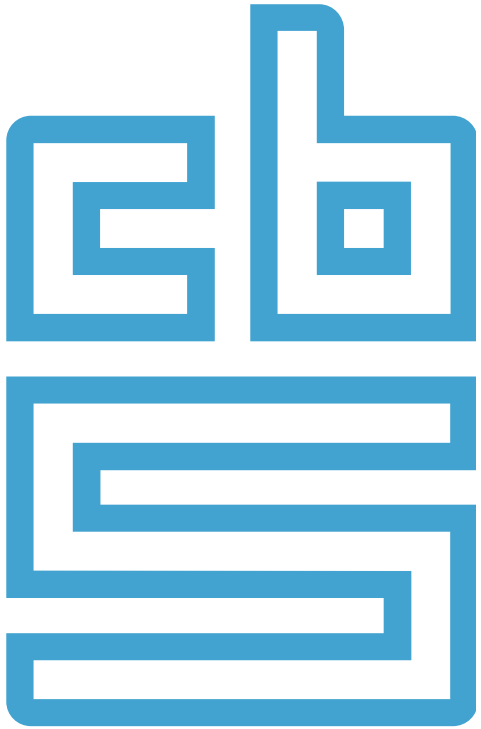


Adaptive Data Collection at Statistics Netherlands with an application to the Health Survey

Workshop on Statistical Data Collection ‘Resourceful Data Acquisition’

Kees van Berkel

10 – 12 October 2018, Geneva, Switzerland

Outline

1. Introduction
   - why Adaptive Data Collection?
2. Methodology
   - random response model
   - reducing nonresponse bias
   - stratification of target population
   - minimization problem
3. Adaptive Data Collection in the Dutch Health Survey 2018
   - about the survey
   - elaboration of the methodology
   - method effects on the survey estimates

Introduction

Why Adaptive Data Collection?

Aim of Adaptive Data Collection:

to get a better balanced response by putting different effort in different groups of the population.

Adaptive Data Collection is effective in:

improving survey results, or reducing survey costs.

$$\hat{y}_{HT} = \sum_{k \in S} y_k / \pi_k$$

Methodology

Methodology

Random response model

1. The sample is a simple random sample of size n.
2. Response follows the ‘Random response model’ in which person k responds with response probability ρ_k.
   Each ρ_k is only known to person k.
3. Answers are independent of the observation mode.

Methodology

Random response model

Aim of survey: estimation of population means
for several target variables.

An estimator for the population mean is the response mean.

In general this estimator is biased, unless all response probabilities ρ_k are equal.

Methodology

Random response model

The bias can be approximated by $\frac{R(\rho,Y)\times S(\rho)\times S(Y)}{\overline{\rho}}$,

with

Y : Target variable,

R : Pearson's correlation coefficient, $|R| \leq 1$,

S : Population standard deviation.

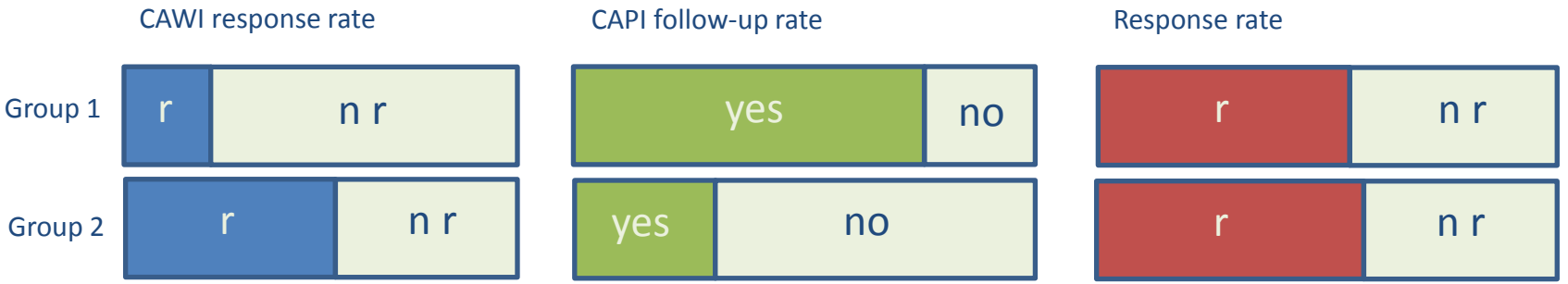
Aim: reduce bias by minimizing $CV(\rho) = S(\rho)/\bar{\rho}$.

Methodology

Adaptive Data Collection

Observation strategy: CAWI → CAPI.

Feature to adapt: CAPI follow-up.



Methodology

Determining target groups

People are divided into target groups based on personal characteristics, so that

- within each group: there is little variation in response behaviour per mode.
- between two groups: there is a big difference in response behaviour for at least one mode.

