Survey Research Methods (2015)
Vol. 9, No. 3, pp. 221-230
doi:10.18148/srm/2015.v9i3.6202

ISSN 1864-3361
http://www.surveymethods.org

Essay

Sunday shopping – The case of three surveys

Jelke Bethlehem
Institute of Political Sciences
Leiden University
The Netherlands

There is a growing discussion about the use of non-probability sampling in survey research. Probability sampling is the preferred method of sample selection, but practical problems like reduced data collection budgets, increasing nonresponse rates, and lack of adequate sampling frames force researchers to use different sampling methods. Particularly, online surveys based on self-selection of respondents have become very popular. Some say that use of such alternative sampling methods is not without risks as often proper inference from sample to population is not possible. Others say that non-probability sampling can produce satisfactory estimates provided effective correction techniques are applied. To obtain more insight in various sample selection methods, it would be nice to be able to compare them in practical situations. This paper describes a case in which three different surveys were carried out on the same topic, at the same time, and with the same questionnaire, but with different sample selection methods: an online panel based on probability sampling, an online survey based on self-selection, and a face-to-face survey in shopping centers. The results of these three polls differ substantially. This is a warning to be careful when choosing a sample selection method.

Keywords: Probability sampling, non-probability sampling, self-selection, web survey

1 Introduction

There are many surveys and polls in the Netherlands. Particularly during election campaigns, polls follow each other in rapid succession. Several poll organizations are active, and sometimes they conduct a new poll each day. Moreover, there are polls of polls that combine the results of a number of polls.

Statistical data collection is also carried out by large government organizations like Statistics Netherlands and the Netherlands Institute for Social Research. It is their core business. Their polls are usually large and complex. Therefore, they are called surveys.

Local authorities in the Netherlands also conduct surveys. A well-known example is the so-called *omnibus survey*, which is repeated regularly in many municipalities. As the name indicates, these surveys ask questions about a wide variety of topics. In the past, the omnibus surveys were conducted with paper questionnaires. Dozens of interviewers delivered blank questionnaires to the homes of the selected people, and after a while they collected the completed questionnaires again. Slot (2009) describes the omnibus survey of the city of Amsterdam. In a period of 25 years approximately 100 omnibus surveys were conducted in this city.

Nowadays, many municipalities have a so-called *citizen panel* (in Dutch: *burgerpanel*). This is a web panel the members of which are inhabitants of the municipality. The panel members have agreed to regularly give their opinion about local current affairs.

What all these surveys and polls have in common is that they use sample data to draw conclusions about the population as a whole. The idea of survey sampling was accepted only after a long period of discussion. This discussion lasted almost 40 years, from 1895 to 1934. Initially, one believed that it was not possible to draw valid conclusions about a population just using sample data. Now there is general consensus that surveys based on probability sampling are a sound means of research. Nevertheless, there are still many surveys in which the principles of probability sampling are not applied. So, there are good and bad surveys.

In 2015 an opportunity presented itself to compare three different types of surveys in the same situation. There was a lot of discussion in the municipality of Alphen a/d Rijn (in the Netherlands) about shopping on Sunday. Should shops be open on Sunday, or should they be closed? The political parties were deeply divided about this. Therefore they decided to ask the inhabitants of the municipality for their opinion.

Contact information: Prof. Dr. Jelke Bethlehem, Albert Verweijstraat 21, 2394 TH Hazerswoude-Rijndijk, The Netherlands (bethlehem@xs4all.nl)

Three different surveys were carried out at the same time, and with the same questionnaire. One survey was based on a probability sample, and the other two used different sample selection techniques. The question was how different the survey results would be. This paper investigates and compares the three surveys.

Section 2 gives a short overview of the historical development of sampling theory, and compares advantages and disadvantages of different approaches. Section 3 describes why there were three different surveys in the municipality of Alphen a/d Rijn, and what the differences were. Section 4 analyses the results of the three surveys. Section 5 draws some conclusions.

2 Historical developments

Data collection for survey research is continuously changing over time, but the basic principles have remained the same. For government statistics, it all started in 1895, the year in which Anders Kiaer, the director of the Norwegian Statistical Bureau, published his *Representative Method*. It was a partial inquiry in which a large number of persons were questioned. This selection should be a "miniature" of the population. Anders Kiaer stressed the importance of representativity. His argument was that, if a sample was representative with respect to variables for which the population distribution was known, it would also be representative with respect to the other survey variables. Kiaer's approach would now be called *quota sampling*. A basic problem of the Representative Method was that there was no way of establishing the accuracy of his estimates. The method lacked a formal theory of inference.

It was Bowley (1906) who made the next step towards a formal theory of survey methodology. He proposed to select samples at random. As a consequence, the theory of probability could be applied. It could be shown that for large samples, selected at random from the population, estimators had an approximately *normal distribution*. The variance of estimators could be estimated, and this variance could be used as an indicator of the precision of estimators.

From this moment on, there were two methods of sample selection. The first one was Kiaer's Representative Method, based on quota sampling, in which representativity played a crucial role, and for which no measure of the accuracy of estimates could be obtained. The second was Bowley's approach, based on random sampling, and for which an indication of the accuracy of estimates could be computed. The discussion about both methods lasted until 1934, in which year the Polish scientist Jerzy Neyman published his now famous paper; see Neyman (1934). Neyman developed a new theory of sampling based on the concept of the confidence interval. He also showed, by making an empirical evaluation of Italian census data, that the Representative Method failed to provide satisfactory estimates of population characteristics.

