O P E R A T I O N S R E S E A R C H A N D D E C I S I O N S
No. 4 2012
DOI: 10.5277/ord120404

Bogumił KAMIŃSKI*
Mariusz KOZAKIEWICZ*
Wit JAKUCZUN*
Małgorzata PÓŁTORAK**

AN OPTIMAL ASSIGNMENT PROCEDURE FOR MULTIPLE ONLINE SURVEYS

The problem of optimal assignment of respondents to internet surveys has been considered. The task is taken from a leading online research company in Central and Eastern Europe, which runs several dozen surveys in parallel. Each survey is assigned a target number of responses to be collected and unique selection criteria for choosing respondents based on their socio-demographic data. A mathematical programming model has been proposed that enables us to find an optimal mechanism for assigning respondents to surveys so as to minimize the required total number of invitations to surveys issued. A side effect of the assignment procedure is that the samples obtained are not representative of the population under survey. Therefore, a sample weighting scheme has been developed that takes this fact into account and allows us to obtain unbiased estimators of the characteristics of the population surveyed.

Keywords: *mathematical programming, optimal allocation problem, online survey management*

1. Introduction

The development of modern communication technologies has led to a growing acceptance of online surveys among scientists and market researchers [8]. There are two main approaches to doing online surveys in open populations: pre-recruited panels and intercept sampling [9]. The approach based on a panel of respondents is the most common due to its technical simplicity. A company maintains a group (panel) of re-

*Warsaw School of Economics, al. Niepodległości 162, 02-554 Warszawa, e-mail addresses: bkamins@sgh.waw.pl, kozakiewicz.mariusz@gmail.com, wit.jakuczun@gmail.com
**Gemius S.A., ul. Postępu 18B, 02-676 Warszawa, e-mail: malgorzata.poltorak@iibr.pl

spondents from which a sample is selected to whom surveys are sent. A more advanced and promising approach is intercept sampling. Using this approach, random respondents are asked to fill in a survey while visiting web pages and the company does not know them in advance.

In this paper, we focus on surveys using intercept sampling. This approach has several appealing characteristics: it is relatively inexpensive, fast and enables a more representative sample of respondents to be collected. Moreover, it allows researchers to reach niche target groups, for example those characterized by deviant or covert behavior [7]. However, it is also associated with numerous difficulties connected with its complex protocol. To understand them, let us start by describing a standard process for conducting an intercept sampling survey:

S1. *An invitation* to participate in the survey is presented to a random sample of respondents visiting a selected set of web pages*.

S2. If the invitation is accepted, *selection criteria* questions are asked; if not, the process is stopped.

S3. If the respondent's *attributes* fit the survey selection criteria, the core survey questions are presented to the respondent and her *response* is collected; if she does not meet the selection criteria, the process is stopped.

Each survey has a required number of responses to be collected. Steps S1–S3 are repeated until this value is reached. The company has to pay for each invitation to participate in a survey presented to a respondent, so it wants to *minimize the number of invitations issued*. Notice, however, that the company has no control over whether a respondent accepts the invitation, so it treats the fact of acceptance as a random event. Therefore its goal can be equivalently stated as *minimization of the number of accepted invitations*. The latter objective will be used in the paper as it leads to a simpler mathematical formulation.

Typically, the company runs several surveys in parallel. The simplest approach to their execution is to run each one separately. In such a case it is relatively easy to control the sample size and ensure an appropriate sampling procedure. Notice that this approach treats respondents in an ineffective manner. If a respondent accepts the invitation but does not meet the selection criteria of the survey, she is assigned to, the invitation is lost. However, it is possible that this respondent meets the selection criteria of some other survey run in parallel by the company. Therefore the company should assign the respondent to a survey after checking which surveys she could respond to. If we take into account that there may be tens of surveys run in parallel, we can expect that this could lead to substantial cost savings, even if we assume that each respondent is assigned to only one survey after accepting an invitation. It would be

*In the case of the company studied, the selected web pages cover around 90% of Polish Internet users. Due to this the company claims that results may be representative for this population.

tempting to assign a respondent to all the surveys that fit her attributes but the company does not apply such a policy, because this would overload the respondent with a large number of core survey questions asked in one session.

The idea of cost reduction by assigning a respondent to a survey after learning her attributes is very simple. An obvious question is why it was not done before by the company. The answer is that there are two serious obstacles to this approach connected with: (1) the optimal assignment of respondents to surveys and (2) representativeness of the samples collected.

Firstly – if a respondent meets the selection criteria for several surveys, it is not obvious to which one of them she should be assigned in order to minimize the total number of issued invitations needed to collect the required number of responses for all the surveys.

The second issue is related to the quality of the samples collected. Because we assign respondents to surveys based on their attributes, the samples collected may not be representative. To see this, consider two surveys run in parallel. In both of them it is required to collect 500 responses. The first survey is targeted only at females and the second survey accepts all respondents. Assume that we have a representative sample of 1000 respondents consisting of 500 males and 500 females that accepted the invitation to participate in the survey. Clearly using this sample we can meet the required number of responses for both surveys. The first survey is not problematic – we assign 500 females to it and this is a representative sample of the female population under study. However, in the second survey we have 500 males. This is clearly not a representative sample of a population of both men and women.

