Task force on quality of BCS data

Analysis of sampling frames

Appropriateness and comprehensiveness of sampling frames, theoretical considerations, empirical evidence on links with data volatility and bias;

August 2013

Krzysztof Puszczak
Agnieszka Fronczyk
Marek Urbański
Sofia Pashova

Table of Contents

1. Introduction

1.a Definitions

Generally sampling frame is a list of all members of a population used as a basis for survey design [1]. There are two types of frames: list frames and area frames. A list frame is a list of all the units in the survey target population (e.g. administrative lists, the electoral roll). An area frame is a complete and exhaustive list of non-overlapping geographic areas. Area frames are less expensive and complex than a list frame. They are also used when no list frame exists and it would be too expensive or complex to create.

A reliable sampling frame should meet the following requirements:

- Well- organized in a logical, systematic fashion
- all units should be accessible – it should contain sufficient information to uniquely identify and contact each unit and other relevant information
- 'up-to-date'
- every element of the population of interest is present in the frame
- every element of the population is present only once in the frame
- no elements from outside the population of interest are present in the frame

The most popular examples of sampling frame in social research are a population register (census) and a telephone directory or random digit dialing. Traditionally the vast majority of national, general-purpose surveys used the most recent population census. Nowadays the sampling frames with telephone numbers have become the most popular type of sampling frame.
As the sampling frame provides the means of accessing the population to obtain a sample, consideration should be given to the quality of the sampling frame. Sampling frames should be evaluated early in the planning stage because a poor-quality frame will bias the final results of the study.

1.b Common problems with sampling frames

Unfortunately in the research practice it is often hard to find an adequate sampling frame that meets the above the expectations. The most common sampling frame errors are as follows:

- Coverage errors : Coverage errors arise from failure to cover adequately all components of the population being studied. Incomplete sampling frames often result in coverage errors.
- Out-of-scope units: Out-of-scope units are units that should not be included in the sampling frame because they do not belong to the target population in the reference period. If included, they cause over-coverage.
- Incompleteness – e.g. Inaccuracy of contact information (address, telephone number)

The sampling frame errors can lead to sampling bias which is outside the scope of statistical theory and impossible to estimate, but in some cases they can strongly bias the final result of the study. It is recommended to identify and minimize any sampling frame errors, however, it is completely impossible to avoid them in practice. Unfortunately all sampling frames used in social research practice are somewhat biased: some are more accurate than others, therefore the sampling error can be minimized by carefully selecting the best available sampling frame. The main causes of sampling frame errors are as follows:

- obsolete information included in sampling frame
- lack of contact information – some units are hard to find
- inappropriate sampling frame - inconsistency between the survey target population and sampling frame population
- the constrains of survey budget - high costs of the adequate sampling frame
- time pressure

The possible solutions to sampling frame problems are as follows:

- increase the sample size at the selection stage and adjust the weights at the estimation stage.
- try to update the frame.
- look for an alternative sampling frame
- combine the current sampling frame with related frames to improve the coverage of the target population.

1.c Telephone-based sampling frames – drawbacks and benefits

Due to huge popularity of telephone surveys in social research nowadays, the special problems with telephone-based sampling frame and the possible solution are investigated in this section.

Traditionally, telephone-based sampling frames were considered to be biased because of the low level of telephone ownership. But nowadays, due to the high level of individual telephone ownership the situation has changed . Additionally the relatively low price of telephone surveys and time-efficiency contributed to their huge popularity.

Although the telephone surveys are the most common and generally accepted method of interviewing now, the quality of telephone-based sampling frames is often inadequate.

- Non-telephone households / respondents

One of the major sources of bias is the exclusion of non-telephone households. It was found that people living in rural areas and those on low incomes were less likely to own telephones than those living in urban areas with higher incomes.

- Mobile only or landline only households /respondents

All over the world it is observed that the rate of mobile only households or respondents increases systematically. The mobile only respondents are distinctly different to respondents with a landline connection only, and the increase in the number of mobile-only people is not uniform across all groups in the community.

- Duplication

On the other hand some respondents are better accessible because of ownership of two or more telephone number (mobile, landline, at home, in the office etc.) and as a consequence the probability of selection of the units is unequal. In this case the partial solution is usage of the adequate weights at the estimation stage e.g. taking into account the number of telephones per respondent.

- Non-household target population

Apart from that the significant differences between sampling frame population and survey target population occurred in the case when the sampling frame population are households and the survey target population are individuals. In this case the partial solution is usage of the adequate weights at the estimation stage.

- Out-of-date

Most problems mentioned above also occur in census-based frames.

1.d Sampling frames used in the DG ECFIN Consumer Survey

In this section the sampling frames used by the institutes participating in the DG ECFIN Consumer Survey are described in more detail and the potential consequences of sampling frame errors for sample error and final results are presented.

The analysis is based on the excel file (*Metadata_checked_by_partners - Consumers.xlsx*) that contains information about the sample design delivered by the institutes. Due to some inaccuracy in this file the presented results should be treated as provisional and need to be supplemented after verification of source data.

Generally majority of institutes use a different kind of telephone directory as a sampling frame (see table below). The diversity of used telephone directories is large therefore in the future it is recommended to investigate their main characteristics. At the current stage it is supposed that different type of telephone directories used in the study can lead to specific bias in the final results of the study.

