Optimal scheduling of contact attempts in mixed-mode surveys

Melania Calinescu, Sandjai Bhulai and Barry Schouten

Discussion paper (201131)

Statistics Netherlands

The Hague/Heerlen, 2011

Explanation of symbols

.	= data not available
*****	= provisional figure
******	= revised provisional figure (but not definite)
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0.0)	= less than half of unit concerned
blank	= not applicable
2010–2011	= 2010 to 2011 inclusive
2010/2011	= average of 2010 up to and including 2011
2010/’11	= crop year, financial year, school year etc. beginning in 2010 and ending in 2011
2008/’09–2010/’11	= crop year, financial year, etc. 2008/’09 to 2010/’11 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands
Grafimedia

Cover
Teldesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

60083201131 X-10

Optimal scheduling of contact attempts in mixed-mode surveys

Melania Calinescu, Sandjai Bhulai and Barry Schouten

Abstract

Survey nonresponse occurs when members of a sample cannot or will not participate in the survey. It remains a problem despite the development of statistical methods that aim to reduce nonresponse. In this paper, we address the problem of resource allocation in survey designs in which the focus is on the quality of the survey results given that there will be nonresponse. We propose a novel method in which the optimal allocation of survey resources can be determined. We demonstrate the effectiveness of our method by extensive numerical experiments.

Keywords: dynamic programming; non-linear optimization; Markov decision theory; survey design; resource allocation.

1 Introduction

Surveys are used all around the world to measure socio-economic status and well-being of people, to test theories, or to make investment decisions, driven by the impossibility of observing the entire population of interest. No matter what the framework of a survey is, its success relies on the active participation of the sampled households and businesses. Unit nonresponse occurs when members of a sample cannot or will not participate in the survey. The impact of nonresponse appears in the inability of computing a full-sample estimator of the population mean. Thus, a bundle of practical issues is created, including bias in point estimates, bias in estimators of precision, and inflation of the variance of point estimators. The error caused by nonresponse is one of the several sources of error in surveys and it has attracted a great deal of interest among researchers across the world, see [3] for an extensive overview on survey nonresponse both in research and practice.

Lately, survey response rates seem to have decreased [3]. This aspect has a significant negative impact on the quality of survey statistics. However, the response rate alone is not a sufficient measure of the impact of nonresponse. The contrast between respondents and nonrespondents with respect to a target variable prevents the survey statistics from reflecting the corresponding characteristics of the target population, also called the representativeness of the respondent sample. An apparent solution to the problem is to increase the frequency of attempts to gather information from the reluctant sample members. Under these circumstances the costs of conducting surveys increase significantly, which leads to new problems, such as budget overruns. Therefore, a constant scientific challenge to survey methodology concerns the improvement of quality statistics by identifying methods that alter the estimation process and adequately handle the occurrence of nonresponse or developing new designs to accommodate the presence of nonresponse.

When people decide to accept or reject the request to participate in sample surveys, there are various factors that influence their decision. Such factors include the social characteristics of the sample unit, the interviewer's behavior, and the attributes of the survey design, e.g., the interview mode, the schedule of the contact attempts, and the language of questionnaire. Traditional survey designs do not use the information that results from studying these factors. Predicting the probability that a sample unit would respond given a list of factors, quantifying the factors' effects on human behavior in terms of costs and data quality can be of great help in designing a quality but cost-effective survey.

While the research in the field is still in its infancy, there have been some design methods investigated or implemented. The first attempts belong to [4], where the authors describe a two-phase sampling framework for nonresponse. This approach has been termed as *responsive survey design*. More advanced methods in the same context of *responsive design* are later presented in [5] and [9], where the main idea is to identify a set of design attributes that potentially influence the survey costs and the errors in the survey estimates and to monitor their influence on costs and errors. This information helps in subsequent phases to alter the design features such that a desired balance between costs and errors is achieved. More on trade-offs between various survey errors can be found in [1] and [8].

More studies on the impact of the timing of call attempts on cooperation of sample units can be found in, for example, [2], where an optimal call scheduling strategy is computed based on historical data. By historical data we understand previous versions of the same survey or similar surveys that can provide the required information. In [6] the relationship between costs, number of call attempts allowed, and cooperation is investigated. The study in [7] indicates what time slots are optimal to schedule calls, emphasizing the importance of optimal timing for the first call. Optimal timing protocols address two causes of nonresponse, unknown non-eligibility and non-contact.

The historical data considered in the aforementioned studies do not depend on person or household characteristics, the use of which could bring improvements in the response rate predictions. Such characteristics are available from administrative registers, financial records, etc. When such information is employed to adjust the design features for a given set of characteristics (i.e., different design features can be applied to different sample units) the resulting survey design is termed *adaptive survey design* and it has been generally introduced in [13] and [11]. The method originates from the field of clinical trials, where treatments are group-specifically set before the start of the trial and change during the trial according to the responses of patients. Therefore, the great advantage of this method lies in its flexibility. It is tailored for groups of sample units, it can be defined before the survey starts and also updated during the data collection based on information regarding the characteristics of respondents and nonrespondents.