In the article, we show how the assignment and sampling problems can be solved. We develop an online stochastic algorithm that iteratively generates optimal assignments of respondents to surveys. We show that the problem can be reduced to the task of solving a linear programming model combined with the random assignment of respondents to surveys. Next we propose a weighting procedure that uses the samples obtained from the assignment algorithm and allows the company to calculate unbiased estimators of responses in a population meeting the survey selection criteria.

To the knowledge of the authors, formulation of the problem presented is a novel one. Standard literature analyzing online surveys concentrates on *single survey problems* and discusses such issues as: questionnaire design, how to sample respondents and the process of survey administration [7–9]. More focused literature discussing problems of the optimal collection of survey data analyzes the problem of adaptive survey design in order to minimize the cost and maximize the quality of the sample [1, 3]. However, in this stream only a *single survey* is considered. A review of papers published in the journal Survey Methodology, a leading journal in this research field, since 2005 has shown no results regarding the problem of analyzing multiple surveys conducted online in parallel. A problem related to the multiple-survey setting that can be found in the literature is an analysis of a single survey using multiple frames, i.e.

several different sampling strategies are applied to one survey [5]. Such lack of analysis in the literature of multiple online surveys using intercept sampling can be explained by the relatively novel nature of this type of surveying method. For example, in Poland only one company focuses its business model on online surveys using intercept sampling and most international research agencies do not offer such surveys at all, concentrating mainly on the panel approach.

The article is divided into three parts. In Section 2, we present theoretical results concerning the optimal assignment of respondents to multiple surveys. The procedure is applied to an example scenario. Unfortunately, the optimal assignment algorithm generates non-representative samples, so next we develop an appropriate response weighing procedure to account for this. Section 3 contains an example application of the proposed procedure using real data that we obtained from the company. The last part of the article contains concluding remarks discussing the practical aspects of applying the presented methodology, its limitations and possible extensions.

The research presented in this article was financed by Gemius S.A. and a grant from Narodowe Centrum Badań i Rozwoju for the project *Innowacyjne techniki i narzędzia badań online, minimalizujące obciążenie użytkowników*. All the data regarding surveys were obtained from Interaktywny Instytut Badań Rynkowych sp. z o.o. being a leading online market research company in Central and Eastern Europe.

2. Model for optimal assignment of respondents to surveys

In this section, we introduce the formal notation used in the paper and present its application to an example scenario. Let us start with a description of the example scenario. We have to specify the respondent population and surveys that are run in parallel. Each respondent is characterized by 2 attributes: employment status and place of residence. Employment can take three levels: unemployed (*UN*), employed in the public sector (*PU*) and employed in the private sector (*PR*). Place of residence can be: village (*V*), small town (*S*) and large town (*L*). We assume that the company knows the joint distribution of these attributes in the general population. This is presented in Table 1.

We assume that the company runs 3 surveys. The first survey needs to be responded to 1000 times, the second 2500 times and the third 500 times. The first survey is targeted at people unemployed or employed in the public sector and living in a village or small town. The second survey requires people employed in the public or private sector and resident in a small or large town. Finally, the third survey is targeted at inhabitants of small towns without restrictions on employment status. The relation

between the attributes of respondents and the selection criteria for these surveys is presented in Table 2.

Table 1. Joint distribution of employment status and place of residence in the example scenario

Level of employment	*V*	*S*	*L*
UN	0.04	0.03	0.03
PU	0.06	0.10	0.14
PR	0.10	0.17	0.33

Source: Authors' work.

Table 2. Selection criteria for the surveys in the example scenario

Level of employment	*V*	*S*	L
UN	1	1, 3	
PU	1	1, 2, 3	2
PR		2, 3	2

The numbers indicate which surveys are appropriate to the given attributes of respondents.

Source: Authors' work.

Now let us move to the formal statement of the multi-survey assignment problem. Consider a company conducting online surveys. Assume that it has orders to run K surveys in parallel and denote $\Gamma = \{1, ..., K\}$. Survey $k \in \Gamma$ has to be responded to r_k times. The company wants to minimize the total number of accepted invitations to surveys. This number is denoted by R. The constraints on the assignment of respondents to surveys are: (1) not every respondent meets the selection criteria of every survey and (2) one invitation can lead to at most one survey response.

Each survey has its own distinct selection criteria. We assume that each respondent is described by T attributes taking values from the finite* sets $A_1, A_2, ..., A_T$. Denote the set of all possible respondent attributes by $A = \times_{i=1}^{T} A_i$. The inclusion criteria require that the attributes of respondents to survey $k \in \Gamma$ belong to a given subset S_k of A. We will say that a respondent with a vector of attributes $a \in A$ meets the selection criteria for survey k if and only if $a \in S_k$.