	Number of countries
Census-based sampling frames	9 HR, DK, FI, DE, LV, LT, PL, PT , SK
Telephone-based sampling frames	17 AT, BE, CY, CZ, EE, EL, FR,IT, LU, MT, NL, SI, ES, SE, TR, UK, IE
No Sampling Frame	3 BG, RO, HU,
Polling Stations Territory	2 ME, MK

Table 1.1. Sampling frames used in Consumer Survey

The second type of sampling frames used by institutes are the official population registers. Nevertheless in this case institutes often interviewed the respondents by phone (CATI) therefore it is possible that some of the registers could be somehow biased in the same way as the telephone directories (see table below).

	Face-to-face	CATI	Unspecified	Total
Telephone directory	1	16	1	18
Census	5	4	1	10
polling stations territory	0	2	0	2
Unspecified	0	1	0	1
Total	6	23	2	31

Table 1 2. Frequency of telephone interviewing in Consumer Survey

Summing up the vast majority of institutes conducted the interviews via telephone, even if they drew the sample from the population register. As it is shown in the previous section the sampling frame error in the case of telephone interviewing sample is potentially large and can bias the result of the study in indeterminate way. For estimation of the bias size in Consumer Survey the further investigations are recommended. The potential directions of the inquiries are suggested below. Due to the inaccuracy of available information it is impossible to draw the final conclusion at the current stage.

The first possible direction of investigation is the usage of mobile phone numbers. Only two institutes have declared openly that their telephone directory includes mobile phones as well (see table below). In most cases the information was unspecified.

	Frequency
Doesn't apply	8
fix only	3
mobile & fix	8
unspecified	12
Total	31

Table 1.3 Mobile and landline telephone numbers used in sampling frames in Consumer Survey

Next possible direction of investigation are the differences between sampling frame population and survey target population in the case when the sampling frame population are households and the survey target population are individuals. In Consumer Survey the institutes quite often used household sampling frames. However, this type of sampling frame errors can be partly minimized by adequate sample design and weighting procedure.

	Frequency
Individuals sampling frames	7
Household sampling frames	10
unspecified	14
Total	31

Table 1.4 Household sampling frames used in Consumer Survey

Finally, the frequency of sampling frame update and the sampling frame coverage need to be investigated. At the current stage the sampling frame coverage is rather misestimated because of the incomplete information about sampling frame size and adequate population size. For household-based sampling frames an adequate population size is the household population size. The available data present only target population size therefore the missing information was partly supplemented at the analysis stage in the following section. However, the supplement needs to be verified.

As far as the sampling frame update is concerned the detailed analysis of this characteristic is presented in the following sections.

2. Quality measures vs. descriptive features of sampling frames

In this section the influence of sampling frames used by the institutes participated in the DG ECFIN Consumer Survey on quality measures is examined. Looking into quality measures, their distribution shows fairly considerable variability (Charts 2.1 – 2.3). This observation lets us assume that there are potential drivers of quality measures in the analysed data.