In the present paper, adaptive survey designs are analyzed from the perspective of resource allocation problems, which constitutes the novelty of this research. Given a budget, a set of household characteristics, and a list of factors that influence the survey costs and quality, we develop a model that computes the allocation of survey resources such that quality is maximized while costs meet the budget constraint. Extracting detailed information from historical data and building a survey design that is both cost-efficient and of high-quality are the main contributions the current paper brings to practice and the research in the field.

The remainder of the paper is structured as follows. Section 2 discusses the mathematical model and Section 3 discusses algorithms to derive optimal adaptive survey design policies. Section 4 presents a range of practical problems that can be handled through this model and solution method. Numerical examples of these situations are given in Section 5. Section 6 concludes the results of the paper and gives directions for future research.

2 Problem formulation

Consider a survey sample consisting of N units that can be clustered into homogeneous groups based on characteristics, such as age, gender, and ethnicity (information that can be extracted from external sources of data). Let $\mathcal{G} = \{1, \dots, G\}$ be the set of homogeneous groups with size N_g for group $g \in \mathcal{G}$ in the survey sample. The survey fieldwork is divided into time slots (e.g., morning, evening, day of the week), denoted by the set $\mathcal{T} = \{1, \dots, T\}$, at which units in a group can be approached for a survey. The survey itself can be conducted using certain interview modes, such as a face-to-face, phone, web/paper survey; the set of different modes is denoted by $\mathcal{M} = \{1, \dots, M\}$. In practice, the size of set $\mathcal{G}$ can vary greatly (e.g., $G = 3$, $G = 100$), depending on the number of characteristics considered for clustering. The size of set $\mathcal{T}$ depends on the interpretation of a time slot. For example, for a 30-day data collection period, $T = 30$ if a time slot denotes one day, $T = 90$ if a time slot denotes a period of the day (e.g., morning, afternoon, evening) or $T = 720$ if the time slot denotes one hour. The size of set $\mathcal{M}$ depends on the number of interview modes available for the given survey.

At each time slot $t \in \mathcal{T}$ one can decide to approach units in group $g \in \mathcal{G}$ for a survey using mode $m \in \mathcal{M}$. In doing so, successful participation in the survey depends on first establishing contact, and then be responsive by answering the questionnaire. From historical data group-dependent contact probabilities $p_g(t, m)$ and cooperation probabilities $r_g(t, m)$ can be estimated, which we consider as given quantities in our problem. These quantities reflect that given the time slot t and the interview mode m, sample units from group g are contacted with probability $p_g(t, m)$ and cooperate with probability $r_g(t, m)$.

Denote by $x_g(t, m)$ a binary 0-1 decision variable that denotes if units in group g are approached for a survey at time t using mode m. Two events are necessary in order to obtain response, successful contact followed by cooperation. Then, the probability of response is given by probability $p_g(t, m)r_g(t, m)$. Similarly, refusal after successful contact occurs with probability $p_g(t, m)\big(1 - r_g(t, m)\big)$. In the current paper, we assume that once contacted, the sample unit either cooperates or refuses to participate in the survey, i.e., we do not allow for the possibility of making an appointment for a future approach. Moreover, we do not take into account refusal conversions. As a consequence, only unsuccessful contacts are considered for a future survey approach; this happens with probability $1 - p_g(t, m)$. Thus, if the unit is approached again at time t' using mode m', then the probability of response is $\big(1 - p_g(t, m)\big)p_g(t', m')r_g(t', m')$, and the probability of a contact failure is $\big(1 - p_g(t, m)\big)\big(1 - p_g(t', m')\big)$. We here assume that both contact and participation probabilities are independent of the contact history. Although such an independence assumption is not uncommon to practice in the case of cooperation probabilities, it may be a too strong assumption for the contact probabilities. Currently, in deriving historical contact probabilities we fit geometrical distributions to call/visit record data. As a consequence, this may lead to pessimistic contact probabilities after the first one, two calls/visits and optimistic probabilities after a larger number of calls/visits. For example, by the geometrical distribution the contact probability goes to 1, whereas in practice, there would always exist non-contacted sample units. In future research, we will investigate extensions of the current model where the independence assumption is relaxed.