*We assume – which is standard practice in survey research – that continuous attributes are discretized before the analysis is performed.

In this research, we assume that the company knows the joint distribution of respondent attributes in the population*. Therefore, if we take a subset $Q \subset \Gamma$ of surveys, the company can calculate the probability that a random respondent meets the selection criteria for surveys in the set Q and does not meet the selection criteria for surveys in the set Γ–Q. The set $S(Q) \subset A$ of respondent attributes meeting this criterion is defined as:

$$S(Q) = \left(\bigcap_{k \in Q} S_k\right) \cap \left(\bigcap_{l \in \Gamma - Q} \left(A - S_l\right)\right) \quad \text{if} \quad 0 < |Q| < |\Gamma| \tag{1}$$

In addition, we define $S(\Gamma) = \bigcap_{k \in \Gamma} S_k$ and $S(\varnothing) = \bigcap_{l \in \Gamma} \left(A - S_l\right)$.

The probability $p(Q)$ that a random respondent $a \in A$ meets the selection criteria for surveys in the set Q and does not meet them for questionnaires in the set Γ–Q can be written as:

$$p(Q) = Pr\left(a \in S(Q)\right) \tag{2}$$

Note that for two subsets of Γ such that $Q_1 \neq Q_2$, the events $a \in S(Q_1)$ and $a \in S(Q_2)$ are mutually exclusive. Moreover $\bigcup_{Q \subset \Gamma} S(Q) = A$, so:

$$\sum_{Q \subset \Gamma} p(Q) = 1 \tag{3}$$

Henceforth, we will call Q the *type of a respondent* because the company can show surveys from the set Q to her and will not show any other surveys. By $I(Q, \mathrm{k})$ we will denote the indicator (characteristic) function of the set Q taking value 1 if $k \in Q$ and 0 otherwise.

Let us now present how this notation is implemented in the example scenario described at the beginning of this section.

We have $K = 3$ surveys, so $\Gamma = \{1, 2, 3\}$ and $r_1 = 1000$, $r_2 = 2000$, and $r_3 = 500$. We have defined $T = 2$ attributes, so $A_1 = \{UN, PU, PR\}$, $A_2 = \{V, S, L\}$. Therefore, we have $A = \{(UN, V), (PU, V)\ (PR, V), (UN, S), (PU, S), (PR, S), (UN, L), (PU, L), (PR, \mathrm{L})\}$. Finally, the inclusion criteria for the surveys given in Table 2 are translated as:

- $S_1 = \{(UN, V), (PU, V), (UN, S), (PU, S)\}$,
- $S_2 = \{(PU, S), (PR, S), (PU, L), (PR, L)\}$,
- $S_3 = \{(UN, S), (PU, S), (PR, S)\}$.

*In practice, this distribution is estimated using past data owned by the company and is additionally updated during the data collection process.

We see that we have eight possible types of respondents (Q): ∅, {1}, {2}, {3}, {1, 2}, {1, 3}, {2, 3}, {1, 2, 3}. The derivation of S(Q) and $p(Q)$ for these types of respondents is performed using the data from Tables 1 and 2. Its results are presented in Table 3.

Table 3. Derivation of the selection criteria $S(Q)$ and selection probability $p(Q)$ in the example scenario

Q	$S(Q)$	$p(Q)$
∅	{(*UN*, *L*), (*PR*, *V*)}	0.13
{1}	{(*UN*, *V*), (*PU*, *V*)}	0.10
{2}	{(*PU*, *L*), (*PR*, *L*)}	0.47
{3}	∅	0.00
{1, 2}	∅	0.00
{1, 3}	{(*UN*, *S*)}	0.03
{2, 3}	{(*PR*, *S*)}	0.17
{1, 2, 3}	{(*PU*, *S*)}	0.10

Source: Authors' own work.

In the next subsection we formulate the optimization problem and solve it for the example scenario.

Optimal assignment to surveys

The company has to decide which surveys it will show to which respondents, remembering that one respondent can be shown at most one survey. Denote by $f_{Q,k}$ the decision variable that is interpreted as the *unconditional* probability that the next individual invited to be surveyed is a respondent of type Q who will be shown survey k. In order for the decision of the company to be admissible, $f_{Q,k}$ may be greater than 0 only if survey k can be shown to a respondent of type Q, i.e. when $I(Q, k) = 1$. Additionally, the sum of all unconditional probabilities for a given respondent type cannot exceed $p(Q)$. This restriction stems from the fact that the unconditional probability that a randomly chosen individual is a respondent of type Q is equal to $p(Q)$. Formally, the described conditions can be written as follows:

$$\forall Q \subset \Gamma : \left(\forall k \in \Gamma : 0 \le f_{Q,k} \le I(Q,k) \right) \wedge \left(\sum_{k \in \Gamma} f_{Q,k} \le p(Q) \right) \tag{4}$$

Using the above notation, the unconditional probability that a randomly selected respondent will be shown survey $k \in \Gamma$ is equal to

$$g_k = \sum_{Q \subset \Gamma} f_{Q,k}$$

The company issues R accepted invitations in total and has to get at least r_k responses for survey k. Therefore, using expected values we obtain the condition $Rg_k \geq r_k$, which ensures that enough responses to survey k will be collected.

