Measuring the coverage bias in landline telephone surveys by comparison of Swiss registry data with commercially available telephone number databases

Klug, Stefan; Arn, Birgit

methods, data, analyses | 2016, pp. 167-194 DOI: 10.12758/mda.2016.007

Measuring the Coverage Bias in Landline Telephone Surveys by Comparison of Swiss Registry Data with Commercially Available Telephone Number Databases

Stefan Klug & Birgit Arn
DemoSCOPE

Abstract

Coverage of the population within the sampling frame is a very important quality characteristic of a study. However, a metrical evaluation of the coverage bias to approach the question of representativeness is usually not possible.

Switzerland stands out in that the federal statistical office (SFSO) has legal access to population registers (person universe) and a full list of landline telephone numbers (phone number universe). However, these data are not available for research institutes, which must rely on commercially available number collections or RDD sampling frames.

This paper wants to quantify the coverage bias of such alternative sampling frames by metric calculation of their congruence with the SFSO universes.

The analysis shows that 85.0% of private phone numbers and 88.9% of the resident population of Switzerland that can be reached via landline by the SFSO definition (non-ALTELs) are included in our exemplarily analyzed commercially available phone number collection. The highest coverage bias is present in the 20-39 age group. The RDD frame covers 97.8% of private phone numbers and 99.8% of non-ALTEL persons. Hence, both available alternative sampling frames are useful for representative studies.

Finally, the potential of use of the Swiss coverage results as benchmarks for other countries is discussed.

Keywords: CATI surveys; coverage bias; RDD; representative; Switzerland; mobile research

1 Introduction

In the early 1990s, landline penetration rates were close to saturation in most European countries (Busse et al, 2012). However, an increase in households and persons available through mobile phones only has taken place in the last decade. Additionally, willingness to publish private landline phone numbers has decreased. Due to this increase of mobile-only households and unlisted landline numbers, the usability of landline phone numbers for high quality surveys has deteriorated noticeably.

This is an international trend that is also observable in Switzerland. Studies showed that only 92% of Swiss households still had a landline phone (Stähli, 2012) and the quality of phone number samples has decreased since then; additionally, Swiss citizens are no longer obliged to list their phone numbers in the public directory. The inevitable coverage bias can lead to a significant error in the survey results, as households and persons are missing completely from the sampling frame. For example, telephone surveys imply bias related to income and household size (Stähli, 2012). According to Schouten and Bethlehem (2009), the sampling frame has to be complete to guarantee a representative response set.

In Switzerland, the SFSO uses a sampling frame called SRPH (SRPH, 2016), which contains the total resident population, for its surveys. The universe of the Swiss resident population is obtained through consolidation of municipal, cantonal and federal registers in one general data warehouse. It reflects the population at a precise reference date and is updated quarterly. SRPH therefore comes very close to a full population person and household sampling frame, as it also contains information on people living together in one household.

Additionally, the SFSO has been granted access to a list of all published and unpublished, private and business phone numbers provided by all operators in Switzerland by law[1]. This list, called Emergency Call Data Base (ECDB), repre-

1 Art. 10, Abs. 3quater of the Bundesstatistikgesetzes (SR 431.01) and Art. 16 of the Registerharmonisierungsgesetz (SR 431.02). Artikel 13a tos 13g of the Statistikerhebungsverordnung (SR 431.012.1), see also: http://www.bfs.admin.ch/bfs/portal/en/index/institutionen/oeffentliche_statistik/rechtliche_grund/bund.html

Direct correspondence to
DemoSCOPE, Klusenstrasse 17/18, 6043 Adligenswil, Switzerland
E-mail: stefan.klug@demoscope.ch / birgit.arn@demoscope.ch

Acknowledgements: We would like to thank the Swiss Federal Statistical Office (SFSO) for its great support. Our special thanks go to Claude Joye and Christoph Freymond for their excellent work and high quality analysis and contributions.
Furthermore, we thank AZ Direct and BIK Aschpurwis + Behrens for their support of our idea, and the possibility to use their data and publish the analysis results.

sents the complete universe of landline phone numbers. At the moment, mobile phone numbers are not part of the database.

SRPH can be used to sample from the complete Swiss resident population, which enables the SFSO to draw truly representative person samples. However, for a CATI survey a phone number for each survey entity is needed. Hence, in a second step, the ECDB is scanned for a phone number for each sampled person. A positive match is usually found for about 72% of the sample. If no phone number match is found, that person is called an ALTEL person by the SFSO.

However, access to SRPH and ECDB is restricted by law and not available for surveys conducted by private market and social research institutes. Data collections from commercial providers, and as a further option RDD sampling frames, must be used by market and social research agencies to conduct research by CATI surveys. Therefore, the agencies rely heavily on these other sources for sampling and the quality of these data.

Within this project, it was possible to compare ECDB and SPRH with a commercially available phone number collection and RDD samples in order to quantify the coverage bias of these sampling frames. Since the SFSO frames do not include mobile phone numbers, this analysis is restricted to landline phone numbers. Nevertheless, the range of the Swiss mobile-only penetration can be estimated from this analysis, as the maximum penetration is given by the share of persons from SRPH where no phone number match in ECDB can be found (ALTEL). As the phone number collection contains also address parameters, a match between these parameters and SRPH was also conducted in a final analysis. Note that several authors tried to access the topic of matching SRPH data and phone number collections by address parameters. However, access was always restricted to a specific sample from SRPH (Lipps et al. 2013, Lipps et al. 2015).

Coverage of a sampling frame can be defined as the percentage of landline numbers or persons in this frame that can also be found in the phone number (ECDB) of person (SRPH) universe. Coverage bias is defined as 100% coverage. As SRPH contains some socio-demographic variables, the qualitative aspects of coverage bias can be described by demographic attributes such as age or canton.

Calculation of the coverage (at a given reference date) for an alternative sampling frame allows researchers to quantify the potential lack of information and barriers to representativeness in this respect. This paper is not intended to pass judgment on 'good' or 'poor' sampling frames. Representativeness is not a dichotomous attribute: it varies from 0% to 100% and is, therefore, a quantitative measure of 'more' or 'less' representative. Ideally, a risk measure of representativeness can be calculated by multiplying the coverage of the sampling frame by the response rate of the samples. The third component – additional bias that originates from the data collection process – cannot be quantified easily and must be taken into account as an estimate of the calculation. This simple calculation can be taken as a quantitative estimate to answer the question of the representativeness of a sample.

In the next section, we will describe examples of other sampling frames most widely used within the Swiss market and social research industry, namely data from AZ Direct (www.az-direct.ch) and a random digit dialing (RDD) frame from BIK Aschpurwis + Behrens (www.bik-gmbh.de).

The available phone number and person sources from the SFSO are described in more detail in Section 3; we will also discuss challenges associated with these SFSO sampling frames. In Section 4, we present the methodology to calculate the coverage bias. Key figures for the comparison of AZ Direct and SFSO data are shown in Sections 5 and 7. We will look at the question of whether the RDD numbers are a useful alternative in Section 6. In Section 8, we discuss the potential and conditions of use of the Swiss coverage results as a benchmark for other countries, and we attempt to analyze the added value of our results for survey researchers outside Switzerland. A general summary of the analysis is provided in Section 9.