By the independence assumption, the probability that a contact would fail at time t' is denoted by $f_g(t')$ given by

$$\begin{aligned} f_g(t') &= \prod_{t=1}^{t'} \prod_{m \in \mathcal{M}} \big[x_g(t, m)\big(1 - p_g(t, m)\big) + \big(1 - x_g(t, m)\big)\big] \\ &= \prod_{t=1}^{t'} \prod_{m \in \mathcal{M}} \big[1 - x_g(t, m)p_g(t, m)\big]. \end{aligned}$$

Note that this is a highly non-linear expression in the decision variables, which can be recursively computed by

$$
\begin{aligned}
f_g(t') &= \prod_{m \in \mathcal{M}} \big[x_g(t',m)\big(1 - p_g(t',m)\big) + \big(1 - x_g(t',m)\big)\big] f_g(t'-1) \\
&= \prod_{m \in \mathcal{M}} \big[1 - x_g(t',m) p_g(t',m)\big] f_g(t'-1),
\end{aligned}
\tag{1}
$$

using the fact that $f_g(0) = 1$. In other words, $f_g(t')$ denotes the proportion of group g that will be handled after time t'. Using this definition, the response rate for group g can then by computed by

$$
\sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{M}} f_g(t-1)\, x_g(t,m)\, p_g(t,m)\, r_g(t,m).
$$

The clustering of the N units usually results in groups that are not of the same size or importance. Therefore, the response rates for the groups are usually weighted by a factor w_g (e.g., $w_g = N_g/N$ is taken in practice). Hence, the overall response rate is computed as

$$
\sum_{g \in \mathcal{G}} \sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{M}} w_g\, f_g(t-1)\, x_g(t,m)\, p_g(t,m)\, r_g(t,m). \tag{2}
$$

In the current paper, we consider that the objective of the decision maker is to maximize the average response as given by (2) by setting the decision variables $x_g(t,m)$ optimally. As mentioned in [11] however, aiming for a high response rate without taking into consideration the quality of the response thus obtained is not desirable in practice. Nevertheless, for the sake of modeling simplicity we prefer the response rate as an objective function. Section 6 gives some intuition on the increase in the model complexity when the quality of the response is considered.

The decision variables are subject to constraints, due to scarcity in resources. In practice, due to resource management constraints, the number of times that a group can be approached by mode m is limited to $\bar{k}_g(m)$ times, leading to the constraint $\sum_{t \in \mathcal{T}} x_g(t,m) \leq \bar{k}_g(m)$. For simplicity, we assume that at time t only one mode can be employed to approach a group, yielding the constraint $\sum_{m \in \mathcal{M}} x_g(t,m) \leq 1$. By combining the objective with all the constraints, we can formulate our optimization problem as a binary programming problem in the following manner.

$$
\begin{aligned}
\max \ & \sum_{g\in\mathcal{G}}\sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{M}} w_g\, f_g(t-1)\, x_g(t,m)\, p_g(t,m)\, r_g(t,m) \\
\text{s.t.} \ & \sum_{t\in\mathcal{T}} x_g(t,m) \le \bar{k}_g(m),\ \forall g\in\mathcal{G},\ \forall m\in\mathcal{M}, \\
& \sum_{m\in M} x_g(t,m) \le 1,\ \forall g\in\mathcal{G},\ \forall t\in\mathcal{T}, \\
& f_g(t) = \prod_{m\in\mathcal{M}} \big[x_g(t,m)\big(1-p_g(t,m)\big) + 1 - x_g(t,m)\big] f_g(t-1),\ \forall g\in\mathcal{G},\ \forall t\in\mathcal{T}, \\
& f_g(0) = 1,\ \forall g\in\mathcal{G}, \\
& x_g(t,m) \in \{0,1\},\ \forall g\in\mathcal{G},\ \forall t\in\mathcal{T},\ \forall m\in\mathcal{M}.
\end{aligned}
\tag{3}
$$

Problem (3) represents the adaptive survey design problem in which survey features with significant influence on the quality of the survey are balanced. In our model, the features are the *interview mode*, the *number of allowed attempts*, and *number of time slots*. The solution of the problem is, however, not trivial. The objective function is a non-convex non-linear function, and the constraints do not form a convex polytope either. As a consequence, our problem is non-tractable even for small-sized problems (e.g., 1 group and 4 time slots!). In the next section, we develop an algorithm that is able to derive optimal solutions by aggregating information in the adaptive survey design problem.

3 Adaptive survey design policies

In this section, we reformulate the adaptive survey design problem such that the problem becomes numerically tractable and consequently applicable in practice. In order to do this, note that at any time t, it is sufficient to know $f_g(t)$ instead of the complete configuration $x_g(t', m)$ for $t' \le t$ for all g. Hence, given $f_g(T)$, the decision at time T is obvious when one also keeps track of the number of times that mode m has been used for each group g. Since the decision at time T is completely determined, one can then calculate the optimal decisions at time $T-1$, and continue working back towards the first time epoch. By keeping track of the time, the contact failure probability, and the utilization of the different modes, the problem becomes completely Markovian and the problem can be cast as a Markov decision problem.