Summing up the above arguments, we can write the decision problem of the company as the following mathematical programming task:

$$R \to \min \text{ subject to: } R\sum_{Q \subset \Gamma} f_{Q,k} \geq r_k, \quad 0 \leq f_{Q,k} \leq I(Q,k), \quad \sum_{k \in \Gamma} f_{Q,k} \leq p(Q) \tag{5}$$

The formulation given by (5) is non-linear, because we multiply R by $f_{Q,k}$ in the first condition. Notice however that, using the substitution $Z = 1/R$, it can be reformulated as the following linear programming problem:

$$Z \to \max \text{ subject to: } Zr_k - \sum_{Q \subset \Gamma} f_{Q,k} \leq 0, \quad 0 \leq f_{Q,k} \leq I(Q,k), \quad \sum_{k \in \Gamma} f_{Q,k} \leq p(Q) \tag{6}$$

Solutions obtained from the above model are given by values of the decision variables $f_{Q,k}$ giving the unconditional probabilities of assigning respondents of different types (Q) to surveys (k).

Let us now present how problem (6) is formulated for the example scenario presented in the previous subsection. The conditions $Zr_k - \sum_{Q \subset \Gamma} f_{Q,k} \leq 0$ give inequalities:

$$\begin{cases} 1000Z - \left(f_{\varnothing,1} + f_{\{1\},1} + f_{\{2\},1} + f_{\{3\},1} + f_{\{1,2\},1} + f_{\{1,3\},1} + f_{\{2,3\},1} + f_{\{1,2,3\},1} \right) \leq 0 \\ 2500Z - \left(f_{\varnothing,2} + f_{\{1\},2} + f_{\{2\},2} + f_{\{3\},2} + f_{\{1,2\},2} + f_{\{1,3\},2} + f_{\{2,3\},2} + f_{\{1,2,3\},2} \right) \leq 0 \\ 500Z - \left(f_{\varnothing,3} + f_{\{1\},3} + f_{\{2\},3} + f_{\{3\}3} + f_{\{1,2\},3} + f_{\{1,3\},3} + f_{\{2,3\},3} + f_{\{1,2,3\},3} \right) \leq 0 \end{cases}$$

The conditions $0 \leq f_{Q,k} \leq I(q, k)$ correspond to 32 inequalities/equations:

$$\begin{cases} f_{\varnothing,1} = f_{\{2\},1} = f_{\{3\},1} = f_{\{2,3\},1} = 0 \\ 0 \leq f_{\{1\},1} \leq 1; \ 0 \leq f_{\{1,2\},1} \leq 1; \ 0 \leq f_{\{1,3\},1} \leq 1; \ 0 \leq f_{\{1,2,3\},1} \leq 1 \\ f_{\varnothing,2} = f_{\{1\},2} = f_{\{3\},2} = f_{\{1,3\},2} = 0 \\ 0 \leq f_{\{2\},2} \leq 1; \ 0 \leq f_{\{1,2\},2} \leq 1; \ 0 \leq f_{\{2,3\},2} \leq 1; \ 0 \leq f_{\{1,2,3\},2} \leq 1 \\ f_{\varnothing,3} = f_{\{1\},3} = f_{\{2\},3} = f_{\{1,2\},3} = 0 \\ 0 \leq f_{\{3\},3} \leq 1; \ 0 \leq f_{\{2,3\},3} \leq 1; \ 0 \leq f_{\{1,3\},3} \leq 1; \ 0 \leq f_{\{1,2,3\},3} \leq 1 \end{cases}$$

The conditions $\sum_{k\in\Gamma} f_{Q,k} \le p(Q)$ are expressed as 8 inequalities (the $p(Q)$ values are taken from Table 3):

$$\begin{cases} f_{\varnothing,1} + f_{\varnothing,2} + f_{\varnothing,3} \le 0.13 & f_{\{1\},1} + f_{\{1\},2} + f_{\{1\},3} \le 0.10 \\ f_{\{2\},1} + f_{\{2\},2} + f_{\{2\},3} \le 0.47 & f_{\{3\},1} + f_{\{3\},2} + f_{\{3\},3} \le 0.00 \\ f_{\{1,2\},1} + f_{\{1,2\},2} + f_{\{1,2\},3} \le 0.00 & f_{\{1,3\},1} + f_{\{1,3\},2} + f_{\{1,3\},3} \le 0.03 \\ f_{\{2,3\},1} + f_{\{2,3\},2} + f_{\{2,3\},3} \le 0.17 & f_{\{1,2,3\},1} + f_{\{1,2,3\},2} + f_{\{1,2,3\},3} \le 0.10 \end{cases}$$

The solution to this problem is given in Table 4 and yields the following assignment rules:

- assign all respondents of types {1} and {1, 2, 3} and 58,33% of the respondents of type {1, 3} to survey 1,
- assign all respondents of type {2} and 43.38% of the respondents of type {2, 3} to survey 2;
- assign 41,67% of the respondents of type {1, 3} and 56,62% of the respondents of type {2, 3} to survey 3.