2 Commercially Available Sources for Phone Number Samples

In the following chapter, we will describe two important sources of landline phone numbers used by survey agencies in Switzerland: a phone number collection from AZ Direct and RDD data from BIK Aschpurwis + Behrens.

2.1 AZ Direct Data

Switzerland has historically had and still has excellent landline telephone provision (Stähli, 2012). Address management companies can continuously update their databases by gathering information from a multiplicity of sources.

The most frequently used database within the market research industry is that provided by AZ Direct. This company offers a sampling frame consisting of Swiss phone numbers and a file containing data on persons and households in Switzerland. This person directory is an enriched database containing hard data and additional person and household attributes generated by means of statistical methods and data mining tools. These two sources will be labeled ‘AZ Direct Numbers’ and ‘AZ Direct Person Plus’, respectively.

Important characteristics in the AZ Direct Numbers file are the type of entry (i.e. private, business or private, and business phone number) and whether a phone number is active. This flag signals the current availability of the number in published registers. A further important feature is the language code, which allows people or businesses to be addressed in their most likely first language. This is an important issue in a multi-language country such as Switzerland. The file also contains address information and a PersonID. Hence, it is possible to identify all numbers that belong to the same person. For all analyses in this paper, we used AZ Direct data from second quarter 2014.

The AZ Direct Numbers database consists of 5.2 million telephone numbers. Excluding numbers that are not landline numbers and keeping the Swiss numbers only (excluding Liechtenstein), we have 4.6 million landline phone numbers available for our analysis. About 3.0 million (66.0%) of these are stated as active numbers. Furthermore, private and business numbers are flagged. Note that *numbers (2.3 million, 43.8%) are not allowed to be called for marketing (i.e. sales) purposes, but can be called for market research by specific research institutes. It is a great advantage to Swiss market and social research institutes to have access to those people whose willingness to participate in CATI surveys has not been spoilt by telemarketing activities.

For 69% of entries in the AZ Direct Numbers file, AZ Direct offers additional information that can be used to restrict the selection of samples (‘AZ Person Plus’). This dataset is predicated on the basis of persons rather than phone numbers. The additional information consists of address-based information, but also information on person or household attributes; e.g. the economic status of the head of the household. The PersonID is the unique link between AZ Direct Numbers and AZ Person Plus.

Note that the AZ Person Plus file cannot be used to draw a representative sample of the population. Register-based information is not accessible to private organizations such as AZ Direct and so it has to be assumed that certain selection characteristics apply to the data collection routines of AZ Direct.

2.2 RDD Sampling Frames from BIK

As an alternative to landline phone number collections, RDD (random digit dialing) offers access to a theoretically fully covered phone number sampling frame. Phone numbers in Switzerland in general are structured in such a way that the region of the landline number (or the provider of the mobile number) can be identified by the three-digit area code. Numbers can be attributed to telephony operators by number blocks and this information is publicly available. Note, however, that today a telephone number can be taken to another region or provider, and thus the system does not follow this rule any longer.

Different methods of generation of RDD numbers are described in Gabler & Häder (2007A, 2007B). Pure random digit dialing has a low hit-rate and is, therefore, inefficient. Hence, they propose a strategy in which those randomized two-digit randomization blocks are identified where at least one registered telephone number can be found. Subsequently, every possible number in these two-digit blocks is generated. This increases the hit-rate and, furthermore, each telephone number is equally probable.

As an alternative strategy, BIK Aschpurwis + Behrens proposes that all the number blocks assigned to telephony providers in Switzerland are used. The 10,000 blocks can be downloaded from an official website (Number Blocks, 2016). Note that BIK Aschpurwis + Behrens extracted only those blocks assigned to private operators. This extraction results in a universe of 37 million phone numbers.

A further idea is to compare the performance of 10 (one-digit randomization), 100 (two-digit randomization), 1,000 (three-digit randomization) and 10,000 (four-digit randomization) randomized blocks. This means that one to four of the last digits are cut from the known blocks and complete phone numbers with all possible digit combinations are generated. This method can be applied both to the Gabler & Häder and the BIK method.

The larger the randomized block (10, 100, 1,000, 10,000), the larger the quantity of generated numbers and, hence, the larger the necessary dialing effort. However, the larger the randomization block size, the higher the coverage of the frame. So it is important to find the right trade-off between the amount of numbers and the coverage of the frame.

The dialing effort of RDD samples can be decreased through use of predictive dialing. Predictive dialing is a specific routine of the computerized dialer that predicts agent availability on interview length and other parameters. Based on this prediction, the dialer starts more calls than the number of agents that are actually available. However, predictive dialing is not a necessary prerequisite for RDD sampling. It is a potentially helpful technique that allows high quality RDD samples when sampling costs have to be reduced. For the integration of mobile phone numbers in dual-frame sampling in particular, RDD mobile sampling is the only reliable sample source, and predictive dialing is required to contact all sampled numbers within time and cost limitations (Klug et al., 2014).

See Table 1 for the comparison of the Gabler & Häder method with the BIK Aschpurwis + Behrens method. All published phone numbers in 2013 were used for the Gabler & Häder method. The number blocks for the BIK method were downloaded on July 1, 2014. The number of blocks for both methods is compared with one to four-digit randomization blocks. It can be seen that the number of blocks is always higher for the BIK method. Hence, the coverage of this method might be better than that of the Gabler & Häder method. However, a larger amount of numbers must be generated and dialed.

For the analysis in Section 6, RDD data generated by the Gabler & Häder method were used. RDD numbers have the drawback that they do not contain any address information. Although the first contact language can be roughly estimated from the area codes, true information for regional stratification is not available.

Table 1 Number of blocks for different randomization sizes, BIK and Gabler & Häder method

Block (randomization) size	# Blocks BIK method	# Blocks Gabler & Häder method
10 (one-digit)	3,728,000	820,611
100 (two-digit)	372,800	125,718
1,000 (three-digit)	37,280	19,048
10,000 (four-digit)	3,728	3,455

Based on published phone numbers, regional spreads of RDD numbers can be estimated and used for stratification.

It has been shown in numerous studies that an invitation letter together with an a-priori incentive will increase participation rates and reduce non-response bias (O'Toole et al., 2008). This is not possible with randomly generated numbers. The advantage of a potentially lower coverage bias is therefore diminished by the disadvantage of a higher non-response bias (higher as if with invitation letters).

3 SFSO Sampling Frames for Phone Numbers and Persons

The SFSO has access to a full list of landline phone numbers, including those that are not listed in public directories (ECDB). Additionally, SRPH contains a list of the Swiss resident population at a reference date.

As noted in Section 1, the list of all phone numbers is called Emergency Call Database (ECDB). Using data from the second quarter 2014, it contains about 4.1 million phone numbers, private, business and administrative. Additionally, the ECDB contains address variables and a regional/cantonal identifier for each telephone number.

The SFSO also works with a subset of the ECDB that contains all private numbers and which is relevant for population survey samples. It contains approximately 3.0 million numbers, 73.3% of all numbers in the ECDB. This SFSO sampling frame is called CASTEM (Cadre de sondage pour le tirage d'échantillons de ménages). A Venn diagram and the exact sizes of the subsets can be found in Figure 1 (left).

The identification of private numbers is made by an algorithm developed by the SFSO: all numbers where the address parameters contain a first name are judged as private numbers. This procedure might in rare cases lead to incorrect allocations.