The history of opinion polls goes back to 1824. In that year, two newspapers, the Harrisburg *Pennsylvanian* and the Raleigh *Star*, attempted to determine political preferences of voters prior to the presidential election. These early polls did not pay much attention to sampling aspects. Therefore, the accuracy of results could not be established. It took until the 1920s before it was realized that the sampling mechanism was important. Then it was George Gallup who started to use quota sampling for his polls. Gallup sent out hundreds of interviewers across the country. Each interviewer was given quota for different types of respondents: so many middle-class urban women, so many lower-class rural men, etc. So, the approaches of Kiaer and Gallup were similar.

The presidential election of 1936 turned out to be decisive for sampling in opinion polls. The two main active polling organizations were *Gallup* and the *Literary Digest Magazine*. The Literary Digest Magazine conducted regular "America Speaks" polls. It based its predictions on returned questionnaires that were sent to addresses obtained from telephone directories and automobile registration lists. The sample size of these polls was very large: over two million people. Gallup's poll was based on a quota sample of "only" 50,000. Gallup correctly predicted Franklin Roosevelt to be the new president, whereas Literary Digest incorrectly predicted that Alf Landon would beat Franklin Roosevelt. The explanation was a fatal flaw in the sampling procedure of the Literary Digest's poll. The automobile registration lists and telephone directories where not representative samples. In the 1930s cars and telephones were typically owned by the middle and upper classes. More well-to-do Americans tended to vote Republican, and the less well-to-do were inclined to vote Democrat. Therefore, Republicans were overrepresented in the Literary Digest sample. As a result of this historic mistake, the Literary Digest magazine ceased publication in 1937. And opinion researchers learned that they should rely on more scientific ways of sample selection.

Gallup's quota sampling approach turned out to work better than Literary Digest's haphazard selection approach. Jerzy Neyman had shown already in 1934, however, that quota sampling can lead to invalid estimates. Gallup was confronted with the problems of quota sampling in the campaign for the presidential election of 1948. Harry Truman was the Democratic candidate and Thomas Dewey was the Republican candidate. The sample size of Gallup's poll was 3,250 persons. At the same time, Leslie Kish selected a probability sample of less than 1,000 people, and concluded that Truman would win. Kish was correct; see Pace (2000). Gallup incorrectly predicted that Thomas Dewey would win the election. The cause of this error was that Gallup used quota samples instead of random samples.

Quota samples are not based on random selection. Interviewers are instructed to select groups of people in the right proportions. But this can only be achieved for a limited

number of variables, such as gender, age, level of education and race. Making a sample representative with respect to these variables, does not automatically guarantee representativity with respect to other variables, like voting behavior. The best way to produce a sample which is at least approximately representative with respect to all survey variables is to apply random sampling and give everyone in the population the same selection probability. This ensures that no group is systematically over-represented or under-represented.

The theory of probability sampling was more or less completed by Horvitz and Thompson (1952). They showed that unbiased estimators of population characteristics can always be constructed provided samples are selected by means of probability sampling and every person in the population has a known and strictly positive probability of being selected. Moreover, under these conditions standard errors of estimates, and thus confidence intervals, can be estimated. Therefore it is possible to quantify the accuracy of estimates.

After a long period of discussion, probability sampling became the accepted method for sample selection. The theory of survey sampling was written down in standard works like Cochran (1953) and Kish (1965). Since then, the paradigm of probability sampling has shown to work well in social research, official statistics, and market research. It has allowed researchers to produce valid and reliable survey results. For more about the history of sampling see, for example, Bethlehem (2009), Lienhard (1997), Utts (1999), and Kish (1995).

A vital ingredient of probability sampling is the availability of a sampling frame. This is a list of all members of the target population of the survey. Fortunately, there is a population register in the Netherlands. Government organizations like Statistics Netherlands and Netherlands Institute for Social Research can use this register as a sampling frame for a probability sample. The municipalities have their own local version of the population register. So they can also select random samples from their populations.

And then came the rise of the Internet. With the introduction in 1995 of version 2.0 of the markup language HTML, it became possible to use the World Wide Web for filling in forms, and thus for completing survey questionnaires; see Bethlehem and Biffignandi (2012, chapter 1), for more details. Web surveys rapidly became very popular among survey researchers. This is not surprising as web surveys seem to have (at first sight) some attractive advantages in terms of costs and timeliness:

- Now that so many people have internet access, a web survey is a simple means to get access to a large group of potential respondents. For example, internet coverage in the Netherlands is over 95%;
- Questionnaires can be distributed at very low costs. No interviewers are needed, and there are no mailing and printing costs;
- Surveys can be launched very quickly. Little time is lost between the moment the questionnaire is ready and the start of the fieldwork. It is sometimes even possible to do a web survey in one day.

A web survey seems to be a fast and cheap means for collecting large amounts of data, but there are also methodological issues. One such issue is sample selection for a web survey. Ideally there is a list of e-mail addresses of all persons in the population. A random sample can be selected from the list, and then an e-mail with a link to the questionnaire is sent to all selected persons. Unfortunately, such an e-mail list is almost never available. An alternative sample selection procedure in the Netherlands could be to select a sample from the population register, and to send a letter (by ordinary mail) with a link to the questionnaire to the selected persons. This makes a web survey more expensive and more time-consuming. With this, some of the advantages of a web survey are lost. It should also be noted that not every researcher has access to the population register, privacy laws forbid this.

