation
   - Randomly select a valid value based on the percent distribution of reported values

Weighted Mean Imputation

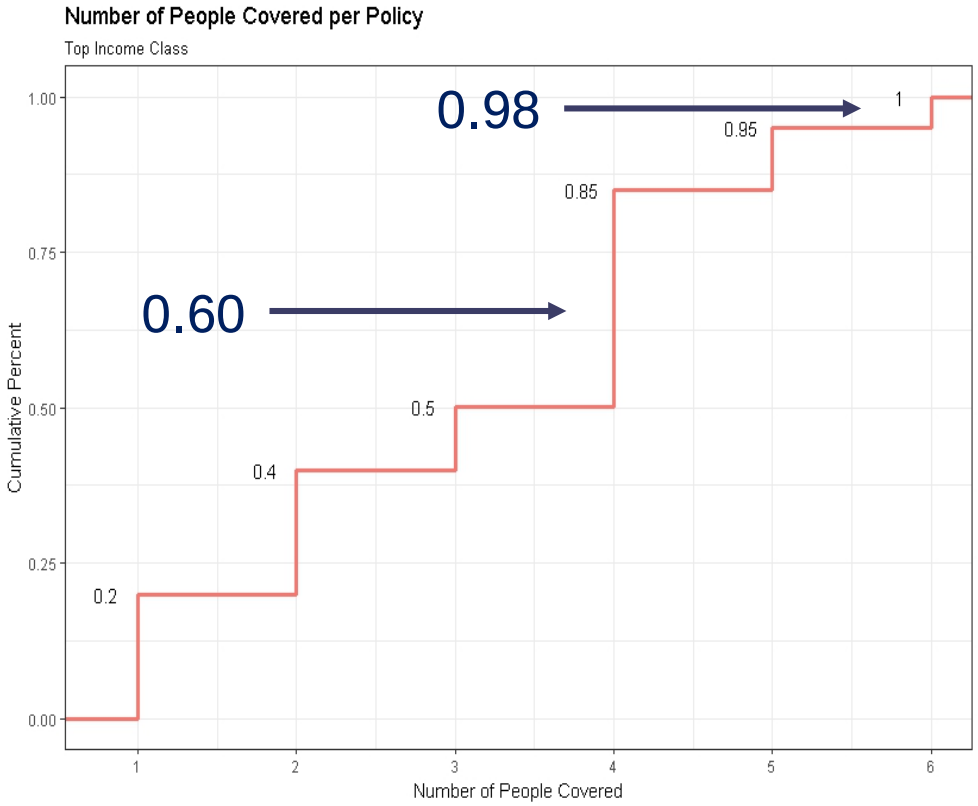
- Use valid records with similar characteristics to define cells
- Calculate the weighted mean of that cell
- Assign the weighted mean of reported expenditures within a given cell to missing or invalid expenditures in the same cell

Hot Deck Imputation Example

- A respondent reports buying a men's jacket, but does not know the cost
- Imputation steps:
  - Select a valid random men's jacket expenditure from all such purchases with the same:
    - Region
    - Area Type
    - Income Class
  - The selected record's expenditure amount is copied to the record being imputed