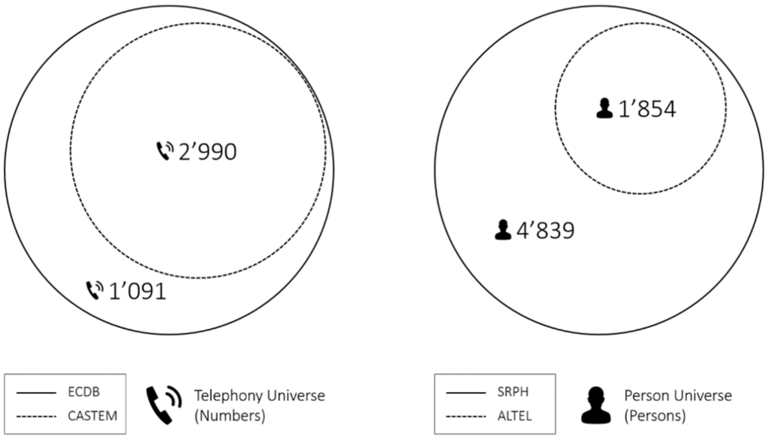

Figure 1 Telephony and person universe and the respective subsets. Numbers are in thousands and add up to the total

A final separation of business and privately used phone numbers cannot be done without direct contact with the number holder.

From 1850 to 2000, the 10-year census provided important information on the structure of the Swiss population. In 2010, a fundamental change took place. Since then, the census has been conducted and evaluated on an annual basis in a new form by the SFSO. In order to ease the burden on the population, the information is drawn primarily from population registers and supplemented by sample surveys. Only a small proportion of the population is surveyed in written form or by telephone. Thus, Switzerland now has a modern statistical system that enables observation of the development of the population and household structure, as well as the structure of buildings and dwellings.

Thanks to this new census system, the SFSO was able to build up the SRPH (SRPH, 2016). For each person in the SRPH, the following variables (in addition to name and address) are known: age, sex, language, nationality, residence permit and canton. The data from all census sources are consolidated and stored in a data warehouse (DWH), see Figure 2. This data warehouse is also the basis of the new SFSO sampling frame.

SRPH, however, does not contain phone numbers. So if a CATI survey has to be conducted, the link between SRPH and ECDB/CASTEM must be constructed. This is done by the SFSO through use of an elaborate matching algorithm that compares how many characters in the address variables of ECDB and SRPH are identical.

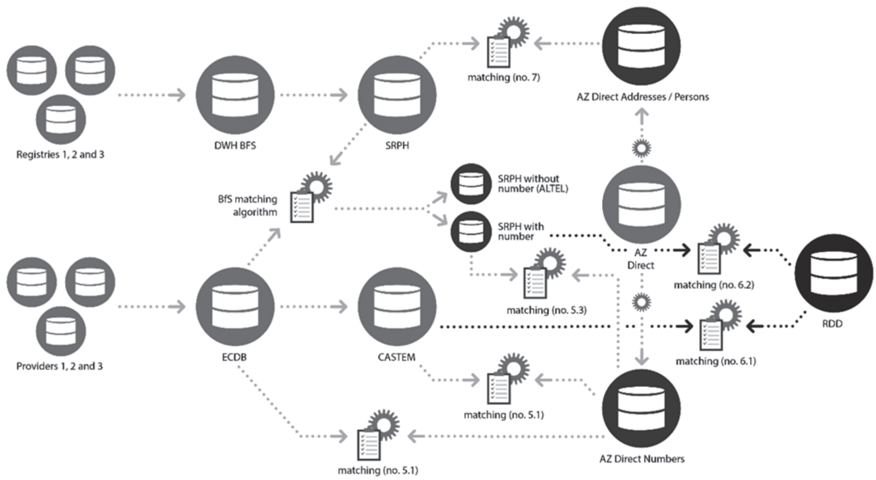

Figure 2 Schema of all data and the section (no.) with calculation of the important key figures

For data from the second quarter 2014, the number matching is possible for only 72.3% of persons in SRPH. Different reasons exist as to why a matching may not be possible (ALTEL persons); for example, if a person does not have a landline phone no match with the ECDB will be found, or the address parameters in the ECDB might be incorrect or obsolete. A Venn diagram and the exact sizes of the subsets can be found in Figure 1 (right). Variations in the matching success can be found; for example, in canton or age (see Figures 7 and 8). The matching percentage is lowest in canton Ticino (TI), whereas people living in Jura (JU) are easiest to identify in the ECDB. Furthermore, identification of phone numbers for persons aged between 20 and 39 is the most difficult. In the following sections, we distinguish between ALTEL and non-ALTEL persons. No sample from the SRPH is taken for our analysis, but we use the whole SRPH frame to compare it with a commercially available landline phone database and an RDD sample frame.

4 Comparison Methodology

Below we describe how coverage bias resulting from use of landline phone number collections or RDD samples can be calculated through comparison of these sets with available SFSO telephony and person universes.

CASTEM (as a subset to the ECDB) is the most complete collection of listed and unlisted private phone numbers. By matching a landline phone number database with CASTEM, the telephone number coverage of this phone number collection (PNC) can be calculated. In our analysis, this calculation is made using AZ Direct Numbers as an example, see Section 5.1. Telephony coverage for private numbers is then defined as $\frac{|PNC \cap CASTEM|}{|CASTEM|}$, where $|\cdot|$ is defined as the size of a set. In Section 6.1, the coverage of an RDD sample is calculated as $\frac{|RDD \cap CASTEM|}{|CASTEM|}$, using the RDD sample from BIK Aschpurwis + Behrens as an example. When calculating the telephony coverage as a key characteristic, the enumerator of the ratio is always the available telephony universe: ECDB for all phone numbers and CASTEM for all private numbers.

In this paper, SRPH defines the total population available for sampling. The coverage of other sampling sources in terms of persons can be calculated by matching them with the persons in SRPH. However, as only landline phone numbers are available, this matching can be done only for those persons in SRPH for whom a phone number can be identified. As we know from Section 3, a phone number can be found only for 72.3% of people in SRPH (non-ALTEL). Hence, for our exemplary alternative sampling sources, person coverage can be calculated as $\frac{|PNC \cap non-ALTEL|}{|SRPH|}$ and $\frac{|RDD \cap non-ALTEL|}{|SRPH|}$, respectively, see Sections 5.3 and 6.2. Note, that the size of $|PNC|$ and $|RDD|$ in this ratio is not defined as the number of phone numbers, but the number of persons. In order to obtain the total person coverage, the enumerator of these ratios must be the size of SRPH. To obtain the coverage for all persons identified by the SFSO, the enumerator can also be the size of the non-ALTELs.

ALTEL persons – as a part of SRPH – are those where no landline phone number from the ECDB can be assigned. However, this does not necessarily mean that no landline phone numbers for these persons exist. As noted above, the address parameters in ECDB connected with a phone number can be incorrect or obsolete. Hence, in Section 7 a matching is made between all SRPH persons and addresses from AZ Direct Numbers and AZ Person Plus. In this analysis, it is particularly interesting to see if the AZ Direct data can add information to the ALTELs for primary contacts via landline phone.

The calculated ratios and coverages are precise and do not need statistical correction.

Figure 2 illustrates all planned analyses and the connection between the different data sources.