Let the state space of the Markov decision problem be denoted by $\mathcal{S} = \mathcal{T} \times [0,1]^G \times \mathbb{N}^{G \cdot M}$, where $s = (t, \vec{f}, K) \in \mathcal{S}$ has components t, denoting the time at which the process resides, $\vec{f} = (f_1, \ldots, f_G)$, denoting the vector storing the probability of contact failure up to time t, and $K = \big(k_g(m)\big)_{g \in \mathcal{G}, m \in \mathcal{M}}$ denoting that mode m can still be used $k_g(m)$ times for group g. The action space $\mathcal{A}_s$, that depends on the state s the process is in, is given by

$$\mathcal{A}_s = \big\{ \big(a_g(m)\big)_{g \in \mathcal{G}, m \in \mathcal{M}} \in \{0,1\}^{G \cdot M} \mid a_g(m) \leq k_g(m), \sum_{m \in \mathcal{M}} a_g(m) \leq 1 \big\}.$$

The transition probabilities are defined as follows

$$p(s, a, s') = \begin{cases} 1 & \text{if } s' = (t', \vec{f'}, K') \\ 0 & \text{for all other } s', \end{cases}$$

where $t' = t+1$, $\vec{f'}$ is computed component-wise as $f'_g = \prod_{m \in \mathcal{M}} \big[1 - a_g(m) p_g(t,m)\big] f_g$, and K' is given by its components $k'_g(m) = k_g(m) - a_g(m), \forall\, g \in \mathcal{G}, m \in \mathcal{M}$. The rewards are given by

$$r(s, a) = \sum_{g \in \mathcal{G}} \sum_{m \in \mathcal{M}} w_g a_g(m) f_g p_g(t,m) r_g(t,m).$$

The tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, with components $\mathcal{A} = (\mathcal{A}_s)_{s \in \mathcal{S}}, \mathcal{P} = \big(p(s,a,s')\big)_{s, s' \in \mathcal{S}}$ and $\mathcal{R} = \big(r(s,a)\big)_{s \in \mathcal{S}}$, completely defines the Markov decision problem (see also [10]).

The Markov decision problem can be solved by dynamic programming (or backward recursion) by formulating the recursion equations for state $s = (t, \vec{f}, K)$ given by

$$\begin{aligned} V(s) &= \max_{a \in \mathcal{A}_s} \big[r(s,a) + \sum_{s' \in \mathcal{S}} p(s,a,s') V(s') \big] \\ &= \max_{a \in \mathcal{A}_s} \Big[\sum_{g \in \mathcal{G}} \sum_{m \in \mathcal{M}} w_g a_g(m) f_g p_g(t,m) r_g(t,m) + \\ &\qquad V\big(t+1, \big(\prod_{m \in \mathcal{M}} \big[1 - a_g(m) p_g(t,m)\big] f_g \big)_{g \in \mathcal{G}}, \big(k_g(m) - a_g(m)\big)_{g \in \mathcal{G}, m \in \mathcal{M}}\big) \Big], \end{aligned} \tag{4}$$

and by setting $V(s) = 0$ for all $s = (T+1, \vec{f}, K)$. Note that the algorithm only needs T iterations, and in each iteration only $2^{G \cdot M}$ actions need to be considered. Hence, for values of realistic size, the algorithm is computationally feasible and is guaranteed to converge to the optimal solution. The weighted response rate is then given by $V(s)$ with $t = 1$, $\vec{f} = \big(f_g = 1\big)_{g \in \mathcal{G}}$, and $K = \big(\bar{k}_g(m)\big)_{g \in \mathcal{G}, m \in \mathcal{M}}$.

4 Features of the model

In the previous section, we formulated the adaptive survey design problem in which the focus was on the quality of the survey results modeled by maximizing the weighted response rates. However, the model formulation is sufficiently flexible to include other features as well, such as budgetary constraints or capacity restrictions. In this section, we discuss how these features can be integrated within our existing framework.

First, we consider a constraint on the budget. There are various categories of costs that occur throughout a survey run, e.g., sample handling, interviewer training, interviewer salary, traveling costs. For a survey organization that runs multiple surveys at the same time, such as Statistics Netherlands, rigorously accounting for these costs is a highly complex task. Building a cost function that computes the total costs for a given survey for a given set of values of decision variables becomes then a daunting task. For this reason, in the current paper we assume that the costs have been aggregated and depend only on the interview mode and the outcome of the approach. Denote by $b^s(m)$ the costs that are incurred by using mode m with a successful outcome. For the costs that are incurred by mode m that results in a failure, we distinguish two types of costs: $b^{f_c}(m)$ when the failure occurs due to failure of contact, and $b^{f_r}(m)$ when the failure occurs due to failure to respond. Let B be the total budget that is available for the survey. An approach at time t using mode m bears the following costs

$$p_g(t,m)\big[r_g(t,m)b^s(m) + \big(1 - r_g(t,m)\big)b^{f_r}(m)\big] + \big(1 - p_g(t,m)\big)b^{f_c}(m).$$