Problems with selecting a random sample for a web survey have caused many researchers to avoid probability sampling. Instead, they rely on *self-selection*. The questionnaire is simply put on the web. Respondents are those people who happen to have internet, visit the website and decide to participate in the survey. So the researcher is not in control of the selection process. Selection probabilities are unknown. Therefore, no unbiased estimates can be computed, nor can the accuracy of estimates be determined.

Self-selection web surveys have a high risk of not being representative. For many of these surveys, people outside the target population can also participate. Sometimes it is possible to complete the questionnaire more than once. It is even possible that certain groups in the population attempt to manipulate the outcomes of the survey. Here are three examples that occurred in the Netherlands:

- In 2005, the Book of the Year Award, a high-profile literary prize in the Netherlands was was determined by means of an online survey. People could vote for one of the nominated books or mention another book of their choice. More than 90,000 people participated in the survey. The winner turned out to be the new Bible translation published by the Netherlands and Flanders Bible Societies. This book was not nominated, but nevertheless an overwhelming majority (72%) voted for it. This was the result of a campaign launched by (among others) Bible societies, a Christian broadcaster and Christian newspaper.
- A group of people tried to influence opinion polls conducted during the campaign for the parliamentary elections in 2012. The group consisted of 2,500 people.

They intended to subscribe to an online opinion panel. Their idea was to behave themselves first as Christian Democrats (CDA). Later on they would change their opinion and vote for the elderly party (50PLUS). They hoped this would affect the opinion of other people too. Unfortunately for them, and fortunately for the researcher, their attempt was discovered when suddenly so many people at the same time subscribed to the panel; see Bronzwaer (2012).

- In January 2014 there were local elections in The Netherlands. A public debate between local party leaders was organized in Amsterdam. A local newspaper, *Het Parool*, conducted a web survey to find out who won the debate. Campaign teams of two parties (the Socialist Party and the Liberal-Democrats) discovered that after disabling cookies it was possible to fill in the questionnaire repeatedly. So the campaign teams stayed up all night and voted as many times as possible. In the morning, both the party leaders had a disproportionally large number of votes. The newspaper realized that something was wrong and cancelled the survey. It accused the two political parties of manipulating the survey. However, it was the newspaper that was responsible for setting up a bad survey; see also Bethlehem (2014).

Self-selection is a form of non-probability sampling. There seems to be a growing discussion about the use of various forms of non-probability sampling. This is because researchers encounter more and more practical problems drawing probability samples. One problem is the lack of proper sampling frames. For online surveys, it would be ideal to have a sampling frame of e-mail addresses. Unfortunately, such sampling frames do not exist (for general population surveys). Telephone surveys also have their problems as there are no telephone lists that cover the population. As a results, one has to rely on some form of random digit dialing, which also has its problems. Another problem of probability sampling are decreasing response rates. The lower the response rate, the more probability sampling resembles non-probability sampling. Response rates of online surveys are often below 40%. And according to Rivers (2007) the response rates of many telephone surveys in the US are even below 20%. Although this was all completely within the rules of the survey, the group of voters was clearly not representative of the Dutch population.

Given the problems with probability sampling, there is an ongoing discussion whether *non-probability sampling* can be used as an alternative. The most obvious example is the AAPOR report on non-probability sampling (see Baker et al., 2013). Gelman (2013), Gelman and Rothschild (2014), and Wang, Rothschild, Goel, and Gelman (2015) take a more favorable position towards non-probability sampling. They claim that the lack of representativity caused by non-probability sampling can be repaired by applying appropriate correction techniques.

A well-known non-probability sampling method is *quota sampling*. The population is divided into strata. Interviewers must select a predetermined number of persons in each stratum, and they are free to choose anyone as long as the person meets the requirements of the stratum. People who are not willing to participate are simply replaced by other people who are willing.

Moser and Stuart (1953) concluded from experiments that quota sampling could produce good results. Others, however, demonstrate that quota sampling cannot be regarded as an acceptable alternative to probability sampling.

Rivers (2007), Vavreck and Rivers (2008), and Rivers and Bailey (2009) propose *sample matching* to reduce nonresponse problems by linking a random sample from a sampling frame to a self-selection web panel. According to Bethlehem (2015), however, sample matching is not better than other correction techniques, like post-stratification. Bethlehem (2010) shows that the potential size of the bias due to nonresponse in a probability survey is much smaller than the self-selection bias in in a non-probability survey.

This discussion makes clear that one has to be careful when using the results of self-selection surveys. Such surveys may easily lead to wrong conclusions being drawn from the results. This can be shown by applying statistical theory (see, for example Bethlehem & Biffignandi, 2012, chapter 9). Sometimes it also possible in practice to compare good and bad surveys. This was the case in the municipality of Alphen a/d Rijn in the Netherlands. In January 2015, there were three different surveys about the same topic (shopping on Sunday), at the same time, and with the same questionnaire. The three surveys had different modes of data collection. In the next sections, these three surveys will be compared in more detail.

3 Three surveys

There has always been a lot discussion in the Netherlands about *shopping on Sunday*. Should shops be open on Sunday, or should they be closed on this day? Politicians have opposing views. On the one hand, liberals believe that shop keepers should be able to decide for themselves whether their shop should be open on Sunday or not. On the other hand, Christian parties want the shops closed, because Sunday is the day of the Lord. It is a day of rest, of going to church, and it is not a day for economic activities. In 1996 there was a new law on Sunday shopping giving all (approximately) 400 municipalities in the Netherlands the possibility to make their own rules.