Methodology

Minimization problem

Minimize $CV(\rho)$ under constraints on

- budget,
- response numbers or rates,
- sample sizes per mode.

Solution: cawi sample size,
capi sampling fractions per target group,
estimate of $CV(\rho)$.

Adaptive Data Collection

in the Dutch Health Survey 2018

Application

Health Survey

Dutch Health Survey

- aim: describing developments in health, medical care and lifestyle
- target population: people living in the Netherlands
- sampling design: simple random sample of 1500 people per month
- observation strategy: CAWI → CAPI
- desired number of respondents: 9500 per year

Application

Determining target groups

The main personal characteristics used in determining the target groups are

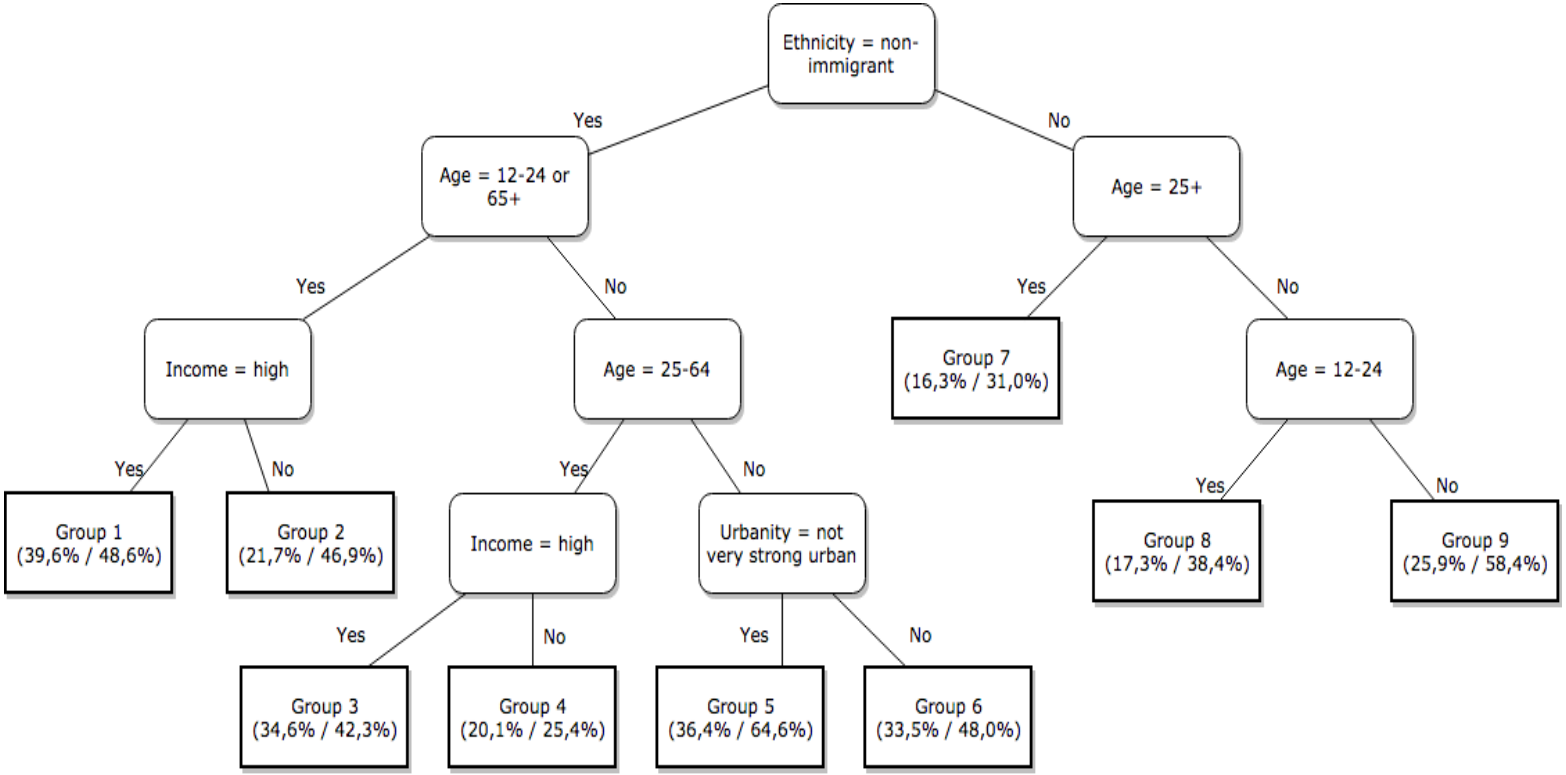
- ethnicity
- urbanization
- age
- income
- ethnicity of parents
- marital status
- educational level
- gender
- place in household
- type of household
- wealth
- home ownership

Dataset: Health Survey, January – June 2017.