5 Comparison of AZ Direct Numbers with ECDB/CASTEM and SRPH

In the following sections, we calculate the coverage of AZ Direct Numbers in terms of phone numbers (ECDB/CASTEM) and persons (SRPH).

5.1 Coverage Concerning ECDB and CASTEM

The AZ Direct Numbers collection contains 4,614,606 numbers that can be used for a match with ECDB (# numbers: 4,081,041) and CASTEM (# numbers: 2,989,632). In general, a match of sets containing numbers results in three subsets of numbers:

1. Numbers in set A only
2. Numbers in sets A and B
3. Numbers in set B only

A match of AZ Direct Numbers and ECDB shows that 1,517,319 numbers are found in AZ Direct only, which is 32.9% of all AZ Direct numbers, see Figure 3. Of these, 488,056 are flagged as active numbers, so in theory these numbers should also appear in the ECDB. A total of 3,097,287 numbers are contained in AZ Direct and ECDB frame, see Table 2 and Figure 3. Hence, 75.9% of the ECDB numbers are also represented in our exemplary landline phone number database.

If we look at the AZ Direct numbers flagged as active in detail, we see that active numbers cover 62.2% of ECDB numbers (see Table 2). This means that coverage of 13.3% (541,101 numbers) is missing, if inactive numbers are excluded from the sampling. Thus, for a market or social research company targeting high representativeness, it is important to also include numbers flagged as inactive. By extension, it is obviously good practice to provide information on formerly active numbers and keep it in the database and sampling frame. We know from previous research that people can be reached by telephone behind inactive numbers, even if at a much lower response rate than if sampling from active numbers only (Diekmann and Bruderer, 2013).

About 38.6% of the numbers flagged with an asterisk (* numbers) can be found in the ECDB. Hence, it is a clear advantage for market and social research companies that such numbers can be sampled and contacted by law.

CASTEM contains 2,989,632 numbers. Hence, the AZ Direct Numbers collection contains many more numbers than CASTEM. Note, however, that the AZ Direct database was not reduced to private numbers using the same reduction logic as for CASTEM. In CASTEM, non-private numbers are selected by filtering addresses with no first name; in the AZ Direct number database, this is done by a flag that separates business and private use of the number. Assuming a combined private and business usage of phone numbers in small businesses (which are most

Figure 3 Resulting subsets from a match of AZ Direct Numbers with ECDB and CASTEM, respectively. Numbers are in thousands and add up to the total

businesses), it makes sense to keep business numbers in the AZ Direct sampling frame and clarify usage in the interview.

The AZ Direct Numbers collection covers 85.0% of the CASTEM frame, see Table 2 and Figure 3. Note that 14.7% of the matching numbers are inactive numbers. Hence, the coverage of AZ Direct numbers is higher if we look at private phone numbers only. For the exact coverage of active and asterisk-flagged numbers and numbers with additional household information in AZ Direct, see Table 2.

Table 2 Coverage of AZ Direct Numbers in terms of ECDB, CASTEM and SRPH

		All	Flagged as active	Asterisk	With additional data
Numbers	$\lvert AZ\rvert$	4,614,606	3,044,242	2,282,966	3,607,221
	$\lvert AZ \cap ECDB\rvert$	3,097,287	2,556,186	1,575,584	2,698,647
	$\frac{\lvert \boldsymbol{AZ \cap ECDB}\rvert}{\lvert \boldsymbol{ECDB}\rvert}$ ($\lvert ECDB\rvert = 4{,}081{,}041$)	**75.9%**	**62.6%**	**38.6%**	**66.1%**
	$\lvert AZ \cap CASTEM\rvert$	2,542,806	2,168,033	1,386,864	2,426,647
	$\frac{\lvert \boldsymbol{AZ \cap CASTEM}\rvert}{\lvert \boldsymbol{CASTEM}\rvert}$ ($\lvert CASTEM\rvert = 2{,}989{,}632$)	**85.0%**	**72.5%**	**46.4%**	**81.2%**
Persons	$\lvert AZ \cap non-ALTEL\rvert$	4,303,048	3,810,267	2,463,010	4,161,634
	$\frac{\lvert \boldsymbol{AZ \cap non-ALTEL}\rvert}{\lvert \boldsymbol{non-ALTEL}\rvert}$ ($\lvert non-ALTEL\rvert = 4{,}838{,}986$)	**88.9%**	**78.7%**	**50.1%**	**86.0%**
	$\lvert AZ \cap non-ALTEL\rvert$	4,303,048	3,810,267	2,463,010	4,161,634
	$\frac{\lvert \boldsymbol{AZ \cap non-ALTEL}\rvert}{\lvert \boldsymbol{SRPH}\rvert}$ ($\lvert SRPH\rvert = 6{,}693{,}298$)	**64.3%**	**56.9%**	**36.8%**	**62.2%**

5.2 Regional Coverage

The coverage of AZ Direct numbers within ECDB and CASTEM can be further analyzed by canton (i.e. 26 regions), see Figures 4 and 5. This regional analysis is done for all numbers and the subset of active numbers. Figures 4 and Figure 5 are sorted downwards by coverage of all numbers within cantonal regions; therefore, the order varies.

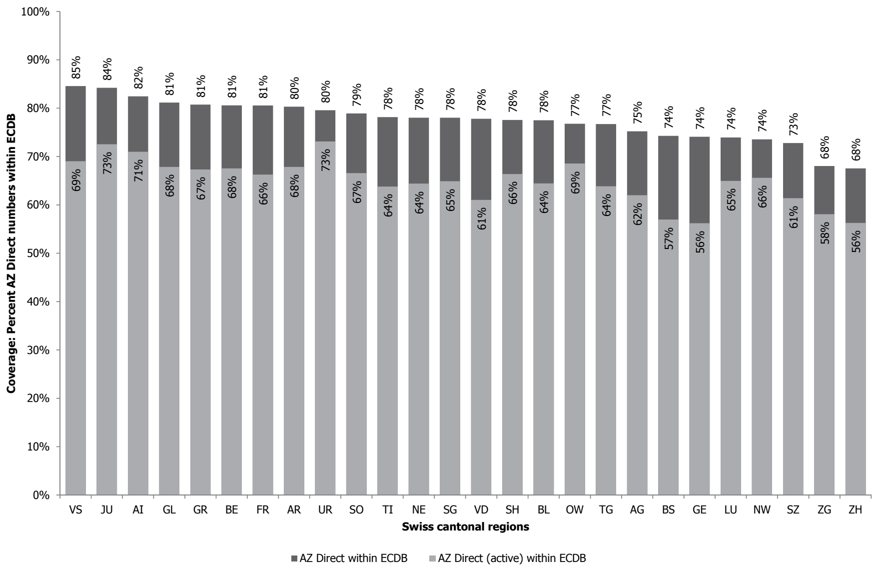

Figure 4 Coverage of AZ Direct Numbers within ECDB for all numbers and active numbers only, by canton