In Alphen a/d Rijn, a municipality in the western part of the Netherlands, there was also a discussion about rules for Sunday shopping. And also here there were opposing views. Local politicians were not able to find a compromise. So in

the end they decided to ask the inhabitants of the municipality for their opinion.

Having no knowledge of the methodological aspects of good surveys, the local politicians decided to conduct face-to-face interviews in the shopping centers on Saturday afternoon. After a survey-methodologist pointed out that this would probably not lead to a representative sample, they decided to use the *AlphenPanel* for another survey. This is a web panel of inhabitants of the town of Alphen a/d Rijn. The panel members were recruited for a large part by means of probability sampling. Some early members (approximately 450) were recruited by means of an opt-in procedure, but later the panel was refreshed and extended to 1,600 people. All these people were recruited by means of a random sample from the population register of the municipality. The new version of the panel was approximately representative with respect to gender, age and town (in the municipality). Young people were somewhat under-represented and the elderly were a little over-represented. Also there somewhat more males than females in the panel.

Moreover, the politicians even decided to conduct a third survey. The idea was to offer a questionnaire on the internet. There were no restrictions; everybody could complete the questionnaire, even more than once. This self-selection survey was mainly offered as a means to give all inhabitants the possibility to express their opinion.

So an interesting situation occurred in which three surveys were carried out at the same time, with the same target population, and with the same questionnaire. This made it possible to compare the three modes of data collection. It could be assessed whether indeed probability sampling leads to better results in this case. This comparison is the topic of this paper.

Initially, the local politicians in Alphen a/d Rijn decided to do a face-to-face survey in shopping centers, where respondents would be recruited on two Saturdays (10 and 17 January 2015), between 11 AM and 3 PM. Unfortunately, it rained on 10 January. So in the end interviewing only took place on 17 January. Interviews were not carried out by professional interviewers, but by members of the political parties. The interviewers had caps and shawls showing their party membership. One wonders if this helps to create the impression of an objective survey. Another problem was that the interviewers were not very well instructed. For example, they did not know what to do with shoppers living outside Alphen a/d Rijn. Should they be included in the survey or not? Some interviewers included them, and others did not.

It is very unlikely that this data collection approach will result in a representative sample from the population of all inhabitants of Alphen a/d Rijn. At most one can say that the sample is representative of all Saturday afternoon shoppers in town. This is a different target population than the target population of all inhabitants. For example, it does not include people who, for whatever reason, do not shop on Saturday.

There are other examples of surveys producing biased results because of the method of data collection. One such example is a radio listening poll of a local radio station in the Netherlands. They also did their survey on a Saturday afternoon in the local shopping centre. One of the outcomes of the survey was that almost no one listened to the sports program on Saturday afternoon. This is not surprising if one does data collection on the same Saturday afternoon in a shopping centre. The conclusion is wrong because listeners to the sports program are excluded from the survey.

After having consulted a survey methodologist, the politicians realized that this was not the best way to do a survey. This expert suggested to use the *AlphenPanel* for their survey. Many Dutch municipalities have so-called *citizen panels*. These web panels usually contain a few hundred citizens. The panels are consulted a few times per year about current local policy issues. One of the objectives is to bring politicians and citizens closer together. The AlphenPanel is the citizen panel of Alphen a/d Rijn. At the time approximately 1,600 members.

The usefulness of this panel depends, for a large part, on its representativity. It was not completely clear how representative the AlphenPanel was. The distribution over variables like gender, age and town in the municipality looked reasonably good. However, this does not automatically mean that the panel is also representative with respect to other variables, like the opinions about Sunday shopping . Fortunately, panel recruitment was, for a large part, based on a random sample from the population register of the town of Alphen a/d Rijn. It was, however, also possible to sign up for the panel spontaneously. So there was also some self-selection. This may have affected representativity. Nevertheless, it was considered likely that the web panel would produce better results than the shopping center survey.

The local politicians decided to carry out yet another survey, and this was a self-selection survey on the internet. Everyone could complete the questionnaire. The basic idea behind this survey was that all inhabitants should be given the opportunity to express their opinion about the issue of shopping on Sundays. The questionnaire could be accessed through the municipality website.

An online self-selection survey has a number of important disadvantages. The first one is that everyone can participate in the survey, even people from outside the target population. For example, inhabitants from neighboring municipalities could fill in the questionnaire of the Alphen a/d Rijn survey, and, in this way, influence the outcomes of the survey. There is anecdotal evidence that this actually happened.

A second disadvantage of self-selection surveys is that they are usually not representative. The participants are typically people who like doing surveys or are interested in the topic of the survey. Research has also shown that some

groups in the population are under-represented in web surveys, like the elderly, low educated people, and people with an ethnic minority background (see, for example Bethlehem & Biffignandi, 2012, chapter 9).

A third disadvantage of self-selection surveys is the possibility to manipulate their outcomes. Some examples of survey manipulation were mentioned in section 2. It was possible to complete the Sunday shopping survey more than once. Some people admitted they did the survey more than 10 times.

The self-selection survey about shopping Sundays was also affected by attempts to influence the outcomes. The consistory of the Reformed Church in Boskoop (one of the towns in the municipality) wrote on her website:

> We would like to call on you to participate in this survey. There is a trend in our society to see Sunday more and more as a normal day instead of a holy day, a day of seclusion and tranquility. Therefore, let us together take responsibility, while prayerfully looking to the Lord who rules and governs everything.