Clustering is carried out through the R package rpart with which a classification tree is generated.

Application

Classification tree



Application

Target groups

Four characteristics are selected and merged into larger groups:

ethnicity: Western, non-Western
age: 0-11, 12-24, 25-64, 65+
income: low, not low
urbanization: very strongly urban, not very strongly urban

Western target groups

	income	not low		low	
age	urbanization	1	2-5	1	2-5
0-11		5	6	5	6
12-24		1	1	2	2
25-64		3	3	4	4
65+		1	1	2	2

Non-Western target groups

age	target group
0-11	9
12-24	8
25+	7

Application

Response rates per target group

Target group	P(cawi)	P(capi)
1	39.6	48.6
2	→ 21.7	46.9
3	34.6	42.3
4	→ 20.1	25.4
5	36.4	64.6
6	33.5	48.0
7	→ 16.3	31.0
8	→ 17.3	38.4
9	25.9	58.4

Application

Optimization problem

Minimize $CV(\rho) = S(\rho)/\bar{\rho}$ under constraints

1. CAWI sample size ≤ 18000.
2. Expected response size ≥ 9628.
3. CAPI sample size ≤ 8039.
4. One CAPI sampling fraction per target group.

From constraints 1 and 2 it follows that $\bar{\rho} \geq \frac{9628}{18000} = 53.5\%$.

Application

Optimization problem

Problem is solved with the R package Alabama.

The package uses the Augmented Lagrangian Adaptive Barrier Minimization Algorithm for optimizing smooth nonlinear functions with constraints.

The algorithm may end up in a local minimum, so different starting values were used and the best solution was selected.

Application

Solution

group	n cawi	r cawi	% r cawi	n elig	n capi	% n capi/ n elig	r capi	% r capi	r total	% r total
1	4550	1939	43	2567	1448	56	703	49	2642	58
2	786	195	25	553	526	95	248	47	442	56
3	7373	2770	38	4767	3405	71	1441	42	4211	57
4	728	168	23	551	551	100	140	25	308	42
5	1474	580	39	933	406	44	263	65	843	57
6	333	121	37	221	147	66	70	48	192	58
7	1276	246	19	1040	1040	100	320	31	566	44
8	411	83	20	341	341	100	132	39	215	52
9	363	105	29	266	175	66	102	58	207	57
total	17295	6208	36	11238	8039	72	3420	43	9628	56

Application

Solution

Quality indicators

Adaptive Data Collection	$\bar{\rho}$	$S(\rho)$	$CV(\rho) = \frac{S(\rho)}{\bar{\rho}}$
	%		
No	64.4	10.2	15.8
Yes	55.7	6.4	11.6

Application

Effect of adaptive data collection on survey results?

Bootstrapping:

Samples with replacement were drawn from the 2016-sample, with the correct numbers for cawi en matching numbers per target group for capi.

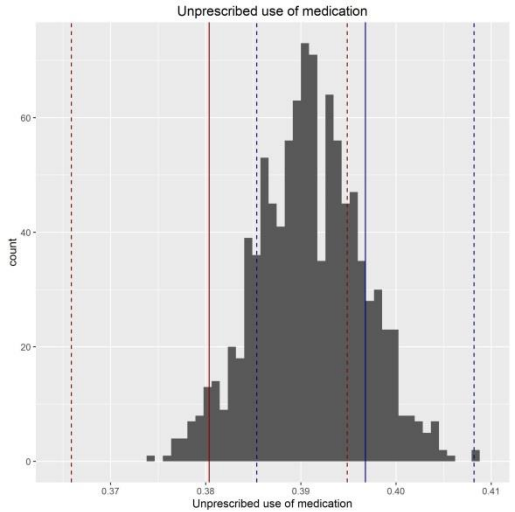
Estimates were made for the core variables of the Health Survey.

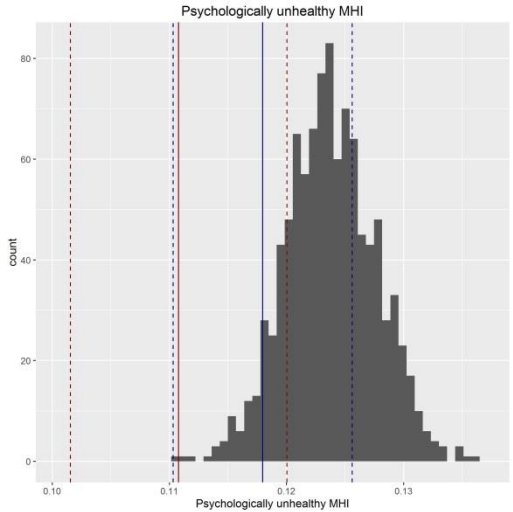
Application

Survey results

Most of the survey results with adaptive data collection do not differ much from those without adaptation.

The greatest shifts:





End of talk