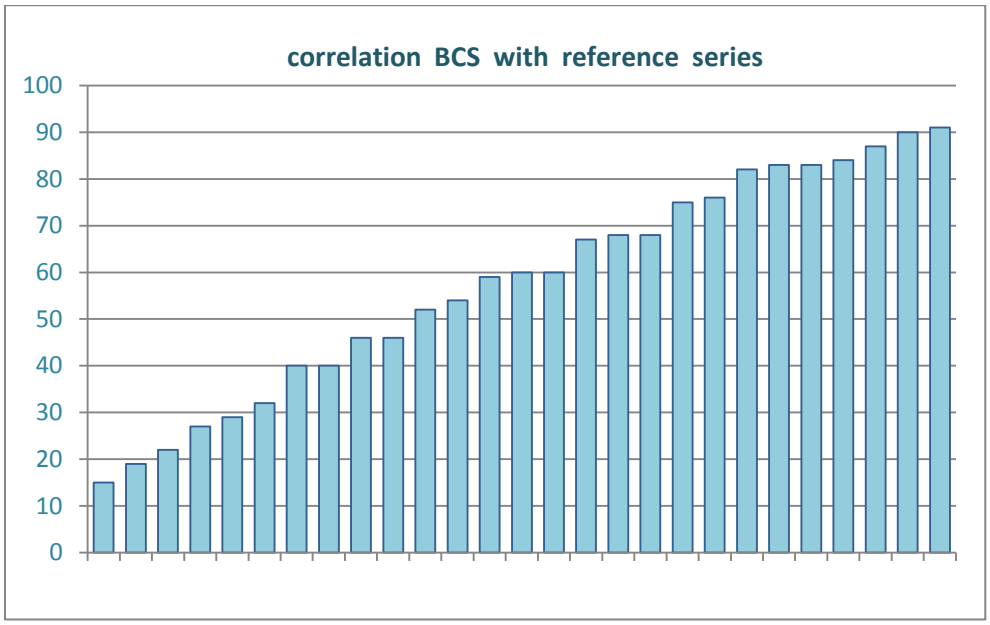

Chart 2.1. distribution of correlation BCS with reference series, n=27

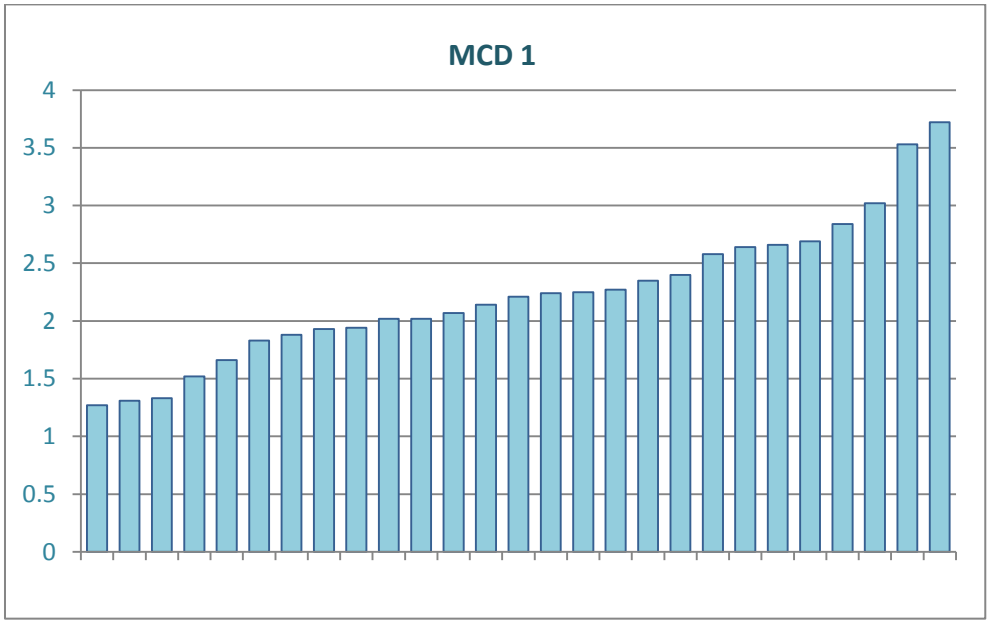

Chart 2.2. distribution of MCD 1, n=27

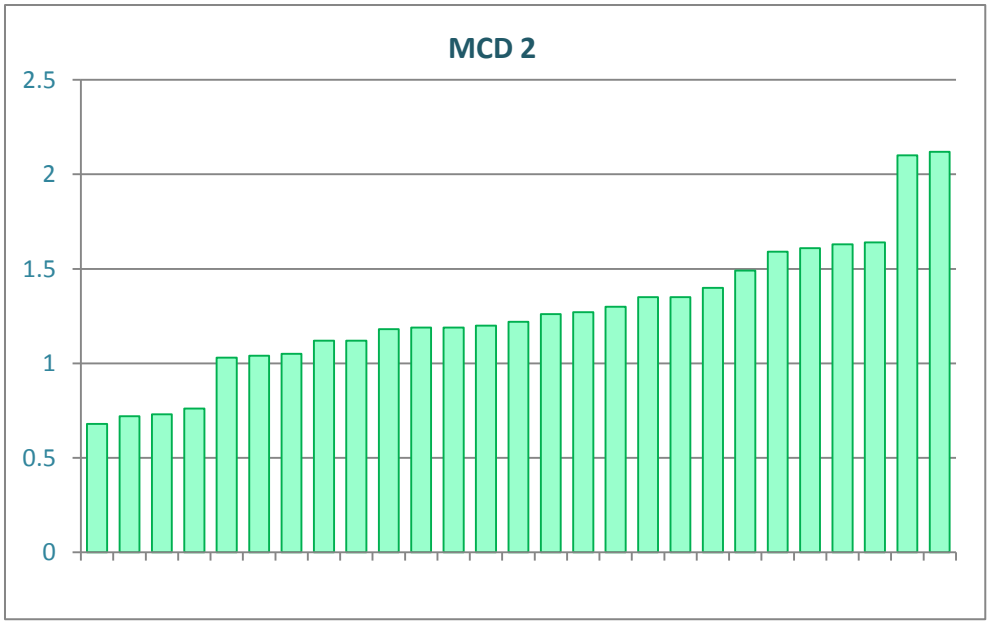

Chart 2.3. distribution of MCD 2, n=27

Having the general view on various sampling frames in customer surveys of BSC project, some variables were selected to verify their possible impact on volatility of the data (volatility) and the tracking performance with respect to statistical reference series (consistency). Based on qualitative analysis of original description of SFs the following features were considered:

- Telephone access as a unit of SF.
- Individuals as units of SF.
- Update frequency, i.e. how often a frame list is updated.
- SF coverage

As it was mentioned in part 1, interpretation of given information is not always straightforward. Especially direct descriptions of SFs in open-ended form leave space for some subjective interpretations or speculations. Incomplete information in some cases does not support analyses either. Regarding all these issues one should be aware of possible misinterpretation of some data. All further presented findings are aimed at discovery of hypothetical causes of measurement quality, i.e. volatility and consistency in BCS results.