The Dutch Reformed Church in the town of Benthuizen (another town in the municipality) also appealed to her members to vote in the survey:

> On the website of the municipality you can complete a questionnaire about the topic of shopping Sundays. The board of the church recommends completion of this questionnaire wholeheartedly.

Moreover, the municipal council invited inhabitants to complete the survey questionnaire more than once:

> You are also cordially invited to participate in this survey, even if you already completed a questionnaire in a survey in one of the shopping centers.

Attempts to motivate specific groups of people to participate in the survey do not help to obtain a representative sample. To the contrary, they will increase the lack of representativity. As a result, outcomes will be seriously biased.

Looking at the three data collection designs (the face-to-face survey in the shopping centers, the survey from the AlphenPanel, and the self-selection survey on the internet), it is likely the panel comes closest to a representative sample. Therefore, this survey can best be used to get a good idea of the opinion of the people in Alphen a/d Rijn about shopping Sundays. The other two survey will probably produce biased results.

An often encountered misunderstanding is that problems in surveys will disappear if the sample size is increased. Unfortunately, this is not the case. Biases due to shortcomings in the sample design will remain whatever the sample size. Therefore, it is not a very good idea to combine the data of the three shopping Sundays surveys into one data set. It would only "pollute" the reasonably good data set of the panel survey. In the end, the politicians in Alphen a/d Rijn decided wisely to only use the panel survey data for policy decision making.

Not using two of the three surveys means throwing away a lot of data. That is a waste of what can be precious information. So one could consider applying correction techniques to improve the representativity of the two "bad" survey by applying some kind of weighting technique. This only works, however, if sufficient effective weighing variables are available. This was not the case here.

4 The results

The AlphenPanel was an initiative of the municipality of Alphen a/d Rijn. The first surveys from this panel took place in 2011. Practical management of the panel and conducting the surveys was in the hands of the market research company I&O Research. Also the self-selection survey was carried out by I&O Research. The organization and the fieldwork for the face-to-face survey in the shopping centers were done by the politicians themselves.

The results of the three surveys were published on 3 March 2015. It turned out that 754 people had completed the questionnaire in the shopping centers. The self-selection survey produced 1,550 completed forms. In the AlphenPanel 857 members completed the questionnaire, out of a total of 1600 members who were invided to participate. This comes down to a response rate of 54%. Taking into account the topic of the survey, and the fact that all panel members agreed to participate in surveys, one would have expected a higher response rate.

The municipality of Alphen a/d Rijn has 107,000 inhabitants. Approximately 66% of the people live in the urban town with the same name. The other 34% live in seven small rural towns around the urban area (source: municipality of Alphen a/d Rijn). For the panel survey and the self-selection survey, the town in which the respondents live, was recorded. This made it possible to compare the distribution of the towns in these surveys with the distribution in the population. Table 1 contains the data.

The response distribution in the panel resembles the distribution in the population. The largest difference is with respect to the percentage of people in the town of Alphen a/d Rijn. 66% of the population lives in this town whereas 70% of the panel respondents come from this town. This is a difference of 4 percentage points. For all other towns the difference is at most 1 percentage point. So one can conclude that

Table 1
Distribution of the respondents over the towns (in %)

Town	Panel	Self selection	Population
Aarlanderveen	2	1	1
Alphen a/d Rijn	70	55	66
Benthuizen	3	13	3
Boskoop	13	18	14
Hazerswoude-Dorp	4	8	5
Hazerswoude-Rijndijk	4	2	5
Koudekerk a/d Rijn	3	2	4
Zwammerdam	1	1	2
Total	100	100	100

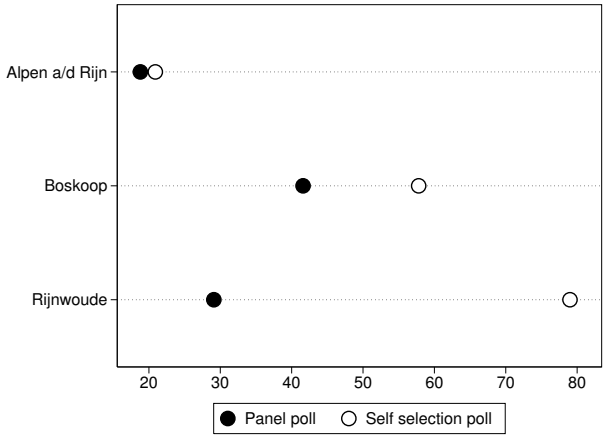

Figure 1. Percentage opponents of Sunday shopping in the regions of the municipality (unweighted estimates)

the panel survey is reasonably representative with respect to town of residence.

There are problems with the representativity of the self-selection survey. There are substantial differences between the percentages in the population and the percentages in the survey. For example, only 3% of the population lives in the town of Benthuizen, but no less than 13% of the self-selection respondents are from this town. There is also an over-representation of people from the towns of Boskoop (18% instead of 14%) and Hazerswoude-Dorp (8% instead of 5%). A logical consequence of the over-representation of these three towns is that one or more other towns are under-represented. This is indeed the case for the town of Alphen a/d Rijn. 66% of the population lives in this town but in the self-selection survey it is only 55%.