Percent Distribution

- A respondent does not say how many people are covered by an insurance plan



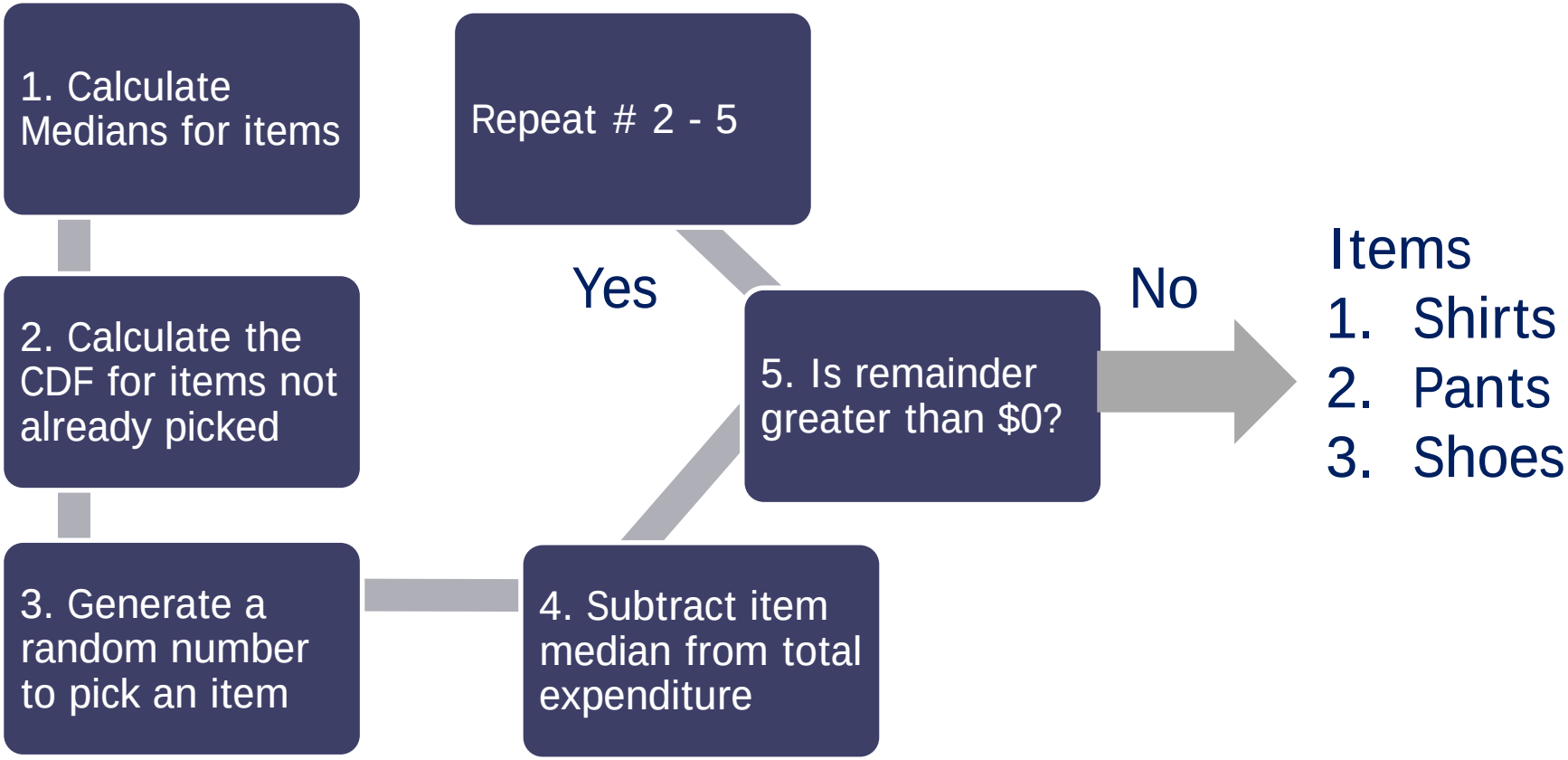
Steps

1. Create CDF
2. Get Random Number
3. Assign value

Allocation

- Example: Respondent reported spending $500 on clothing
- We need two pieces of information:
  1. Targets – shirts, pants, and shoes
  2. Allocation Proportions
     - 45% on shirts
     - 35% on pants
     - 20% on shoes