2.a Type of units in sampling frames - analysis of impact

Telephone access as a unit of SF. There are 3 levels of this variable:

- SF units are not telephone numbers
- SF units are fixed telephone numbers
- SF units are both fixed and mobile telephone numbers

Below there some charts showing level of quality measures in different SF

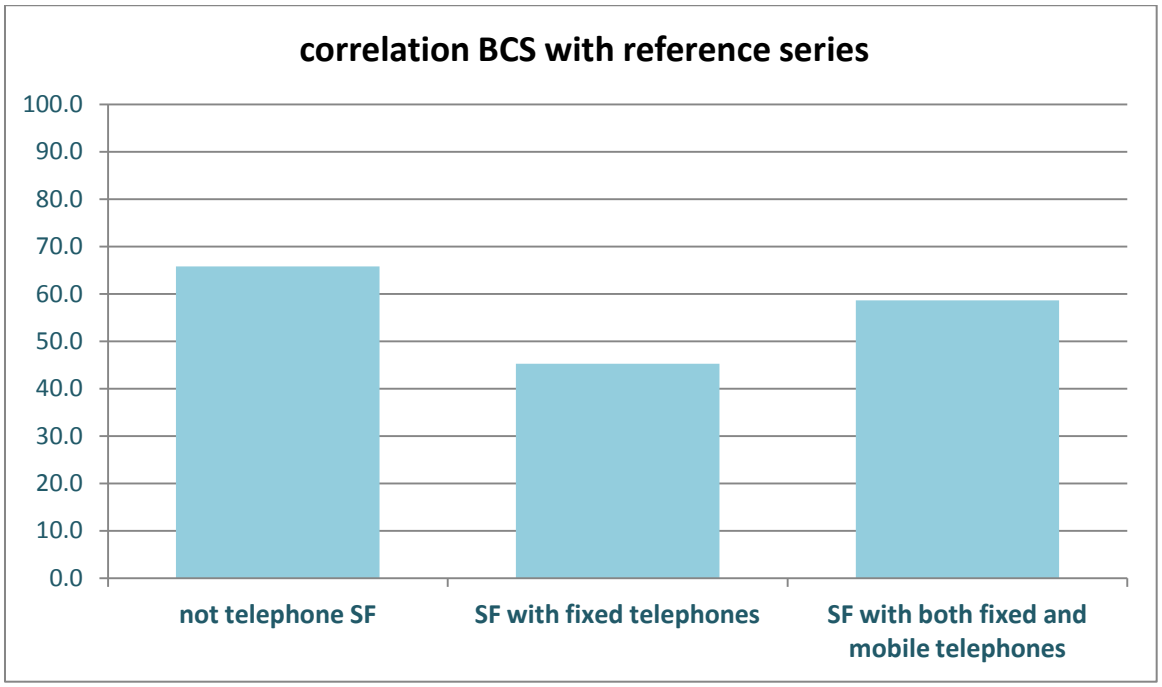

Chart 2.4. mean correlation BCS with reference series in 3 groups of SFs, n=27

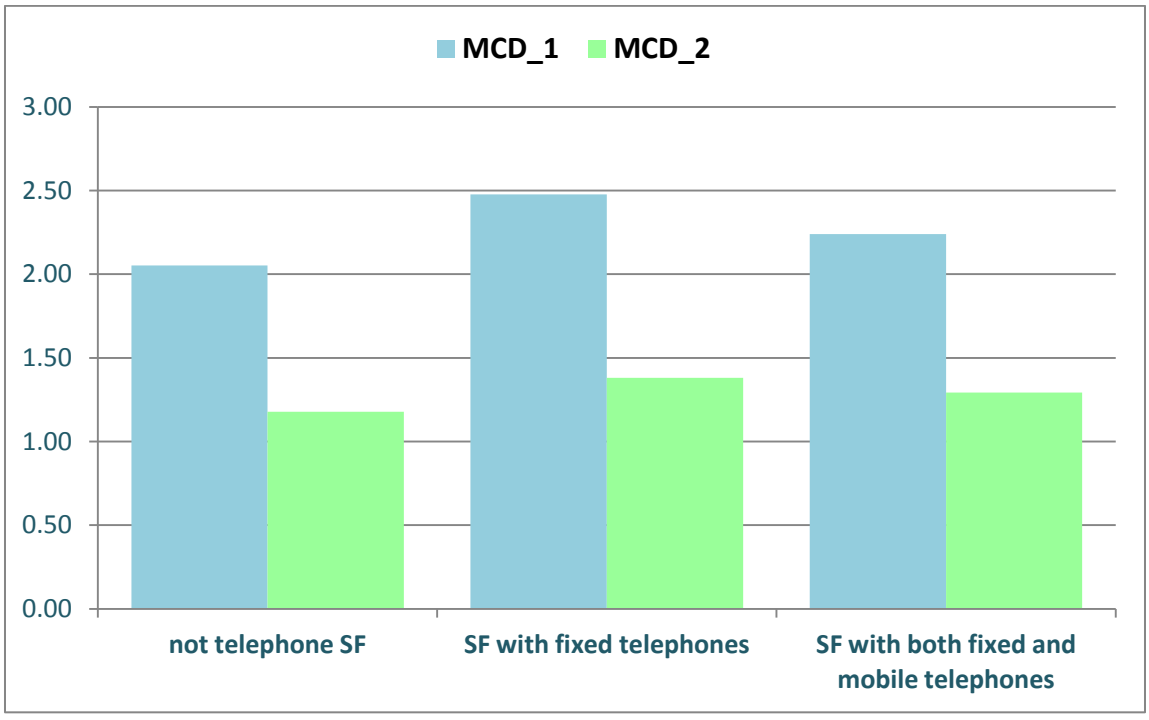

Chart 2.5. means for MCD_1 and MCD_2 in 3 groups of SFs, n=27

Individuals as units of SF. There are 2 levels of this variable:

- SF units are not individuals
- SF units are individuals

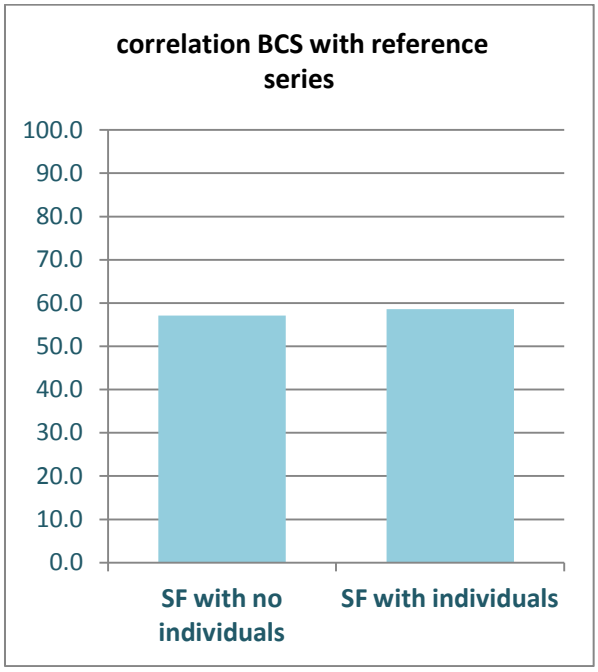

Chart 2.6. mean correlation BCS with reference series in 2 groups of SFs, n=27

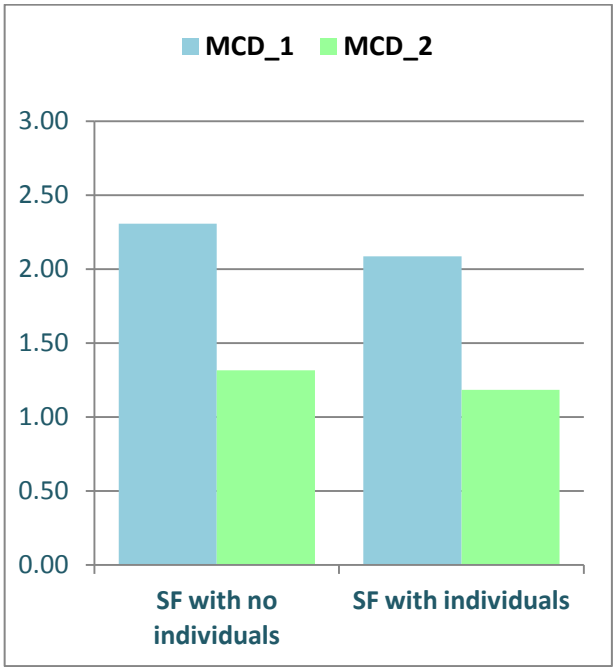

Chart 2.7. means for MCD_1 and MCD_2 in 2 groups of SFs, n=27

To consolidate both variables, two major types of SFs, was defined:

- SF with telephones
- SF with individuals

In case of 6 countries SF cannot be defined. In case of 2 countries the given information can classify SF as both telephones and individuals. Both cases were a priori classified as individuals.