In general, the costs $b_g(t,m)$ at time t using mode m depend on the contact failures before time t. These costs can be written as follows

$$b_g(t,m) = x_g(t,m) \cdot$$
$$f_g(t-1)\Big[p_g(t,m)\big[r_g(t,m)b^s(m) + \big(1-r_g(t,m)\big)b^{f_r}(m)\big] + \big(1-p_g(t,m)\big)b^{f_c}(m)\Big],$$

with $f_g(t)$ given by (1). Hence, using this definition, the budgetary constraint that needs to be added to problem (3) is given by

$$\sum_{g\in\mathcal{G}}\sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{M}} N_g b_g(t,m) \leq B.$$

A capacity constraint can be addressed in a manner analogous to the constraint on the budget. Let C be the available capacity, measured by the number of interviewer hours available to survey the sample. For the sake of simplicity, our current model does not take into consideration the scheduling of the interviewer. Similar to the budgetary cost structure, the required capacity depends on the interview mode and on the outcome of each approach. Denote by $c^s(m)$, $c^{f_c}(m)$, and $c^{f_r}(m)$ the capacity utilized when the approach is successful, or has failed due to contact failure, or failed due to a non-respondent, respectively. Following the same steps as above, the capacity constraint to be added to problem (3) is given by

$$\sum_{g\in\mathcal{G}}\sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{M}} N_g c_g(t,m) \leq C,$$

with $c_g(t,m)$ defined as

$$c_g(t,m) = x_g(t,m) \cdot$$
$$f_g(t-1)\Big[p_g(t,m)\big[r_g(t,m)c^s(m) + \big(1-r_g(t,m)\big)c^{f_r}(m)\big] + \big(1-p_g(t,m)\big)c^{f_c}(m)\Big].$$

Note that if the budgetary constraint and the capacity limitation are added to the model, then the maximum number of attempts $\bar{k}_g(m)$ is no longer required. However, a constraint that enforces a maximum number of visits can easily be added to the model. Hence, the binary programming problem becomes

$$
\begin{aligned}
\max \quad & \sum_{g\in\mathcal{G}}\sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{M}} w_g\, f_g(t-1)\, x_g(t,m)\, p_g(t,m)\, r_g(t,m) \\
\text{s.t.} \quad & \sum_{g\in\mathcal{G}}\sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{M}} N_g b_g(t,m) \le B, \\
& \sum_{g\in\mathcal{G}}\sum_{t\in\mathcal{T}}\sum_{m\in\mathcal{M}} N_g c_g(t,m) \le C, \\
& \sum_{m\in M} x_g(t,m) \le 1, \ \forall g\in\mathcal{G}, \ \forall t\in\mathcal{T}, \\
& f_g(t) = \prod_{m\in\mathcal{M}} [x_g(t,m)(1-p_g(t,m)) + 1 - x_g(t,m)] f_g(t-1), \ \forall g\in\mathcal{G}, \ \forall t\in\mathcal{T}, \\
& f_g(0) = 1, \ \forall g\in\mathcal{G}, \\
& b_g(t,m) = x_g(t,m) f_g(t-1)\Big[p_g(t,m)\big[r_g(t,m)b^s(m) + \big(1-r_g(t,m)\big)b^{f_r}(m)\big] + \\
& \qquad\qquad \big(1-p_g(t,m)\big)b^{f_c}(m)\Big], \ \forall g\in\mathcal{G}, \ \forall t\in\mathcal{T}, \ \forall m\in\mathcal{M}, \\
& c_g(t,m) = x_g(t,m) f_g(t-1)\Big[p_g(t,m)\big[r_g(t,m)c^s(m) + \big(1-r_g(t,m)\big)c^{f_r}(m)\big] + \\
& \qquad\qquad \big(1-p_g(t,m)\big)c^{f_c}(m)\Big], \ \forall g\in\mathcal{G}, \ \forall t\in\mathcal{T}, \ \forall m\in\mathcal{M}, \\
& x_g(t,m) \in \{0,1\}, \ \forall g\in\mathcal{G}, \ \forall t\in\mathcal{T}, \ \forall m\in\mathcal{M}.
\end{aligned}
\tag{5}
$$

Note that in this formulation, we have chosen to model the budgetary constraint and the capacity restriction as a global constraint over all the groups. However, it is quite easy to divide the budget B into budgets B_g for each group g, and then have a constraint per group. A similar remark holds for the capacity restriction as well.