Why are the three towns Benthuizen, Hazerswoude-Dorp, and Boskoop over-represented in the self-selection survey? A plausible explanation is that these three towns are part of, or close to, the Dutch Bible Belt. This is strip of land across the Netherlands that is inhabited by a high percentage of conservative Protestants. For example, in the town of Benthuizen almost 50% voted for conservative Protestant parties in the local elections of 2010, while the average in the country was around 6%. There are many conservative Protestants in Benthuizen, Hazerswoude-Dorp and Boskoop, and they were asked by their churches to participate in the self-selection survey. So one can expect these people to be over-represented in this survey.

To obtain more insight into the effects of the lack of representativity, the percentage of opponents of Sunday shopping was computed in various regions of the municipality. The result is the dot chart in figure 1.

The three regions (Alphen a/d Rijn, Boskoop and Rijnwoude) correspond to the old municipalities that merged into the new municipality of Alphen a/d Rijn on 1 January 2014. The "old" Alphen a/d Rijn consisted of the urban town Alphen a/d Rijn, and the rural towns of Aarlanderveen and Zwammerdam. Rijnwoude consisted of the four rural towns Benthuizen, Hazerswoude-Dorp, Hazerswoude-Rijndijk and Koudekerk a/d Rijn. The region Boskoop just contains the rural town of Boskoop.

The percentage of opponents of shopping on Sunday in Rijnwoude is only 29.1% in the panel survey, and as much as 79.0% in the self-selection survey. One can conclude that the opponents are heavily over-represented in the self-selection survey. There are also differences between the panel survey and the self-selection survey in Boskoop. The percentage of opponents in the self-selection survey is 57.8% whereas it is only 41.6% in the panel survey.

Figure 2 zooms in on the outcomes of the self-selection survey in the Rijnwoude region. Almost all people in Benthuizen (94%) are opponents of shopping Sundays. The percentage of opponents is also very large in Hazerswoude-Dorp (87%). To the contrary, the percentage of opponents is very small in Hazerswoude-Rijndijk (9%) and Koudekerk a/d Rijn (19%).

This analysis shows the risk of carrying out a survey based on self-selection. Such a survey will often lack representativeness. As a consequence, estimates may be seriously biased. There are strong indications that conservative Protestants are over-represented in the self-selection survey on Sunday shopping. Therefore, estimates for the percentage of opponents of Sunday shopping will be systematically too high.

Surveys may be selective for various reasons. Self-selection is one of them. Another reason is nonresponse. To remove, or at least reduce, a possible bias due to nonresponse, usually some kind of weighting adjustment procedure is carried out. Weights are assigned to respondents in such a way that under-represented groups get a larger weight, and over-represented groups get a smaller weight. This can be done if proper auxiliary variables are available. These are variables that are measured in the survey and for which the

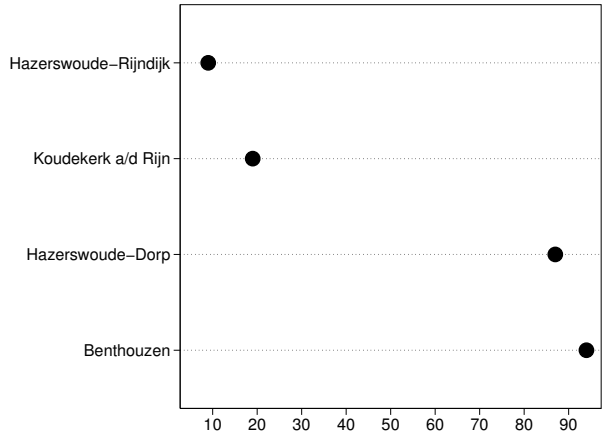

Figure 2. Percentage of opponents in the towns of the region Rijnwoude ; self-selection survey (unweighted estimates)

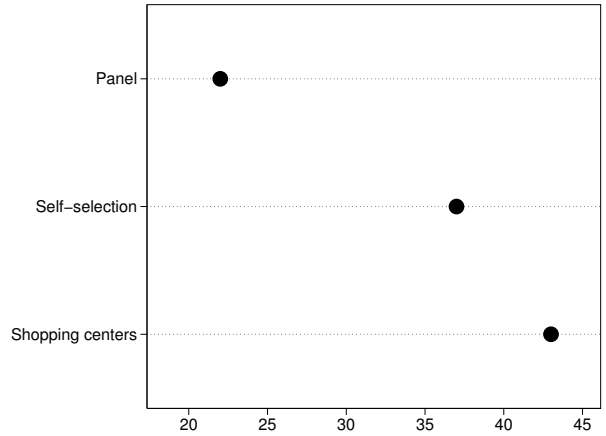

Figure 3. Percentages of opponents of shopping Sundays

distribution in the population is available. Weighting adjustment is only effective if the auxiliary variables are strongly correlated with the target variables of the survey. It is often difficult to find such variables. One source could be the sampling frame (for example a population register), a statistical institute (typically for demographic variables), and sometimes such variables can be extracted from administrative sources. Also *paradata* (data about the data collection process) may be available; see Krueger and West (2014). For more about weighting adjustment see Bethlehem, Cobben, and Schouten (2011).

Adjustment weighting was applied both for the panel survey and the self-selection survey. In both cases, three auxiliary variables were used: gender, age (in four age classes), and region of the municipality (Alphen a/d Rijn, Boskoop, Rijnwoude). One may wonder whether such a limited set of auxiliary variables is capable of removing possible biases.

The objective of the surveys was to obtain more insight in the opinion of the inhabitants about Sunday shopping. Not surprisingly, one of the questions was whether one favored or opposed shopping on Sunday. Figure 3 shows the percentages of opponents of shopping on Sundays. Note that weighting adjustment was applied to the panel survey and the self-selection surveys, and not to the face-to-face survey in the shopping centers.