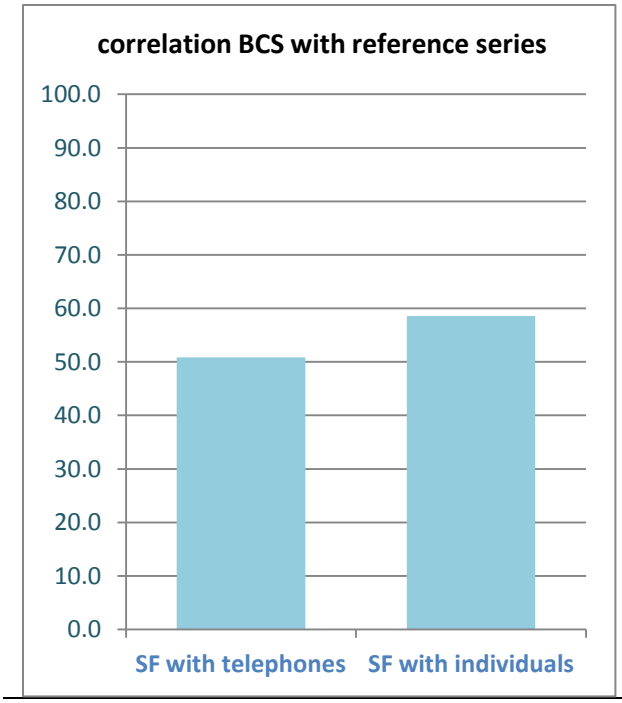

Chart 2.8. mean correlation BCS with reference series in 2 groups of SFs, n=23

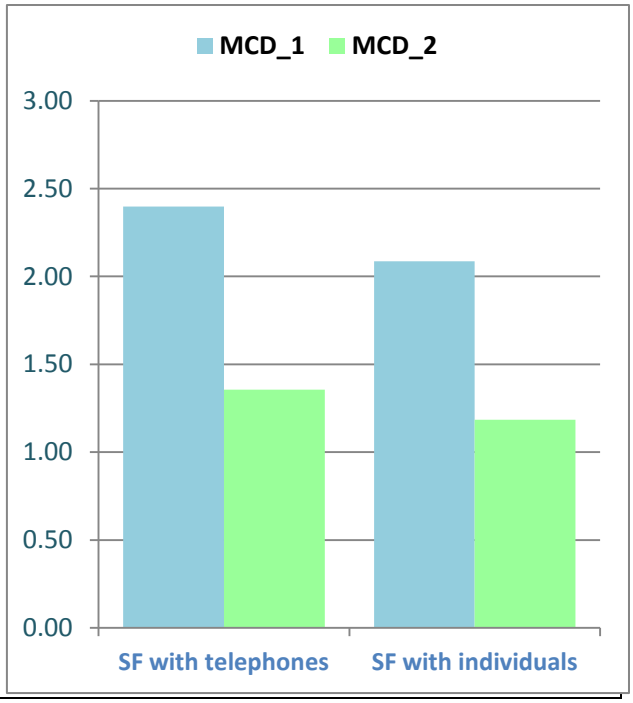

Chart 2.9. means for MCD_1 and MCD_2 in 2 groups of SFs, n=23

As analysed data do not meet criteria of quantitative measurement all findings are recommended to be interpreted as working hypothesis. Nevertheless some conclusions appear to be drown. The general tendency is that all quality measures (volatility and consistency) are coherent. As far as type of units in SF are concerned the best quality parameters are observed in countries with SF based on individuals. In case of countries where SF are based on telephone access, a slightly better quality parameters are connected with SF including both fixed telephones and mobiles in contrary to SF with fixed telephone access only.

2.b Update frequency of sampling frames - analysis of impact

Update frequency, i.e. how often frame list is updated, has 7 categories as below:

- 0.003 – daily/ continuous update
- 0.08 – monthly update
- 0.25 – quarterly update
- 1.0 – annual update
- 2.0 – update every two years
- 3.0 – update every three years
- 10.0 – update every ten years

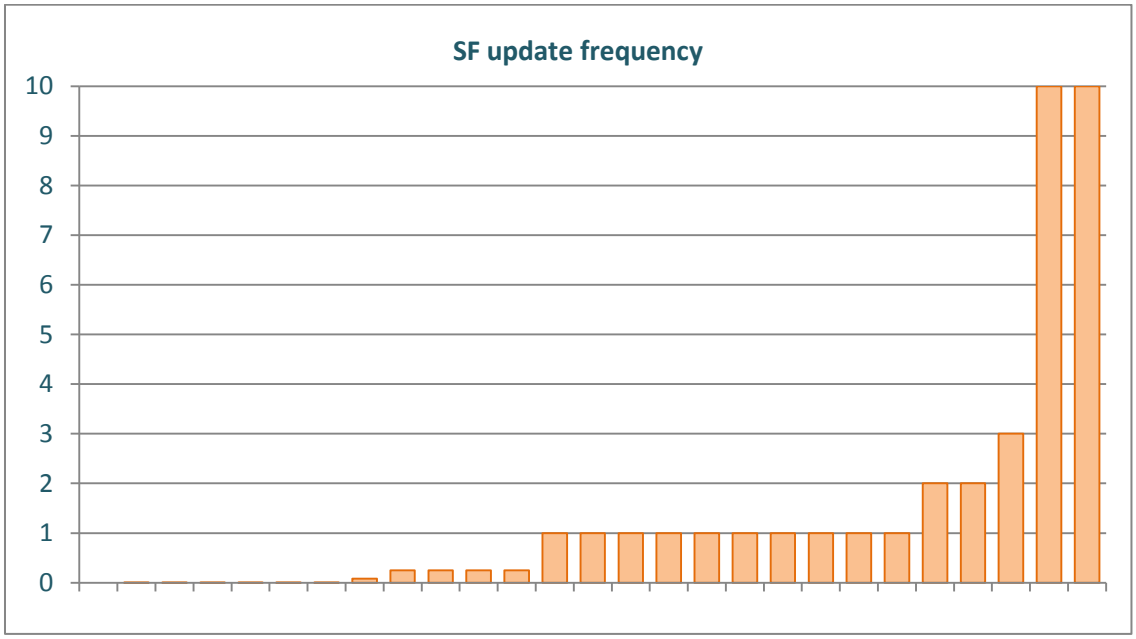

Chart 2.10. distribution of SF update frequency (in years), n=27

In order to discover any visible pattern of possible relations three scatter plots were prepared (Chart 2.11 – 2.13):