In order to incorporate the budgetary constraint and the capacity restriction in the Markov decision problem, we need to add state variables b and c for both the budget and the capacity, respectively. In each state $s = (t, \vec{f}, K, b, c)$, these variables denote the budget and the capacity that are left for the rest of the survey. At time t, the budget and the capacity after taking a decision $a_g(m)$ are decreased by $\sum_{g\in\mathcal{G}}\sum_{m\in\mathcal{M}} a_g(m)b_g(t,m)$ and $\sum_{g\in\mathcal{G}}\sum_{m\in\mathcal{M}} a_g(m)c_g(t,m)$, respectively. This can only be done as long as the budget and the capacity remain non-negative. This requirement is added to the action set. Hence, the dynamic programming backward recursion equations become

$$V(s) = \max_{a\in\mathcal{A}_s} \Big[\sum_{g\in\mathcal{G}} \sum_{m\in\mathcal{M}} w_g a_g(m) f_g p_g(t,m) r_g(t,m) + V\Big(t+1, \big(\prod_{m\in\mathcal{M}} \big[1 - a_g(m)p_g(t,m)\big] f_g\big)_{g\in\mathcal{G}}, \big(k_g(m) - a_g(m)\big)_{g\in\mathcal{G}, m\in\mathcal{M}}, b - \sum_{g\in\mathcal{G}} \sum_{m\in\mathcal{M}} a_g(m) b'_g(t,m), c - \sum_{g\in\mathcal{G}} \sum_{m\in\mathcal{M}} a_g(m) c'_g(t,m)\Big)\Big], \tag{6}$$

with $\mathcal{A}_s = \{a_g(m) \in \{0,1\}^{G\cdot M} \mid a_g(m) \leq k_g(m), \sum_{m\in\mathcal{M}} a_g(m) \leq 1, b - \sum_{g\in\mathcal{G}} \sum_{m\in\mathcal{M}} a_g(m) b'_g(t,m) \geq 0,$ and $c - \sum_{g\in\mathcal{G}} \sum_{m\in\mathcal{M}} a_g(m) c'_g(t,m) \geq 0\}$, in which we defined $b'_g(t,m)$ and $c'_g(t,m)$ to be $b'_g(t,m) = f_g p_g(t,m)\big[r_g(t,m) b^s(m) + \big(1 - r_g(t,m)\big) b^{f_r}(m)\big] + \big(1 - p_g(t,m)\big) b^{f_c}(m)$ and $c'_g(t,m) = f_g p_g(t,m)\big[r_g(t,m) c^s(m) + \big(1 - r_g(t,m)\big) c^{f_r}(m)\big] + \big(1 - p_g(t,m)\big) c^{f_c}(m)$.

5 Numerical examples

The previous sections dealt with the theoretical models to solve the problem of resource allocation within adaptive survey designs. In this section, we give two numerical examples to illustrate our methodology.

Our first example shows that the solution of the basic unconstrained model is indeed optimal, although, counterintuitive. Consider a survey sample in which all units belong to the same group g. The set of available interview modes is $\mathcal{M} = \{\text{face-to-face, phone}\}$. The survey fieldwork is divided in $T = 6$ time slots. Table 1 gives the contact and cooperation probabilities $p_g(t,m)$ and $r_g(t,m)$ as estimated from previous such surveys and the maximum number of attempts $\bar{k}_g(m)$.

Note that there is a clear preference for contact at time slots t_3 and t_6 for both interview modes. For response, on the other hand, the probabilities indicate more than 50% chance for positive response except for an attempt by face-to-face at t_3 and by phone at t_5. Therefore, it is not obvious what time slots should be chosen in order to maximize the total reward. Using the algorithm from Section 3, we obtain the solution depicted in Table 2.

Let us analyze this solution. It looks surprising that for the first time slot

Table 1: Input data for group g

Mode	Probability	t_1	t_2	t_3	t_4	t_5	t_6	$\bar{k}_g(m)$
Face-to-face	$p_g(t,m)$	0.3	0.4	0.8	0.2	0.3	0.7	2
	$r_g(t,m)$	0.9	0.7	0.3	0.8	0.8	0.6	
Phone	$p_g(t,m)$	0.4	0.5	0.9	0.4	0.4	0.8	4
	$r_g(t,m)$	0.8	0.5	0.7	0.6	0.4	0.6	

Table 2: Optimal solution – original setting

Time slot	t_1	t_2	t_3	t_4	t_5	t_6	Response rate
Mode	F2F	F2F	Ph	Ph	0	Ph	0.753

Table 3: Optimal solution – different cooperation probability at t_1

Time slot	t_1	t_2	t_3	t_4	t_5	t_6	Response rate
Mode	Ph	F2F	Ph	Ph	F2F	Ph	0.736

mode F2F is chosen and not Ph, although the immediate reward is higher for Ph. Two observations are necessary to explain this behavior. First, the cooperation probability $r_g(t_1, \text{F2F})$ is higher than $r_g(t_1, \text{Ph})$. The situation changes when $r_g(t_1, \text{F2F}) < r_g(t_1, \text{Ph})$. For example, take $r_g(t_1, \text{F2F}) = 0.7$. Then, the new optimal solution (see Table 3) uses *phone* as first approach interview mode. Second, given the formula in (1), a lower contact probability in the first time slot leads to a higher future reward.

