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

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.

Why Adaptive Data Collection?
Methodology

\[ \hat{y}_{HT} = \sum_{k \in S} \frac{y_k}{\pi_k} \]
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 \( \rho_k \). Each \( \rho_k \) 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 $\rho_k$ are equal.
Methodology

Random response model

The bias can be approximated by

\[
\frac{R(\rho, Y) \times S(\rho) \times S(Y)}{\bar{\rho}}
\]

with

\[Y\] : Target variable,
\[R\] : Pearson’s correlation coefficient, \(|R| \leq 1,\)
\[S\] : Population standard deviation.

Aim: reduce bias by minimizing

\[
CV(\rho) = \frac{S(\rho)}{\bar{\rho}}.
\]
Methodology

Adaptive Data Collection

Observation strategy: CAWI → CAPI.

Feature to adapt: CAPI follow-up.

CAWI response rate

Group 1
- r
- n
- r

Group 2
- r
- n
- r

CAPI follow-up rate

- yes
- no

Response rate

- r
- n
- r

- yes
- no

- r
- n
- r
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

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
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


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

Classification tree

Ethnicity = non-immigrant

Age = 12-24 or 65+
- Yes
  - Income = high
    - Yes
      - Group 1 (39.6% / 48.6%)
    - No
      - Group 2 (21.7% / 46.9%)
- No
  - Age = 25-64
    - Yes
      - Group 4 (20.1% / 25.4%)
    - No
      - Group 3 (34.6% / 42.3%)

Age = 25+
- Yes
  - Group 7 (16.3% / 31.0%)
- No
  - Age = 12-24
    - Yes
      - Urbanity = not very strong urban
        - Yes
          - Group 6 (33.5% / 48.0%)
        - No
          - Group 5 (36.4% / 64.6%)
    - No
      - Group 8 (17.3% / 38.4%)

Group 9 (25.9% / 58.4%)
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**

<table>
<thead>
<tr>
<th>income</th>
<th>not low</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>urbanization</td>
<td>1</td>
</tr>
<tr>
<td>0-11</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>12-24</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>25-64</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>65+</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

**Non-Western target groups**

<table>
<thead>
<tr>
<th>age</th>
<th>target group</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-11</td>
<td>9</td>
</tr>
<tr>
<td>12-24</td>
<td>8</td>
</tr>
<tr>
<td>25+</td>
<td>7</td>
</tr>
</tbody>
</table>
### Response rates per target group

<table>
<thead>
<tr>
<th>Target group</th>
<th>P(cawi)</th>
<th>P(capi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>39.6</td>
<td>48.6</td>
</tr>
<tr>
<td>2</td>
<td>21.7</td>
<td>46.9</td>
</tr>
<tr>
<td>3</td>
<td>34.6</td>
<td>42.3</td>
</tr>
<tr>
<td>4</td>
<td>20.1</td>
<td>25.4</td>
</tr>
<tr>
<td>5</td>
<td>36.4</td>
<td>64.6</td>
</tr>
<tr>
<td>6</td>
<td>33.5</td>
<td>48.0</td>
</tr>
<tr>
<td>7</td>
<td>16.3</td>
<td>31.0</td>
</tr>
<tr>
<td>8</td>
<td>17.3</td>
<td>38.4</td>
</tr>
<tr>
<td>9</td>
<td>25.9</td>
<td>58.4</td>
</tr>
</tbody>
</table>
Application

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

1. CAWI sample size $\leq 18000$.
2. Expected response size $\geq 9628$.
3. CAPI sample size $\leq 8039$.
4. One CAPI sampling fraction per target group.

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

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

### Quality indicators

<table>
<thead>
<tr>
<th>Adaptive Data Collection</th>
<th>$\bar{\rho}$</th>
<th>$S(\rho)$</th>
<th>$CV(\rho) = \frac{S(\rho)}{\bar{\rho}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>64.4</td>
<td>10.2</td>
<td>15.8</td>
</tr>
<tr>
<td>Yes</td>
<td>55.7</td>
<td>6.4</td>
<td>11.6</td>
</tr>
</tbody>
</table>
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