Picking the Target Items

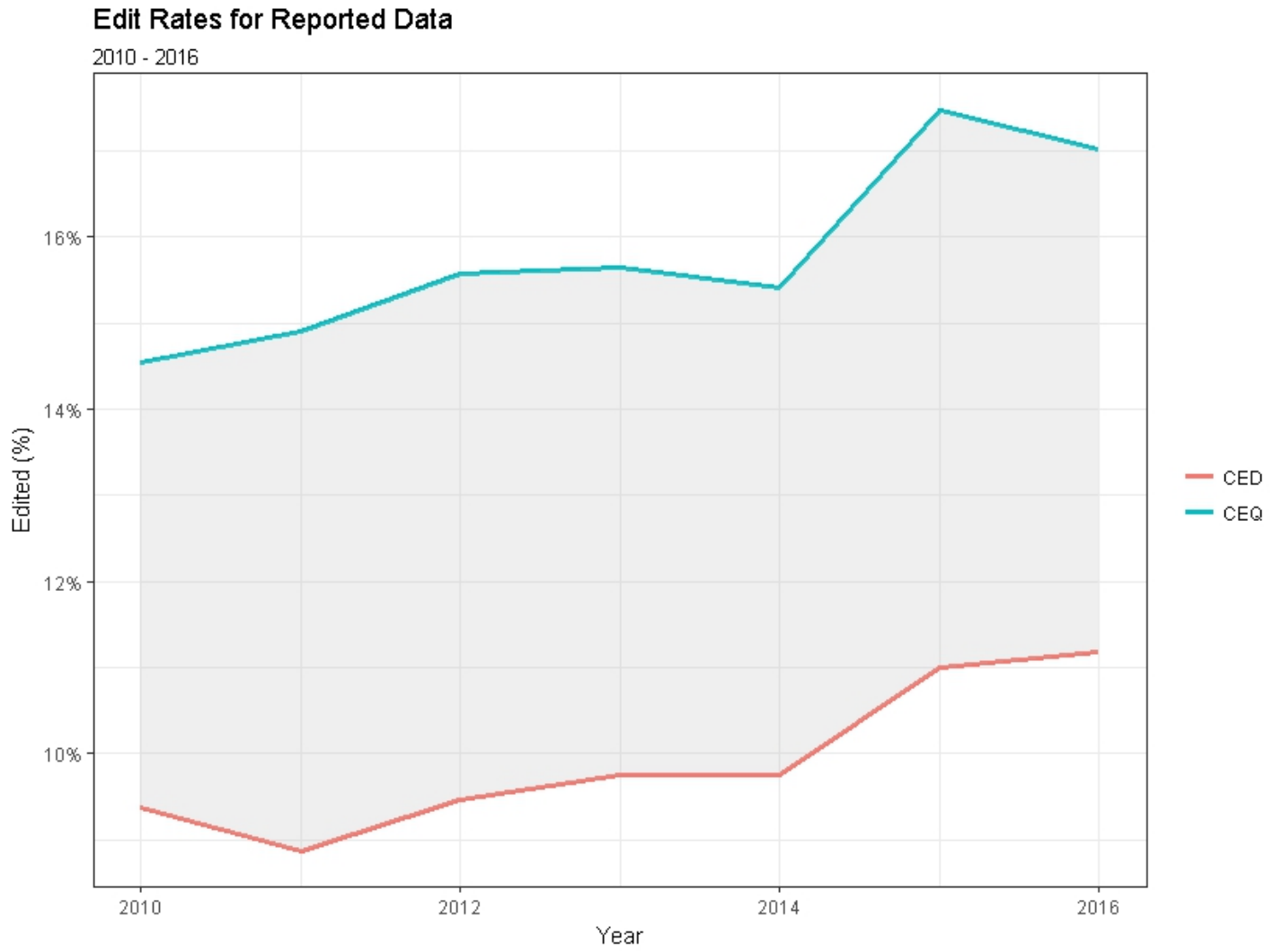


Allocating the amounts

1. Get the weighted means for each item selected
2. Calculate the share of the sum of the means
3. Derive allocation amounts

Item	Mean ($)	Share (%)	Allocation ($)
Shirts	$35.00	21.88%	$109.40
Pants	$67.00	41.87%	$209.35
Shoes	$58.00	36.25%	$181.25
Total	$160.00	100.00%	$500.00

Imputation and Allocation Rates



Why Impute and Allocate?

Benefits

- Meet internal needs for mapping
- Provide complete datasets to users

Concerns

- Our methods rely on MAR assumption
- Potential for underestimated variance

Contact Information

Clayton Knappenberger
Economist
Division of Consumer Expenditure Surveys
www.bls.gov/cex
202-691-6236
knappenberger.clayton@bls.gov