Table 4. Values of $f_{Q,k}$ in the optimal solution of problem (6) in the example scenario

k/Q	∅	{1}	{2}	{3}	{1, 2}	{1, 3}	{2, 3}	{1, 2, 3}
1	0	0.1	0	0	0	0.0175	0	0.1
2	0	0	0.47	0	0	0	0.07375	0
3	0	0	0	0	0	0.0125	0.09625	0

Source: Authors' own work.

Using such a scheme we can expect to issue $R \approx 4598$ accepted invitations, which is more than the theoretically possible minimum of 4000 = 1000 + 2500 + 500 accepted invitations because 13% of respondents have type ∅ and cannot be assigned to any questionnaire. On the other hand, if the company did not use the optimization procedure and ran the surveys independently, we can calculate using the data from Tables 1 and 2 that it would require ca. 8717 accepted invitations to meet the required number of collected responses, so the gain from the optimization procedure is substantial.

Sample weighting procedure

Let us move on to the analysis of the representativeness of the samples created using the assignment procedure described. First notice that under representative sam-

pling from the surveyed population the proportion of respondents to survey k who are of type Q will be equal to:

$$Pr_{\text{rep}}\left(Q\middle|k\right)=\frac{I\left(Q,k\right)p\left(Q\right)}{\sum\limits_{B\subset\Gamma}I\left(B,k\right)p\left(B\right)} \tag{7}$$

The above condition states that survey k is carried out when $k \in Q$ and in such a case a respondent is of type Q with probability $p(Q)$ divided by the probability that any respondent meets the inclusion criteria for survey k.

However, if we apply the assignment weights resulting from the procedure described by equation (6), the probability that a respondent is of type Q given survey k is carried out is equal to:

$$Pr_{\text{opt}}\left(Q\middle|k\right)=\frac{f_{Q,k}}{\sum\limits_{B\subset\Gamma}f_{B,k}} \tag{8}$$

Therefore, after collecting the sample for survey k, we have to weight all the observations from respondents of type Q by the ratio $Pr_{\text{rep}}\left(Q\middle|k\right)/Pr_{\text{opt}}\left(Q\middle|k\right)$ in order to obtain unbiased estimators of the characteristics of the surveyed population.

Notice that we must ensure that $Pr_{\text{rep}}\left(Q\middle|k\right) > 0 \Rightarrow f_{Q,k} > 0$. If this condition is not met, we encounter division by 0. The economic interpretation of this case is the following: if it is possible that in representative sampling a respondent of type Q is assigned to survey k, then it also must be possible under optimal sampling. However, the formulation of the optimization problem (6) does not ensure that this condition is met in general.

In the example scenario presented in the previous section we encounter this problem. For example, all respondents of type $\{1, 2, 3\}$ were assigned to survey 1. Therefore, we have $Pr_{\text{opt}}\left(\{1, 2, 3\}\middle|2\right) = 0$ but $Pr_{\text{rep}}\left(\{1, 2, 3\}\middle|2\right) \approx 13.51\%$ so it would be impossible to apply the weighting procedure for survey 2.

In such cases – in order to ensure that the sample can be reweighted – we have to add some minimal threshold value $\alpha > 0$ for $f_{Q,k}$ in situations where $Pr_{\text{rep}}\left(Q\middle|k\right) > 0$. Therefore, the optimization problem (6) should be augmented into the following form:

$$\begin{gathered} Z \to \max \text{ subject to: } Zr_k - \sum_{Q\subset\Gamma} f_{Q,k} \le 0, \quad I\left(\left]0;1\right]\right), \quad Pr_{\text{rep}}\left(Q\middle|k\right)\right) \\ \alpha \le f_{Q,k} \le I\left(Q,k\right), \quad \sum_{k\in\Gamma} f_{Q,k} \le p\left(Q\right) \end{gathered} \tag{9}$$

Notice, however, that α must not be too large. Denote the number of elements of Q by $|Q|$. Then we can see that α must be less than or equal to $\min_{Q\subset\Gamma;\, Q\neq\varnothing} p(Q)/|Q|$, otherwise the constraints $I\left(]0;1],\, Pr_{\text{rep}}\left(Q|k\right)\right)\alpha \le f_{Q,k}$ and $\sum_{k\in\Gamma} f_{Q,k} \le p(Q)$ cannot be jointly met.

In the example scenario if we take $\alpha = 0.01$, then we obtain the optimal solution presented in Table 5. In particular, observe that $f_{\{1,2\},1}$, $f_{\{1,2\},2}$ and $f_{\{3\},3}$ are 0 because p({1, 2}) and $p(\{3\})$ are 0. For this solution we obtain $R = 5000$. Note that the required number of accepted invitations increases due to the additional constraints.