The estimates differ substantially. They range from 22% to 43%. Given the ways in which the three surveys were conducted, one can expect the 22% of the panel survey closer to the true value in the population than the other two estimates. Therefore, 22% is the best guess for the percentage of opponents. However, it must be taken into account that the estimate is based on a sample from the population. So there are margins of error. For the estimate of 22%, the margin is approximately 3 percentage points. This means that with a high probability the true value will be between 19% and 25%. It must also be taken into account that the panel suffered from nonresponse in the recruitment process. This may cause the estimate to have some bias.

The estimate based on the self-selection survey is much higher: 37% instead of 22%. So there is a difference of 15 percentage points. Note that no statistical test was carried out to determine whether the difference was significant. This is not possible because the underlying distribution of the self-selection poll was unknown. However, the difference is so large that it cannot be attributed to sampling error. Even after weighting adjustment, this estimate is much higher. Apparently, the conservative Protestants are still over-represented. Weighting by region does not sufficiently help to reduce the over-representation of towns like Benthuizen and Hazerswoude-Dorp.

The face-to-face survey in the shopping centers produces an even larger estimate: 43%. It is almost double the value from the panel survey (22%). Without more research no clear explanation can be given for this large value. Maybe Saturday shoppers do not have a need to shop on Sundays. Furthermore, people who like to shop on Sundays because it is not possible for them to shop on Saturday, will not be included in the survey.

The differences between the three surveys are too large to be able to attribute them to random sample fluctuations. There are significant systematic differences. The only conclusion that can be drawn is that the self-selection survey and the face-to-face survey in the shopping centers are wrong. Their results should not be used.

The survey in the shopping centers seems to be the worst survey of the three. It should be noted that this not because it is a face-to-face survey. Usually, face-to-face surveys perform well, because interviewers can persuade people to participate in the survey and they can assist the respondents in giving appropriate answers to the questions. The survey in the shopping centers is bad because the sample selection

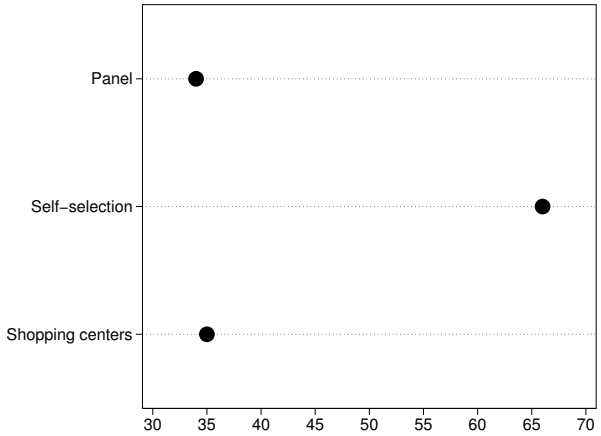

Figure 4. Opposed to Sunday shopping because of religion

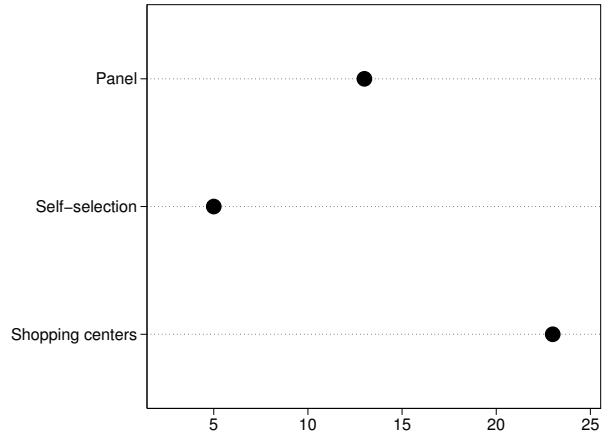

Figure 5. Sunday shopping should only apply to supermarkets, Do-It-Yourself shops, and garden centers?

mechanism produces samples that are far from representative.

More questions were asked in the surveys. For example, the opponents of Sunday shopping were asked about their reasons. One of the answer options was that they were against Sunday shopping Sundays of their religion. Figure 4 shows the estimates of the percentages opposed to shopping because of their religion.

The best estimate is probably that of the panel (34%). One can conclude that one out of three opponents is against shopping on Sunday because of religious reasons. The estimate based on the self-selection survey is almost twice as large (66%). Again, this is an indication that conservative Protestants are over-represented in the self-selection survey. The estimate for the shopping centers is (by chance?) close to the estimate for the panel survey.

A final example is the question of whether Sunday shopping should apply to all shops, or only to supermarkets, Do-It-Yourself shops, and garden centers. This question was asked only of those in favor of shopping Sundays.

Figure 5 shows that estimates go in all directions. There is no clear pattern. Assuming the panel estimate is closest to the true population value, the estimate for the self-selection survey is much too low, and the estimate for the survey in the shopping centers is much too high.

5 Conclusion

Investigation of the public opinion on Sunday shopping in Alphen a/d Rijn made it possible to compare three surveys: a face-to-face survey in the shopping centers, a survey from a representative online panel, and an online survey with self-selection. Although all three surveys were conducted at the same time, and used the same questionnaire, their outcomes were very different. Assuming the panel survey estimates are closest to true population value, the self-selection survey and the shopping center survey must be bad polls.