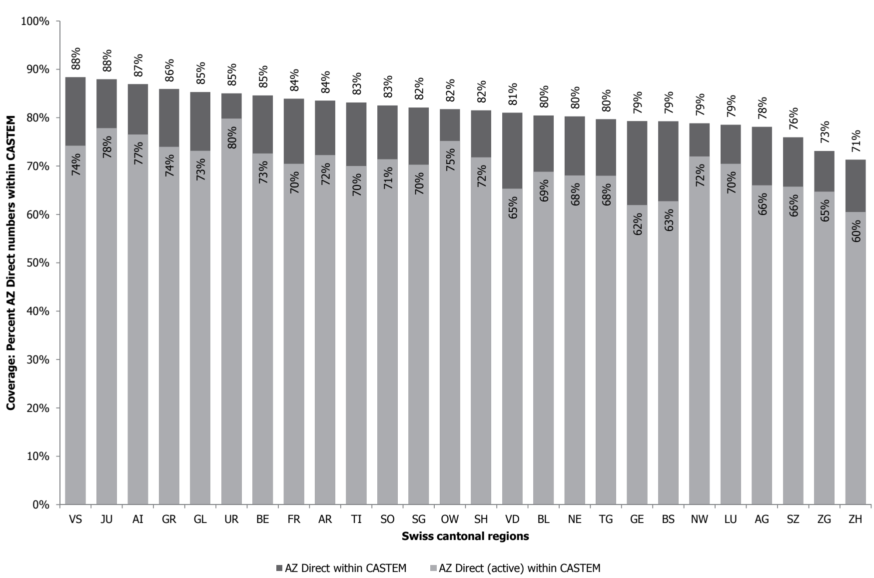

Figure 5 Coverage of AZ Direct Numbers within CASTEM for all numbers and active numbers only, by canton

It shows that coverage of AZ Direct within ECDB and CASTEM varies greatly between cantons. The coverage within CASTEM is always higher than within ECDB. Looking at all phone numbers, numbers in canton Valais (VS) and Jura (JU) are covered the most. The overall pattern for active phone numbers is similar, but on a lower level. In some cantons (Uri (UR), Nidwalden (NW), Obwalden (OW) and Lucerne (LU)), the difference between active and non-active numbers is lower (<10%) than in others.

5.3 Coverage within SRPH

When comparing AZ Direct Numbers and SRPH, the main focus is on person coverage rather than telephony coverage as before. SRPH consists of 6,693,298 persons and 3,525,438 households. As already noted in Section 3, for 27.7% of persons and 30.9% of households in SRPH, no phone number from ECDB can be matched (ALTEL), see Figure 1. This results in 4,838,986 persons (and 2,437,810 households) where a phone number can be matched (non-ALTEL).

When matching AZ Direct Numbers with SRPH persons by the assigned phone number, 88.9% of non-ALTEL persons and 88.2% of non-ALTEL households are covered. When considering the total SRPH sampling frame as the enumerator for the coverage (ALTEL and non-ALTEL), the assigned phone number of 64.3% persons (61.0% households) is part of AZ Direct Numbers. It has to be noted that this value is only approximately 8% lower than the maximum achievable value of 72.3% non-ALTEL persons. The absolute numbers are given in the Venn diagram in Figure 6.

Figure 6 Resulting subsets from match of AZ Direct Numbers with SRPH. Numbers are in thousands and add up to the total

5.4 Coverage by Regional and Demographic Characteristics

Since demographic characteristics are known for the persons in SRPH, further analyses concerning coverage characteristics can be made. As the maximum achievable coverage within SRPH is the share of non-ALTEL persons, analyses show non-ALTEL coverage within SRPH in comparison with AZ Direct's key figures.

Again, the coverage within SRPH varies regionally within cantonal regions, see Figure 7. This is also reflected in coverage differences within language regions, see Figure 8. The Italian-speaking part of Switzerland has less coverage than the German and French-speaking regions. However, the low coverage can also be a result of the high share of non-permanent resident homes in this part of Switzerland and the number of Italian-speaking people who work outside their home region. In canton Ticino (TI), the SFSO faces the same challenges in identification of phone numbers within SPRH: the ALTEL share is highest in this canton (Figure 7).

Men have a lower coverage than women: 62.5% (non-ALTEL 70.1%) compared with 66.0% (non-ALTEL 72.9%). In terms of age, those aged between 20 and 39 years have the lowest AZ Direct coverage within SRPH. For higher age groups, there is almost no gap between coverage of AZ Direct Numbers and non-ALTEL persons within SRPH. The gap between non-ALTEL persons and AZ Direct Numbers is highest for those aged between 30 and 39. Thus, existing phone numbers are particularly hard for AZ Direct to collect in this age group.

For AZ Direct, foreigners living in Switzerland (AZ coverage: 45.6%, non-ALTEL: 57.6%) are harder to collect than Swiss citizens (AZ Direct coverage: 69.8%, non-ALTEL: 75.9%). The reason might be that foreigners are not as willing to publish their phone numbers in a register. The AZ Direct coverage is particularly low for holders of permit B (a time-restricted permit) (AZ coverage: 31.9%, non-ALTEL: 51%); permit C holders (permanent permit) have an AZ Direct coverage of 52% (non-ALTEL: 63%).

Not surprisingly, persons living in single-person households have the lowest coverage within the AZ Direct Numbers (see Figure 8).

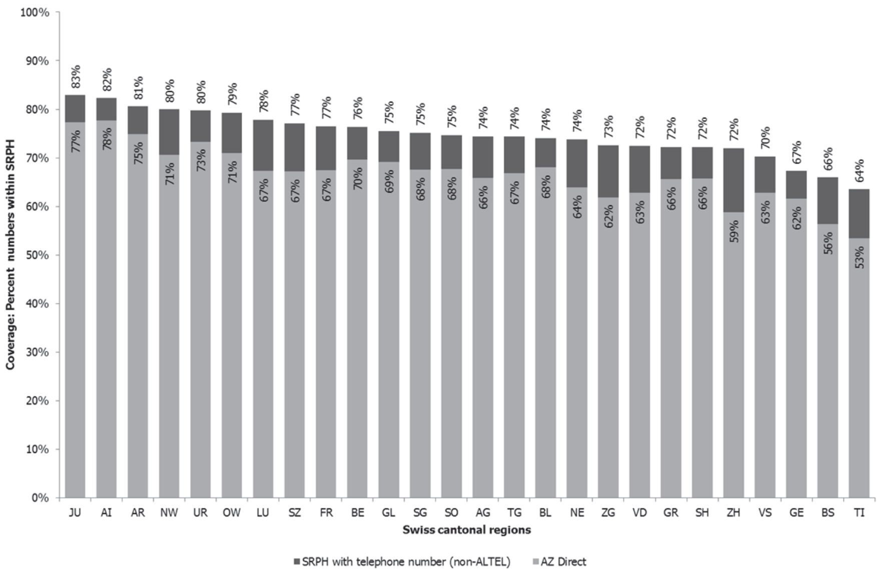

Figure 7 Coverage of AZ Direct Numbers and non-ALTEL persons within SRPH, by canton