Table 5. Values of $f_{Q,k}$ for the optimal solution of problem (9) in the example scenario with $\alpha = 0.01$

k / Q	∅	{1}	{2}	{3}	{1, 2}	{1, 3}	{2, 3}	{1, 2, 3}
1	0	0.1	0	0	0	0.02	0	0.08
2	0	0	0.47	0	0	0	0.02	0.01
3	0	0	0	0	0	0.01	0.08	0.01

Source: Authors' own work.

Finally let us note that if we ensure that $f_{Q,k} > 0$ when $Pr_{\text{rep}}\left(Q|k\right) > 0$, the calculation of unbiased estimators for the surveyed population is simple. For example, assume that in survey k we measure a continuous characteristic X and we want to estimate its expected value. Denote by $\hat{x}_{Q,k}$ the mean of the observed sample of X for customers of type Q who were assigned to survey k under optimal sampling following the results of procedure (9). Note that because we have introduced the condition $I\left(]0;1],\, Pr_{\text{rep}}\left(Q|k\right)\right)\alpha \le f_{Q,k}$, it is always possible to calculate $\hat{x}_{Q,k}$ if $k \in Q$ and $p(Q) > 0$. Thus the following is an unbiased estimator of mean of X in the surveyed population:

$$\sum_{Q:f_{Q,k}>0} \hat{x}_{Q,k} Pr_{\text{rep}}\left(Q|k\right) \tag{10}$$

In order to see this, note that in the optimized sample we assign to questionnaire k in total $R\sum_{B\subset\Gamma} f_{B,k}$ observations of which $Rf_{Q,k}$ observations of type Q. Therefore, given that some observation in the sample has type Q, its weight in the sum (10) equals $Pr_{\text{rep}}\left(Q|k\right)/\left(Rf_{Q,k}\right)$. This weight can be rewritten as $(Pr_{\text{rep}}(Q|k)/Pr_{\text{opt}}(Q|k))/(R\sum_{B\subset\Gamma} f_{B,k})$ and we see that the required weight $Pr_{\text{rep}}(Q|k)/Pr_{\text{opt}}(Q|k)$ is applied to this observation.

Adaptation of the model in practical applications

There are three practical problems with the application of the process described in the previous sections. Firstly, in the model (9) the condition

$$R\sum_{Q\subset\Gamma} f_{Q,k} \geq r_k$$

resulted from a calculation based on expected values and in practice we can expect some random deviation from them. Secondly, the company usually only has access to estimators of the probabilities $p(Q)$ used in the optimization process, so the calculations are only approximate. Thirdly – a respondent might accept an invitation to participate in the survey but later not finish filling in the questionnaire, i.e. drop out during the surveying process. Due to these reasons, in practical applications the exact solution of problem (9) is only an approximation to the truly optimal assignment. In order to cope with this limitation, we propose to update the optimal assignment schedule based on information gathered during the survey collection process. After an invitation is accepted by a respondent and her response is collected, the company can update the values of r_k. If the respondent has type $Q \neq \varnothing$, then she is assigned to one of the surveys $i \in Q$. If her response is successfully collected, then the value of r_i is decreased by 1. On the other hand, if she drops out and does not finish filling in the survey or her type is $\varnothing$, then no update to any r_k is performed. After this, the procedure has either finished (if all $r_k \leq 0$) or the problem (9) is solved again using the updated values of r_k and the procedure is repeated. The process described above is depicted in Fig. 1.

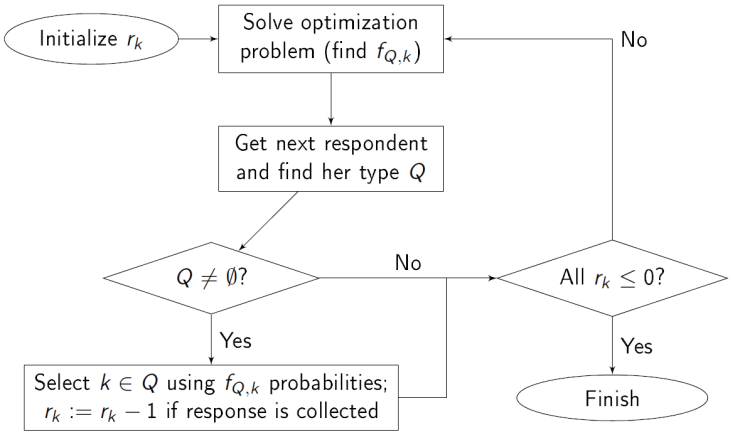

Fig. 1. Recommended procedure for using the algorithm for the optimal allocation of respondents to surveys

Notice that the proposed procedure can be effectively applied in near real time because problem (9) is linear and at each step a decrease in the value of r_k can only be slight, so an earlier optimal solution is admissible and can be used as a starting point in the next optimization step. This feature of the proposed solution is important for the company because the surveys are conducted online, which means that the speed of calculations is important. In this case, we have to remember that the probabilities $f_{Q,k}$ may change at each step of the procedure, so the probabilities given by formula (8) do not remain constant and appropriate weights for further estimation must be calculated for each observation separately.