The structure of the solution given in Table 2 is motivated by the choice of $\bar{k}_g(m)$. From t_3 onward $k_g(\text{F2F}) = 0$, therefore *phone* is the only interview mode left available. Thus, the choice for time slots t_3, t_4, and t_6 is logical. However, taking action 0 at t_5 again looks counterintuitive. Since there are enough attempts left for mode Ph and there are no budget or capacity constraints, it feels natural to choose for an attempt to approach. However, the response obtained at time slot t_6 is higher than the response obtained if approached at both t_5 and t_6. Therefore, action 0 is preferred.

Table 4: Optimal solution – more attempts available

Time slot	t_1	t_2	t_3	t_4	t_5	t_6	Response rate
Mode	F2F	F2F	Ph	Ph	F2F	Ph	0.755

Table 5: Input data for group g_2

Mode	Probability	t_1	t_2	t_3	t_4	t_5	t_6	$k_{g_2}(m)$
Face-to-face	$p_{g_2}(t,m)$	0.8	0.6	0.4	0.6	0.4	0.2	1
	$r_{g_2}(t,m)$	0.9	0.7	0.6	0.8	0.5	0.3	
Phone	$p_{g_2}(t,m)$	0.7	0.6	0.5	0.6	0.5	0.4	2
	$r_{g_2}(t,m)$	0.8	0.6	0.5	0.6	0.4	0.2	

The optimal solution in Table 2 does not employ all attempts available for mode Ph. Therefore, we cannot obtain a different solution if we increase the number of attempts for this mode. On the other hand, if we increase the number of attempts to 3 for F2F, then the response rate improves (see Table 4). The structure of the optimal solution does not change much from the original setting. The only difference appears at t_5 where this time there are enough attempts for mode F2F, and selecting this mode leads to a higher reward.

Our second example depicts the optimization mechanism for two groups. Consider again the setting from the previous example, where Table 1 has the input data for group g_1. Table 5 below gives the corresponding input data for group g_2.

Approaching group g_2 for the survey follows a more intuitive behavior, e.g., high cooperation probabilities correspond to high contact probabilities. It is no surprise that the optimal solution (see Table 6) starts with the choice of t_1 and Face-to-face as interview mode, since this results in the highest immediate reward. The same argument governs the entire structure of the solution.

If the proportion of the two groups in the survey sample is $w = (0.5, 0.5)$, then the overall response rate given by (2) is 0.787. Evidently, the more weight the second group has in the sample, the higher the total response rate. Analogously, the higher the weight for g_1, the lower the overall response rate. For

Table 6: Optimal solution for group g_2

Time slot	t_1	t_2	t_3	t_4	t_5	t_6	Response rate
Mode	F2F	Ph	0	Ph	0	0	0.821

Table 7: Optimal solution for group g_2, $k_{g_2}(\text{F2F}) = 3$

Time slot	t_1	t_2	t_3	t_4	t_5	t_6	Response rate
Mode	F2F	F2F	0	F2F	Ph	Ph	0.85

Table 8: Optimal solution for group g_2, $k_{g_2}(\text{F2F}) = 4$

Time slot	t_1	t_2	t_3	t_4	t_5	t_6	Response rate
Mode	F2F	F2F	F2F	F2F	Ph	Ph	0.851

example, if $w = (0.8, 0.2)$ then the total response rate is 0.767.

Let us take another look at group g_2. If we increase $k_{g_2}(m)$ for Face-to-face to 2, we obtain a group average response of 0.839. This is an improvement of approximately 2.2%. A further increase to 3 improves the average response by 3.6% from the original setting and by only 1.4% from the previous situation (Table 7). From $k_{g_2}(m) = 4$ on the response rate does not improve anymore (Table 8).

The decrease in the response rate gain from 2.2% to 1.4% by allowing for more attempts is due to the general decreasing trend in the input probabilities. Such sensitivity analysis is essential when deciding upon a good value for $\bar{k}_g(m)$. We can conclude that even for an unconstrained problem (with respect to budget or available capacity) admitting larger values for $\bar{k}_g(m)$ does not necessarily yield improved values of the objective function.

6 Conclusions

For every survey that is planned, survey organizations are confronted with the decision over the necessary budget such that the resulting survey quality is above a pre-agreed level. In most cases, the budget proves insufficient due to an increased effort to convince sample units to respond. Temporary solutions can be found in replacing the expensive design with a cheaper one. That, however, has a negative influence on the survey quality. Moreover, the decreasing response prevents expensive designs from performing as well as it is expected. Therefore, a new perspective needs to be taken. Learning the behavior patterns, i.e., the survey features that influence the willingness to participate into surveys, for respondents and nonrespondents, can help with a more useful assignment of resources to surveys. Striving for an optimal resource allocation can help with reducing the budget overruns.