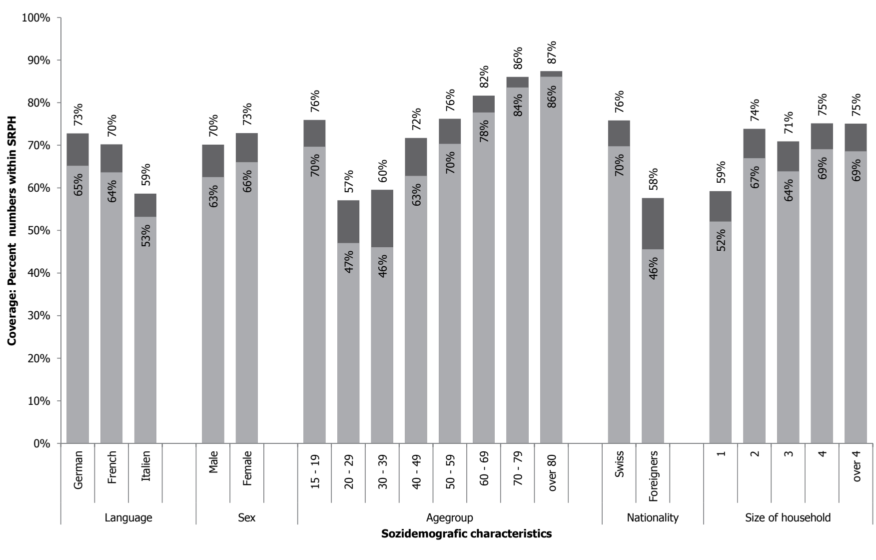

Figure 8 Coverage of AZ Direct Numbers and non-ALTEL persons within SRPH, by language region, sex, age group, nationality and size of household

6 Comparison of RDD with Different Basic Populations

The methodology used to obtain an RDD sample is described in Section 2.2. We compare RDD to CASTEM and SRPH in a similar way as AZ Direct Numbers in Sections 5.1 and 5.3, respectively. As the RDD sample in this analysis contains only private numbers, a comparison with ECDB is not discussed here.

6.1 Coverage of RDD within CASTEM

As described in Section 2.2, RDD numbers can be generated on the basis of one-digit, two-digit, three-digit or four-digit randomization. The coverage of CASTEM by the RDD sample for the different block sizes is shown in Table 3.

In theory, coverage of RDD within CASTEM has to be 100%. When creating an RDD framework, it is interesting to understand why coverage does not reach 100%. We found various explanations, all leading to the fact that valid number blocks were unknown at the time of generation of the numbers. More details are given in Section 6.2.

The coverage from two-digit randomization is 3.7% higher than for one-digit randomization. This is a significant gain in coverage, yielded by an increase of 4,365,700 phone numbers. The gain when using three-digit and four-digit randomization is not as high and many more numbers need to be generated and dialed.

The trade-off between the quantity of numbers and coverage might be best for the two-digit randomization and is also the most widely used approach in RDD sampling. For this reason, the comparison between RDD and SRPH is conducted exemplarily for the two-digit randomization in the next section.

6.2 Coverage of RDD within SRPH

In total, about 12.6 million RDD numbers are generated by the two-digit randomization, see Table 3. SRPH contains 4,838,986 persons and 2,437,810 households where a telephone number can be found in the ECDB. This was discussed in Section 5.3.

About 9.7 million telephone numbers are found in RDD only, see Figure 9. This is expected as RDD will always exceed the number of used numbers, as the approach is to capture all likely numbers by randomization. For 99.8% of non-ALTEL persons and 99.8% of non-ALTEL households, a match between the number from ECDB and the RDD sample is found. Hence, RDD provides an excellent alternative to coverage of non-ALTEL SRPH persons. In total (including ALTELs), the coverage of RDD within SRPH is 72.2% for persons and 69% for households.

Table 3 Coverage of RDD within CASTEM by different block size

Block (randomization) size	# of numbers	Coverage of RDD within CASTEM
10 (one-digit)	8,206,110	94.1%
100 (two-digit)	12,571,800	97.8%
1,000 (three-digit)	19,048,000	98.5%
10,000 (four digit)	34,550,000	98.7%

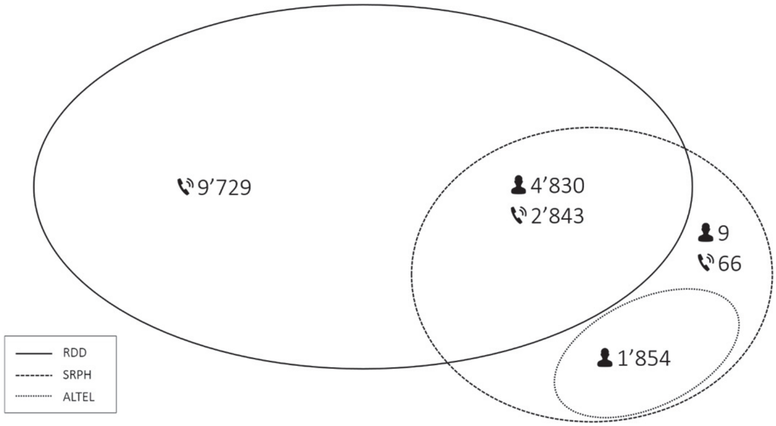

Figure 9 Resulting subsets from match of an RDD sample with SRPH. Numbers are in thousands and add up to the total

Of non-ALTEL phone numbers, no match could be found with RDD for 66,841. For 45.6% of these numbers, no entry existed in the blocks found by the Gabler & Häder method and, hence, no number was generated; i.e. this is the loss of numbers, if we use the Gabler & Häder method instead of all assigned available blocks for RDD number generation. And 54.5% of numbers were not generated because the provider was marked as a business operator (i.e. sells services to legal entities only).

Figure 10 shows the regional variability of the coverage of RDD and non-ALTEL within SRPH.

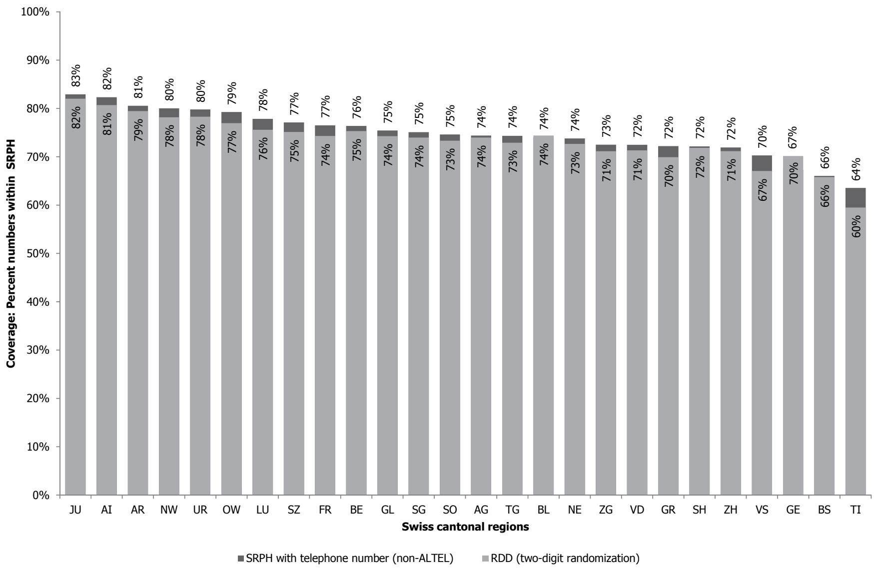

Figure 10 Coverage of RDD and non-ALTEL within SRPH, by canton

7 Supplementary Information within AZ Direct Addresses/Persons

In our last analysis, we match address items from SRPH with address items in AZ Direct to discover if the latter contains contact information by phone not in the SFSO's official phone number and person data sources. The hypothesis is that through the inclusion of not only currently activated numbers but also formerly activated (now inactive) numbers – which is unique to the AZ Direct number database – numbers can be assigned to ALTEL persons.