In the next section a real life example of the application of the proposed procedure is presented.

3. Application of the model to real data

To validate our approach we conducted an experiment with the aid of real data. Using information gathered from real respondents was very important for our study, as the results are to be implemented and used on an everyday basis by the survey company. The goal of the experiment was to evaluate the maximum possible improvement between the current approach (respondents are invited to participate in surveys independently), called the *sequential approach*, and the procedure described in this paper, called the *optimization approach*. The results reported below are only selected summary statistics as the detailed survey parameters must remain confidential due to the survey company's policy and the regulations of the privacy law.

The population studied

The surveyed population in our study consists of respondents who visit a selected range of web pages that together are visited by around 90% of all internet users in Poland. Using this fact, the company assumes that sampling from this population may give a reasonably representative sample of the whole population of internet users in Poland. The target population for a particular survey is a subset of the whole population of Polish internet users.

Data collection procedure

To maximize the realism of our study we prepared a set of $K = 27$ surveys that were typical of a one month period of the company's activity. From those surveys we

extracted the *selection criteria* defining the target groups. They allowed us to define the sets S_k. Next, we prepared a survey including all those selection criteria. The survey was run and we collected a set of 1381 respondents. With this data, we were able to estimate the probabilities $p(Q)$. Using them we prepared the experimental environment to test the feasibility of our optimization approach.

Survey selection

In order to calculate how much of an improvement the optimization approach is over the sequential approach, we selected a subset of 8 surveys out of the 27 surveys collected. The surveys selected had the same constraints on the demographic traits age and sex and one other trait defining the target group, which was specific to each survey. This subset of 8 surveys was selected so as to ensure that their target groups were as pairwise exclusive as possible, in other words $\left|S_i \bigcap S_k\right|$ should be small for all pairs i, k. Such a set of constraints is typically met in practice, so it enables evaluation of potential improvement.

Simulating respondents

For all surveys the task was to collect $r_k = 1000$ responses. Therefore, in order to be able to complete these 8 surveys simultaneously, the set of 1381 responses was too small and had to be enlarged. We decided to resample the respondents we collected in the first step of the experiment using the bootstrap approach [2]. Technically, this means that we assumed that the empirical distribution of attributes for these 1381 respondents is their exact distribution in the general population. This gave us an approximation of the respondent population with the same statistical characteristics as the original set of respondents.

To alleviate the randomization effect of our results, we repeated our experiment 100 times and reported the mean of the results obtained.

Computational results

We performed two computational experiments: one for the sequential approach and the other for the optimization approach. Both experiments were conducted until we collected the mandatory number of respondents for each survey, which was equal to 1000 for each of them.

Table 6 presents the mean results of our computational study for 100 runs. The conclusion is that our optimization approach is around 2.3 times better in terms of the number of accepted invitations required to finish the surveying process.

Table 6. Number of responses required to collect a sample of 1000 responses for each of 8 surveys for both approaches averaged over 100 runs

Approach	Mean number of respondents
Sequential	29 800
Optimization	13 040

Source: Authors' own work.

Figures 2 and 3 present the representativeness problem of the sample for survey No. 3. The black bars show how respondents answered this question in the population surveyed. The grey bars present the answers to this question in the sample collected using the optimization approach. It is clear that the distributions are different. This implies that using observation weighting is crucial in the analysis of the data collected in order to obtain an unbiased estimator of the characteristics of the population surveyed.

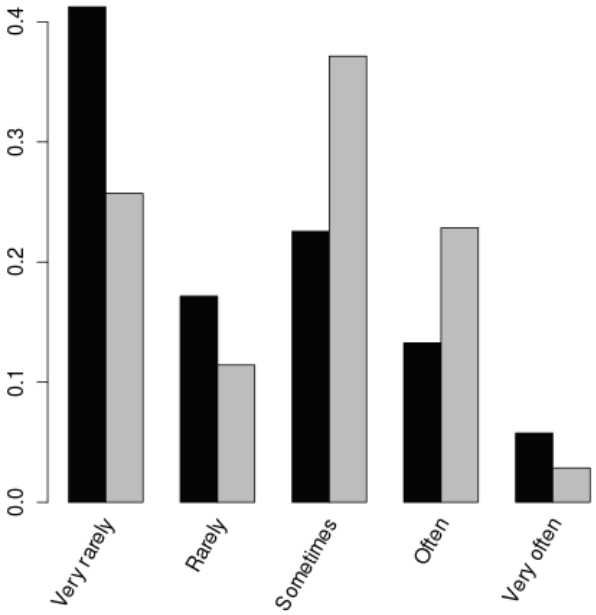

Fig. 2. Distribution of answers to the question *How often do you go to the cinema?* from respondents above 17 years of age. Results for survey No. 3.
Source: Authors' own work