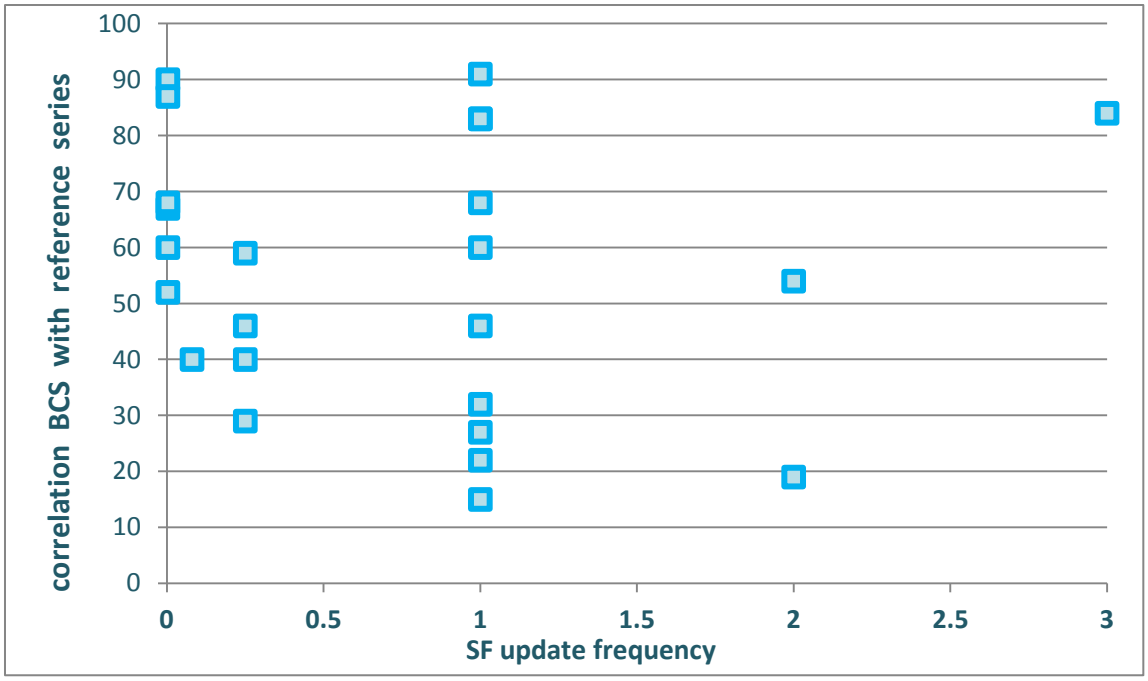

Chart 2.11.scatter plot for SF update frequency (in years) and correlation BCS with reference series , n=25. On X-axis value 10 (“every 10 years”) is not shown.

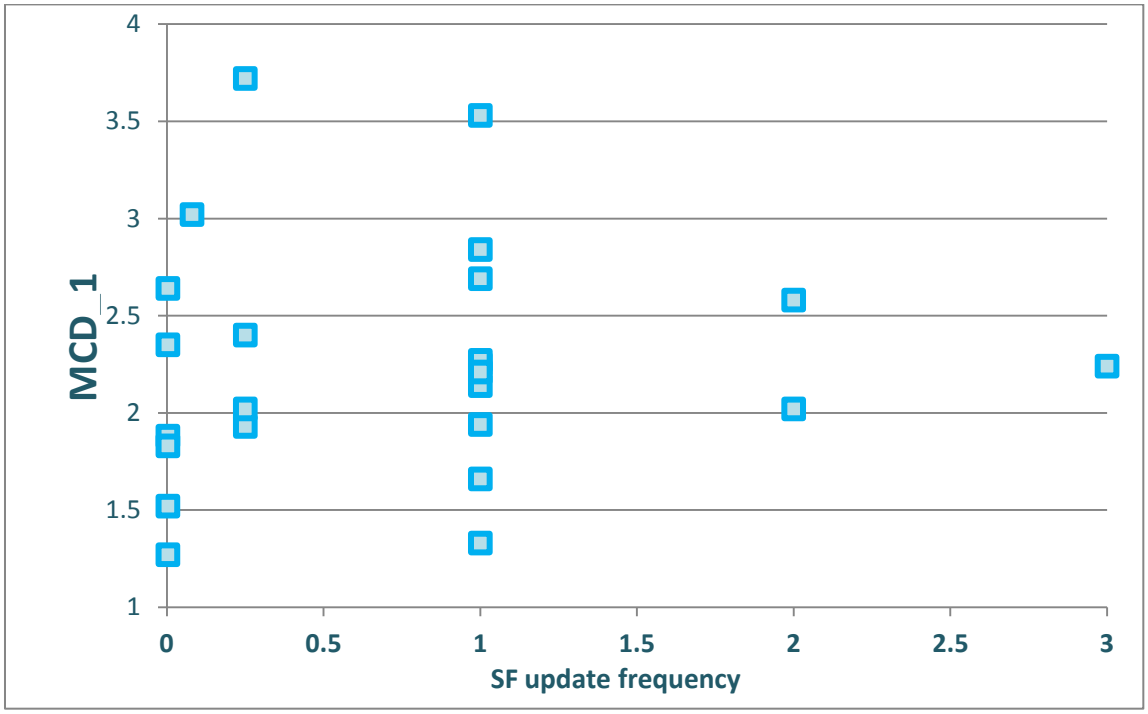

Chart 2.12.scatter plot for SF update frequency (in years) and MCD_1 , n=25. On X axis value 10 (“every 10 years”) is not shown.