Our research aims at optimizing the resource allocation for an adaptive survey design. The main components of such a survey design are interview modes, number of time slots, number of allowed attempts, and the survey sample divided into homogeneous groups according to some given criteria (e.g., demographics). The current paper presents an optimization model that addresses this new adaptive survey design methodology.

We start by analyzing a simpler version of the problem, i.e., with no budget or capacity constraints. In this setting optimizing the resource allocation is translated to choosing a sequence of time slots such that the response rate is maximized given the contact and cooperation probabilities for each group, each time slot, and all available interview modes. The history of past actions that has to be considered at each step when choosing an action is a complex non-linear factor. Section 2 explains why the simplified model is non-scalable and non-tractable even for small problems.

Nevertheless, the optimal solution can be found. We present a method that addresses the non-linearity of the problem. The idea is to use a Markovian decision formulation of the problem, in which the state space is extended such that the contact failure probability is included in the state. Thus, there is no need to store the entire configuration of past actions. Via dynamic programming the new formulation is solved to optimality. We have tested the performance of the method by analyzing various survey settings. Section 5 presents the numerical results, where the optimal solution is not entirely intuitive.

The main advantages of our method are guaranteed optimality and short computational times. Thus, the model can be successfully used as a basis for representation of more complex practical settings. For example, Section 4 deals with the necessary changes in the theoretical framework that adjust our method to accommodate cost and capacity constraints.

Statistics bureaus are directly interested in testing adaptive designs as an alternative to classical survey designs. Simulated results prove the efficiency of our current technique. A direct comparison between the two designs is not yet possible due to some practical aspects not fully covered by our method. For example, estimation of the input probabilities is not discussed here. However, a great deal of attention has to be paid to this phase since the optimization part builds upon this input. Issues such as time-dependency, history-dependency, and repeated shift between interview modes have to be taken into account when estimating the input probabilities. Moreover, implementation of a solution with multiple shifts between certain interview modes might be difficult in practice (e.g., face-to-face, phone), therefore the number of such switches should be constrained.

Flexibility in addressing various objective functions is another aspect of interest. As shown by (4), the method is applicable as long as the objective value has an additive property. However, recent literature on survey methodology argues that aiming for high response rates influences negatively the bias of the estimators (see, e.g., [12] and [9]). Other quality measures, such as low variation in the group response rates (i.e., representativeness of the respondent sample), have been indicated as more suitable. Moreover, the quality of the response should be investigated as well. Data quality measures, such as the measurement error, should be included in the model. Also, feedback on the reasons for nonresponse could help in choosing a more suitable strategy to approach sample units. However, such quality functions do not posses the additive property which makes it difficult to approach by the current method.

Future research aims at tackling these issues in order to develop a model that meets practical needs. Intuitively, taking two survey designs with similar settings, an adaptive design is expected to outperform the classical design since more information becomes available from historical data and the design is tailored such that the group response rate is optimized.

References

[1] D.A. Dillman. *Mail and Internet Surveys: the Tailored Design Method.* John Wiley & Sons, 2007.

[2] B. Greenberg and S. Stokes. Developing an optimal call scheduling strategy for a telephone survey. *Journal of Official Statistics*, 6:421–435, 1990.

[3] R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little. *Survey Nonresponse.* Wiley Series in Probability and Statistics, 2002.

[4] M.H. Hansen and W.N. Hurwitz. The problem of non-response in sample surveys. *Journal of American Statistical Association*, 41(236):517–529, 1946.

[5] S.G. Heeringa and R.M. Groves. Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):439–457, 2006.

[6] W.D. Kalsbeek, S.L. Botman, J.T. Massey, and P-W. Liu. Cost-efficiency and the number of allowable call attempts in the National Health Interview Survey. *Journal of Official Statistics*, 10(2):133–152, 1994.

[7] R. Kulka and M. Weeks. Toward the development of optimal calling protocols for telephone surveys: a conditional probabilities approach. *Journal of Official Statistics*, 4:319–332, 1988.

[8] L.E. Lyberg, P. Biemer, M. Collins, E.D. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin. *Survey Measurement and Process Quality.* Wiley Series in Probability and Statistics, 1997.

[9] C. Mohl and F. Laflamme. Research and responsive design options for survey data collection at Statistics Canada. In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 2007.

[10] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, 1994.

[11] B. Schouten, M. Calinescu, and A. Luiten. Optimizing quality of response through adaptive survey designs. Discussion paper 201118, Statistics Netherlands, The Hague, 2011.

[12] B. Schouten, F. Cobben, and J. Bethlehem. Indicators for the representativeness of survey response. *Survey Methodology*, 35(1), 2009.

[13] J.R. Wagner. *Adaptive Survey Design to Reduce Nonresponse Bias.* PhD thesis, University of Michigan, 2008.