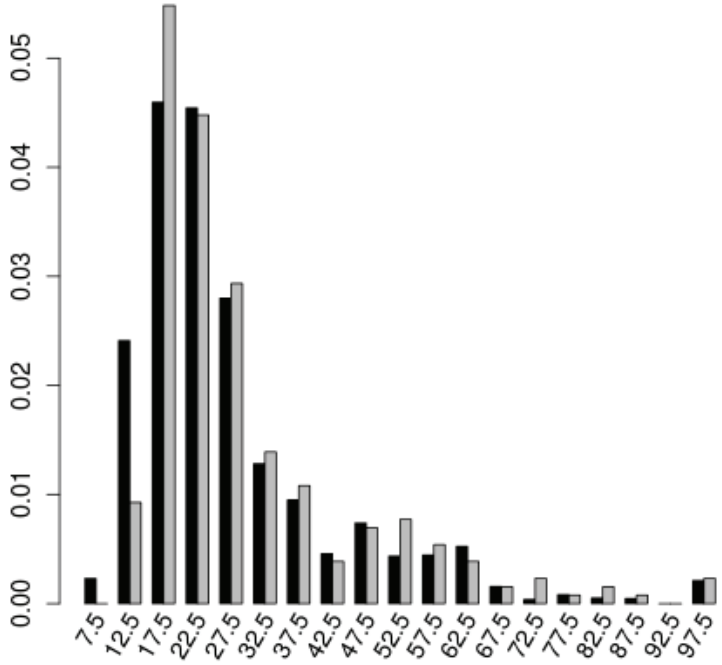

Fig. 3. Histogram of age. Results for survey No. 3.
Source: Authors' own work

Additionally, note that Table 6 indicates that the optimization approach is relatively close to the theoretical minimum number of respondents needed, which is equal to 8000. Unfortunately, this value cannot be achieved due to randomness in the selection process, the fact that some of the respondents do not meet the criteria for any of the surveys and that some surveys target very narrow subpopulations.

4. Concluding remarks

A method for the optimal assignment of respondents to multiple surveys has been presented together with a sample weighing procedure which corrects for the non-representativeness of the samples obtained. The experiment performed on real life data confirms the effectiveness of the procedure.

On the practical side, it should be stressed that the proposed algorithms are computationally efficient. The linear optimization problem that has to be solved is not very complicated and its solutions can be easily iteratively updated as new data are gathered. Therefore, it is possible to implement them in a system that assigns respondents

to surveys in real time. Additionally, the iterative nature of the solution safeguards the procedure from the case where the company's estimates of the joint distribution of demographic traits in the population surveyed are inaccurate. In such a case sampling weights can be automatically updated during the data collection process.

We can see the potential of improving our approach further by doing more research on the online nature of our problem. A promising direction is the online matching algorithm [4], which was successfully implemented for the problem of allocating internet advertisements [10], which is similar to the problem presented in this article.

As the final results of our research are to be implemented and used on a daily basis in practice, we additionally have to cope with numerous technical constraints, such as different sites where surveys can be displayed with different costs, channel capacity and expected rate of accepting invitations. A similar study was done for the optimal scheduling of mixed-mode surveys [1], where Markov decision theory and a dynamic programming approach were used [6]. Inclusion of heterogeneous survey costs and capacity limitations will make the optimization problem stated in formula (9) more complicated, but should not alter its general structure. The only major change will be in the objective function, as we will have to minimize the total cost of running all the surveys. However, after this change the optimization problem remains linear.

References

[1] CALINESCU M., BHULAI S., SCHOUTEN B., *Optimal scheduling of contact attempts in mixed mode surveys,* Statistics Netherlands, Hague, The Netherlands, 2011.

[2] EFRON B., *Bootstrap methods: another look at the jackknife*, The Annals of Statistics, 1979, 7 (1), 1–26.

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

[4] KARP R.M., VAZIRANI U.V., VAZIRANI V.V., 1990, *An optimal algorithm for on-line bipartatie matching*, STOC '90, Proceedings of the twenty-second annual ACM symposium on theory of computing, Baltimore, 13–17.05.1990, 352–358.

[5] LOHR S., RAO J.N.K., *Estimation in Multiple-Frame Surveys*, Journal of the American Statistical Association, 2006, 101, 1019–1030.

[6] PUTERMAN M.L., *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, Hoboken, New Jersey 1994.

[7] SELM VAN M., JANKOWSKI N.W., *Conducting Online Surveys*, Quality and Quantity, 2006, 40, 435–456.

[8] SINGH A., TANEJA A., MANGALARAJ G., *Creating Online Surveys: Some Wisdom from the Trenches Tutorial*, IEEE Transactions on Professional Communication, 2009, 52 (2), 197–212.

[9] SUE V.M., RITTER L.A., *Conducting Online Surveys*, SAGE Publications, Los Angeles 2006.

[10] VEE E., VASSILVITSKII S., SHANMUGASUNDARAM J., *Optimal online assignment with forecasts*, ACM, Cambridge, Massachusetts, USA, 2010.