Many municipalities in the Netherlands have a so-called *citizen panel*. Such a panel can be used for conducting surveys and polls in a meaningful way provided the panel is representative of the population. This is only possible if people are recruited for the panel by means of probability sampling. One way for municipalities to achieve this is to select a random sample from the population register, and invite the selected people to become a member of the panel. Every effort must be made to avoid nonresponse in the recruitment process, as nonresponse may affect the representativity of the panel. If the panel is used to conduct a poll, it might be wise to conduct a weighting adjustment technique to correct for a possible lack of representativity.

An online survey with self-selection is a bad measurement instrument. Representativity can be affected in various ways: people from outside the target population can participate, and in this way "pollute" the survey. Furthermore, people may be able to complete the questionnaire more than once. Moreover, groups of people can attempt to manipulate the outcomes of the survey. In the case of the Sunday shopping survey, conservative Protestants tried to get a majority in favor of keeping shops closed on Sunday.

The face-to-face survey in the shopping centers is also a bad survey. In fact, this approach reduces the target population of all inhabitants of the municipality to only those who shop on Saturday afternoon. There is no guarantee at all that this sub-population is representative of the whole population. Indeed, the estimates differ substantially from those of the panel survey.

A weighting adjustment procedure was applied to the results of the self-selection survey. After weighting, there were still large differences between the estimates of the self-selection survey and the panel survey. So, weighting adjustment did not help to repair the lack of representativity. This

may be caused by the limited set of weighting variables (gender, age and region) that were insufficiently correlated with the important survey variables. This shows that weighting is only effective if proper weighting variables are used.

The three surveys in Alphen a/d Rijn once more show how important it is to apply a proper sampling technique for surveys. And "proper" means that probability sampling must be applied.

References

Baker, R., Brick, J. M., Bates, N., Battaglia, M., Couper, M., Dever, J., ... Tourangeau, R. (2013). *Report of the AAPOR task force on non-probability sampling*. Deerfield: American Association of for Public Opinion Research.

Bethlehem, J. (2009). *The rise of survey sampling*. Discussion Paper 09015, Statistics Netherlands, The Hague/Heerlen.

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, *78*, 161–188.

Bethlehem, J. (2014). Slechte peiling, gemanipuleerde uitslag [bad poll, manipulated result]. PeilingPraktijken. Alles over goede en slechte peilingen, January 13. Retrieved from http://peilingpraktijken.nl/weblog/2014/01/slechte-peiling-gemanipuleerde-uitslag/

Bethlehem, J. (2015). Solving the nonresponse problem with sample matching? *Social Science Computer Review, online first*. doi:10.1177/0894439315573926

Bethlehem, J. & Biffignandi, S. (2012). *Handbook of web surveys*. Hoboken, NJ: John Wiley & Sons.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys*. Hoboken, NJ: John Wiley & Sons.

Bowley, A. (1906). Address to the Economic Science and Statistics section of the British Association for the Advancement of Science. *Journal of the Royal Statistical Society,A*, *69*, 548–557.

Bronzwaer, S. (2012). Infiltranten probeerden de peilingen van maurice de hond te manipuleren [infiltrants attempted to manipulate the polls of maurice de hond]. NRC, September 13. Retrieved from http://www.nrc.nl/nieuws/2012/09/13/infiltranten-deden-hackpoging-peilingen-maurice-de-hond

Cochran, W. (1953). *Sampling techniques*. New York: John Wiley & Sons.

Gelman, A. (2013). Yes, worry about generalizing from data to population. but multilevel modeling is the solution, not the problem. Statistical Modeling, Causal Inference, and Social Science, July 11. Retrieved from http://andrewgelman.com/2013/07/11/19334/

Gelman, A. & Rothschild, D. (2014). When should we trust polls from non-probability samples? The Washington Post, April 11. Retrieved from http://www.washingtonpost.com/blogs/monkey-cage/wp/2014/04/11/when-should-we-trust-polls-from-non-probability-samples/

Horvitz, D. & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*, 663–685.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, *2*, 813–830.

Krueger, B. & West, B. (2014). Assessing the potential of paradata and other auxiliary data for nonresponse adjustments. *Public Opinion Quarterly*, *78*, 795–831.

Lienhard, J. (1997). Gallup poll. engines of our ingenuity. Engines of our ingenuity, No. 1199. Retrieved from http://www.uh.edu/engines/epi1199.htm

Moser, C. & Stuart, A. (1953). An experimental study of quota sampling. *Journal of the Royal Statistical Society, A*, *116*, 349–405.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, *97*, 558–606.

Pace, E. (2000). Leslie Kish, 90; improved science of surveys. The New York Times, October 14. Retrieved from http://www.nytimes.com/2000/10/14/us/leslie-kish-90-improved-science-of-surveys.html

Rivers, D. (2007). *Sampling for web surveys*. Paper presented at the Joint Statistical Meetings, Section on Survey Research Methods, Salt Lake City, Utah.

Rivers, D. & Bailey, D. (2009). *Inference from matched samples in the 2008 U.S. National Elections*. Paper presented at the 64th Annual Conference of the American Association for Public Opinion Research, Hollywood, Florida.

Slot, J. (2009). Vragen? Geen vragen! [Questions? No questions!] Retrieved from http://www.onsamsterdam.nl/

Utts, J. (1999). *Seeing through statistics*. Belmont. California. USA: Duxburry Press.

Vavreck, L. & Rivers, D. (2008). The 2006 cooperative congressional election study. *Journal of Elections*, *18*, 355–366.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting (in press)*. doi:10.1016/j.ijforecast.2014.06.001