The question is the number of ALTEL persons for whom a landline number could be found within the AZ Direct databases using the address items (name, surname, street name, house number, postcode, place name) as matching information. In terms of an algorithm developed and tested by the SFSO, a match is defined as successful when more than 80% of the characters from the address items are identical. Usually, the SFSO conducts this matching with the data from ECDB to find numbers for SRPH persons, see Section 3. We use the same matching algorithm for the comparison between the AZ databases and SRPH.

In order to provide as much person information as possible, we use both data sources from AZ Direct, the file based on phone numbers (AZ Numbers) and the

Figure 11 Resulting subsets from match of AZ Addresses/Persons with SRPH. Numbers are in thousands and add up to the total

file based on person information (AZ Person Plus). In this sense, each identifiable person within these files is linked to a number, resulting in a file structure where the telephone number is not unique. In contrast, in ECDB each number is linked uniquely to one person, the official holder of the telephone number. The resulting file is called ‘AZ Addresses/Persons’, see Figure 2.

In total, 3.3 million identifiable persons are within AZ Direct Addresses/Persons. SRPH contains about 6.7 million persons, see Section 5.3 and Figure 11. The large difference between these two files is not surprising, as register-based person information is, by law, not accessible to private organizations such as AZ Direct.

A match between person data from AZ Direct and SRPH is successful in 308,308 persons, but no match between SPRH and ECDB can be found (i.e. ALTEL persons), see Figure 11.

Additionally, a match can be found between person data from AZ Direct and SRPH in 50,778 addresses, but where the assigned number is different. Using such additional information is an option for the SFSO in order to generate additional landline phone contact information for CATI surveys.

A total match is found for 2,119,028 persons; i.e. 31.7% of SRPH persons. About 1,216,796 persons are in AZ Direct Addresses/Persons only and cannot be found in SRPH through use of SFSO’s matching algorithm. This might be due to incorrect, incomplete or obsolete address parameters, or due to the manner in which the algorithm works. Because of the nature of this personalized information, further investigation is not possible due to data protection laws.

The analysis of the matched addresses shows only little variation in the coverage for cantons and language region. However, significant variations exist in the age group and the variable household size. One-person households are covered best by AZ Direct sources. It is not surprising that non-adult persons and non-heads of individual households cannot be found in non-official data sources. Even though AZ Direct contains some person information, it must be seen as a household or telephony register when used for population sampling.

8 Some Remarks on Potential Generalization

As noted above, Switzerland's unique legal setting facilitates the quantitative, precise calculation of coverage results, as reported in Sections 5 to 7. To our knowledge, such detailed analyses are not possible in other countries due to the lack of access to the phone number and person universe. However, a majority of the obtained results may be generalized and used as benchmarks for other countries when certain conditions apply. Depending on the fulfillment of these conditions, the calculated coverage values may not be useful as point estimates for other countries, but they could be used as upper or lower coverage levels in the country of interest.

In order to discuss the potential for generalization of the obtained results, some considerations on the datasets used in the preceding analyses must be taken into account. These are summarized in Table 4. As a general result, it can be derived that the quality and size of phone number collections (PNCs) and RDD samples depend on the percentage of listed phone numbers available to commercial providers and (for RDD samples only) the availability of published number blocks.

In order to investigate the percentage of listed phone numbers in other countries, we can look at some estimates reported by Heckel and Wiese (2012). They compared the total number of listed (published) private phone numbers with the total number of households in Germany, Italy, the UK, France and Spain, and calculated a percentage of listed phone numbers ranging from 53% to 69%. Hence, the percentage of listed phone numbers in other European countries may be lower than in Switzerland.

Sand (2014) investigated the impact of official sources of assigned number blocks for the GESIS RDD sampling frame for Germany. Depending on the availability, quality and completeness of such sources within other countries, they may or may not be used to generate RDD numbers. An overview for some European countries can be found in Heckel and Wiese (2012).

As mentioned in Table 4, census data are available for a multitude of other countries, but the quality may differ greatly. In general, the census systems can be classified as traditional, register-based, register combined with other sources, and rolling censuses (Valente, 2010). The Swiss census belongs to the combined census type, which is also used in Italy, Germany, Spain and other central European countries. Austria and Scandinavian countries use solely register-based census systems. France is the only country to use a rolling census, whereas the UK, Portugal and most Eastern European countries use traditional census systems. To our knowledge, the use of different sampling frames within specific census types and countries is not documented and cannot be evaluated here.

Based on these preceding remarks, a discussion of the potential concerning the generalization of the results in Sections 5 to 7 in relation to other countries can be found in Table 5.

In order to assess the validity of the results concerning the unconditional person coverage (Sections 5.3 and 6.2) for other countries, the relative landline penetration must be taken into account. For a majority of European countries, this quantity is reported in Heckel and Wiese (2012, p. 111). The landline penetration in Switzerland is about 92% (Stähli, 2012).

Table 4 Remarks concerning the various datasets discussed in the preceding sections

Dataset(s)	Remarks
PNCs	▪ The conditions for commercial providers to collect (landline) phone numbers in Switzerland do not differ from conditions in other countries. The percentage of listed phone numbers may have an influence on the success of data collectors. ▪ Legal conditions and data protection laws to generate and maintain such data collections vary from country to country.
RDD samples	▪ The amount of numbers reached by using the Gabler & Häder method for number generation depends on the quality of the phone number list used as the basis for number generation (i.e. the percentage of listed phone numbers). ▪ The amount of numbers reached by using the BIK method (published number blocks) depends on the number and completeness of the published number blocks.
ECDB/ CASTEM	▪ To our knowledge, access to a complete database of published and unpublished numbers across all telephony providers for official statistics is unique to Switzerland.
SRPH	▪ Census data are available for a multitude of other countries, but the underlying census systems differ.

Table 5 Summary of discussion concerning the generalization of results in Sections 5 to 7 in relation to other countries

Section	Datasets	Remarks/discussion
5.1 & 5.2	AZ Direct Numbers vs. ECDB/CASTEM	▪ Given a similar percentage of listed phone numbers and an equivalent collecting effort/method of the commercial provider, the calculated coverage of all AZ Direct Numbers can be taken as a general benchmark for other countries. ▪ If a country has a lower percentage of listed phone numbers or the data of the commercial provider have a lower quality, the reported coverage can be seen as a maximum level. ▪ According to the results in Section 5.2, it can be taken as a general result that regional variability in coverage is present.
5.3	AZ Direct Numbers vs. SRPH	▪ In addition to the generalization conditions mentioned above, this coverage depends on the landline penetration in the country of interest. ▪ If a country has a lower landline penetration than Switzerland, the reported coverage can be seen as a maximum lower level for unconditional coverage (note that the reported coverage depends on the quality of the matching algorithm between phone numbers and registry data). ▪ The results concerning variations in coverage related to regional and demographic characteristics can be generalized at least qualitatively. For example, the finding that coverage for 30 to 39-year-olds is lowest is, in our opinion, also valid for other countries.
6.1	RDD (Häder & Gabler method) vs. CASTEM	▪ Given a similar basis for number generation as in Switzerland, the reported coverage can be taken as a general benchmark for other countries. ▪ The result that a two-digit randomization offers the best trade-off between quantity of numbers and coverage is a general result that does not depend solely on Swiss conditions.

Section	Datasets	Remarks/discussion
6.2	RDD (Häder & Gabler method) vs. SRPH	▪ In addition to the generalization conditions mentioned above, this coverage depends on the landline penetration in the country of interest. ▪ If a country has a lower landline penetration, the reported coverage can be seen as a maximum lower level for the unconditional coverage (note that the reported coverage depends on the quality of the matching algorithm between phone numbers and registry data).
7	AZ Direct Addresses/ Persons vs. SRPH	▪ As already noted in Section 2.1, we assume that certain selection criteria apply to persons in the AZ Person Plus file. Hence, we do not recommend use of the results reported in Section 7 as a benchmark for other countries or providers of data collections.

9 Conclusions

The purpose of this paper is to calculate reliable measures of coverage of alternative telephone sampling frames; i.e. commercially available alternatives to the databases available to the SFSO (ECDB/CASTEM and SRPH). The examples we use are a landline phone number collection offered by AZ Direct and RDD samples generated by BIK Aschpurwis + Behrens. The intent is not to evaluate these sources in terms of 'can be applied' or 'cannot be applied', as such a decision depends on the content, the purpose of the survey, the survey budget and other restrictions, and is finally the researchers' choice. This paper also does not include a comprehensive comparison of other methods. We assume that few options exist as far as commercial landline phone number databases are concerned. Open sources, such as internet telephone directories, cannot be used for sampling since the underlying lists or databases cannot be accessed and, therefore, randomized sample drawing is not possible.

Among the key findings here is that the exemplarily analyzed AZ Direct Numbers collection covers the population with a rate of approximately 85% concerning the telephony universe (with CASTEM as the reference population) and 64% concerning the person universe (with SRPH as the reference population). Looking at non-ALTEL persons only, the coverage within SRPH is 89%. Non-coverage is influenced by age, sex, household size and region. It must be noted that in AZ Direct Numbers, entries are missing mainly for the 20-39 age group. Additionally, the share of ALTEL within SRPH is above average within this age group. Hence, there is a two-fold gap for this age group, which may lead to substantial bias in survey results.

When using RDD, the two-digit randomization provides the best trade-off between the quantity of generated numbers and coverage (97.8% coverage within CASTEM). In a comparison of RDD and the non-ALTEL persons, a match is found for 99.8% of persons. However, RDD has some drawbacks: an advantage of AZ Direct or SRPH over RDD samples is that households can be addressed by post before the survey starts, leading to higher response rates and, therefore, a trade-off between coverage and response rates. Also, RDD samples need predictive-dialing if research budgets are restricted.

Other non-telephony sampling approaches can be used if the risk of non-coverage bias within telephony samples appears too high for a given research target. It is clear, though, as shown in this paper, that telephone surveys still have a high measurable coverage. It can be concluded from the analyses in Sections 5 and 6 that both commercially available sources are robust sampling frames for representative studies. The choice between these two depends on researchers' risk evaluation of non-coverage against non-response and the intended study design. If no postal information is required, RDD sampling will be the preferred solution.

As discussed in Section 1, representativeness is not a decision between true or false, but studies can be representative up to a certain level. The risks and implications of a (slight) lack of representativeness can be included when the results are published.

Comparison of the AZ Direct Addresses/Persons and the SRPH addresses shows that for 308,308 ALTEL persons, a match is found within the commercially available AZ Direct database, but a further analysis of the validity and usability of this information should be considered.

Except for the analysis in Section 7, the obtained results for Switzerland can be generalized to other countries, taking into account key figures on the percentage of listed phone numbers, the availability of published number blocks (RDD only) and the landline penetration in the country of interest.

The effect on survey estimates when excluding parts of the population in telephone surveys (i.e. the coverage bias) remains an important concern among survey researchers (Massey, 1988). In future, the use of mobile numbers is essential. In particular, the two-fold gap in coverage for the 20-39 age group could be closed by the inclusion of mobile phone numbers in telephone samples. We strongly believe that as a solution dual-frame samples, including RDD mobile numbers, can bring the desired effect to high quality samples by closing coverage and overcoming the non-coverage issues shown and discussed in this paper.

Literature

Busse, B., & Fuchs, M. (2012). The Components of Landline Telephone Survey Coverage Bias. The Relative Importance of No-Phone and Mobile-Only Populations. *Quality and Quantity,* 4, 1209-1225.

Diekmann, A., & Bruderer Enzler, H. (2013). Risikosurvey 2013. Risikowahrnehmung der Schweizerinnen und Schweizer. Retrieved March 30, 2015, from ETH Zürich, Professur für Soziologie website: http://www.socio.ethz.ch/forschung/risikostudie.html

Gabler, S., & Häder, S. (2007A). Überlegungen zu einem Stichprobendesign für Telefonumfragen in Deutschland. *ZUMA-Nachrichten*, 41, 7-18.

Gabler, S., & Häder, S. (2007B). Haushalts- und Personenerhebungen – Machbarkeit von Random Digit Dialing in der Schweiz. *Methodenbericht BFS*, ISBN: 978-3-303-00378-7.

Heckel, C., & Wiese, K. (2012). Sampling Frames for Telephone Surveys in Europe. In S. Häder et al (Eds.), *Telephone Surveys in Europe* (pp. 103-119). Heidelberg: Springer.

Klug, S., Arn, B., & Müller, M. (2014). Mobile Erstkontakte in Dual-frame Stichproben aus RDD Mobile und SRPH. Schweizer Tag der öffentlichen Statistik. Bundesamt für Statistik.

Lipps, O. & Pekari, N., & Roberts, C. (2013). Coverage and nonresponse errors in an individual register frame based Swiss telephone election study. *FORS Working Papers 2013-2*.

Lipps, O. & Pekari, N., & Roberts, C. (2015). Undercoverage and Nonresponse in a List-sampled Telephone Election Survey. *Survey Research Method*s, 9 (2), 71-82.

Massey, J. T. (1988). An overview of telephone coverage. In R. M. Groves et al. (Eds.). *Telephone survey methodology* (pp. 3-8). New York: Wiley.

Number Blocks (2016). Website: http://www.eofcom.admin.ch/eofcom/public/searchEofcom_e164Allocated.do

O'Toole, J., Sinclair, M., & Leder, K. (2008). Maximising response rates in household telephone surveys. *BMC Medical Research Methodology*, 8, 71.

Sand, M. (2014). Überarbeitung des GESIS Auswahlrahmens für Telefonstichproben: Führt die Anreicherung durch die Angaben der Bundesnetzagentur zu einer Verbesserung der Auswahlgrundlage. *Dresdner Beiträge zur Soziologie*, 5, 13-38.

Schouten, B., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 101-113.

SRPH (in German: Stichprobenrahmen für Personen- und Haushaltserhebungen) (2016). Website: http://www.bfs.admin.ch/bfs/portal/de/index/news/00/08.html

Stähli, M. E. (2012). Switzerland. In S. Häder et al. (Eds.), *Telephone Surveys in Europe* (pp. 25-26). Heidelberg: Springer.

Valente, P. (2010). Census taking in Europe: how are populations counted in 2010? *Population & Societies*, 467, 1-4.