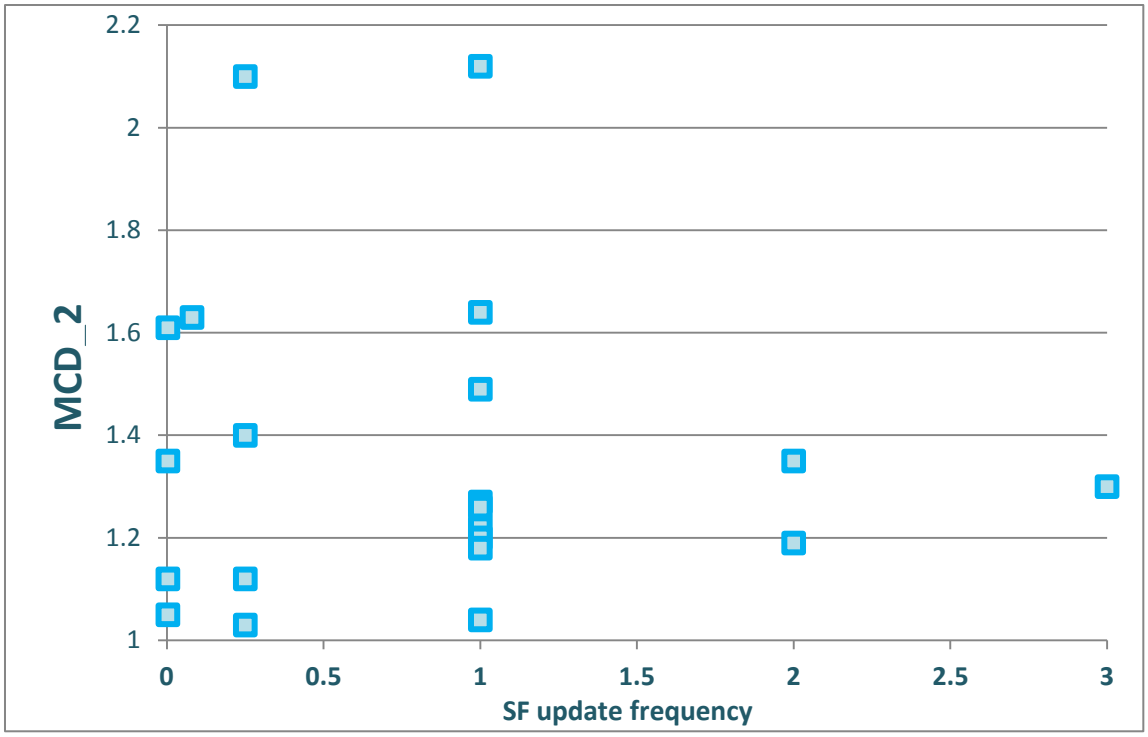

Chart 2.13.scatter plot for SF update frequency (in years) and MCD_2 , n=25. On X axis value 10 ("every 10 years") is not shown

It is quite difficult to find any rule that could shed light on character of relations between update frequency and quality parameters. That was the reason why other charts are presented. Another perspective shows how average quality parameters change along with joining SFs with smaller frequency (Chart 2.14 - 2.15).

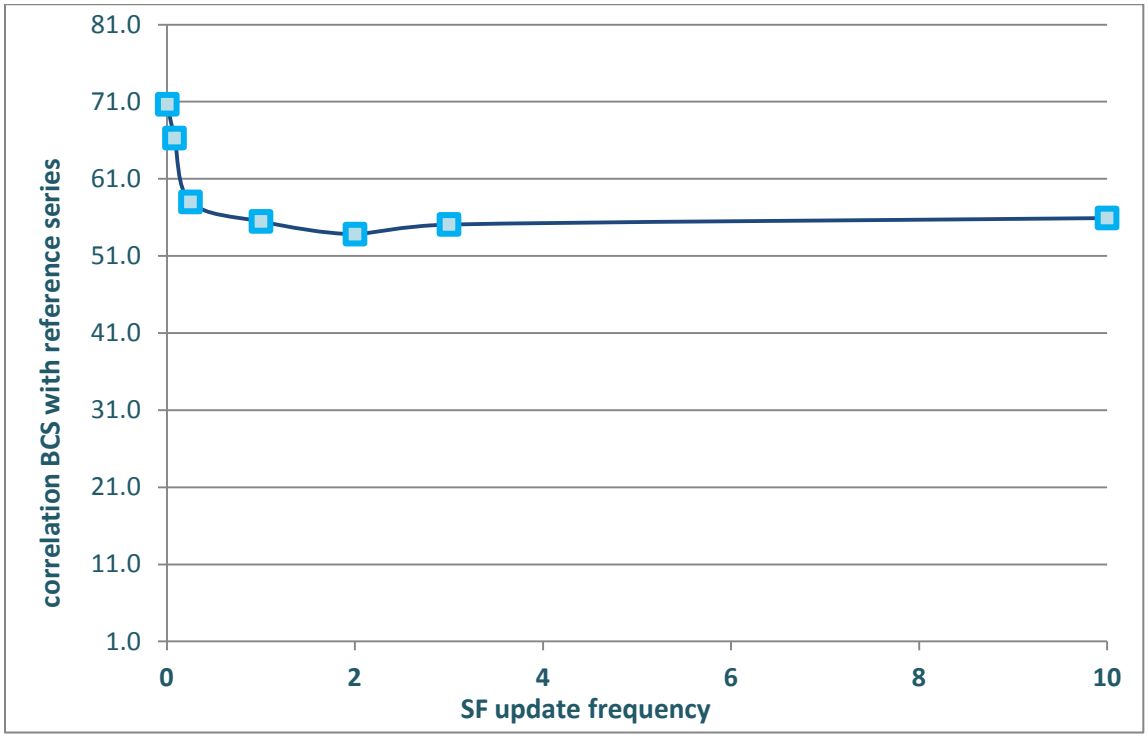

Chart 2.14. X-axis: grouped SFs with at least "x" update frequency. Y-axis: means of correlation BCS with reference series for grouped SFs, n=25

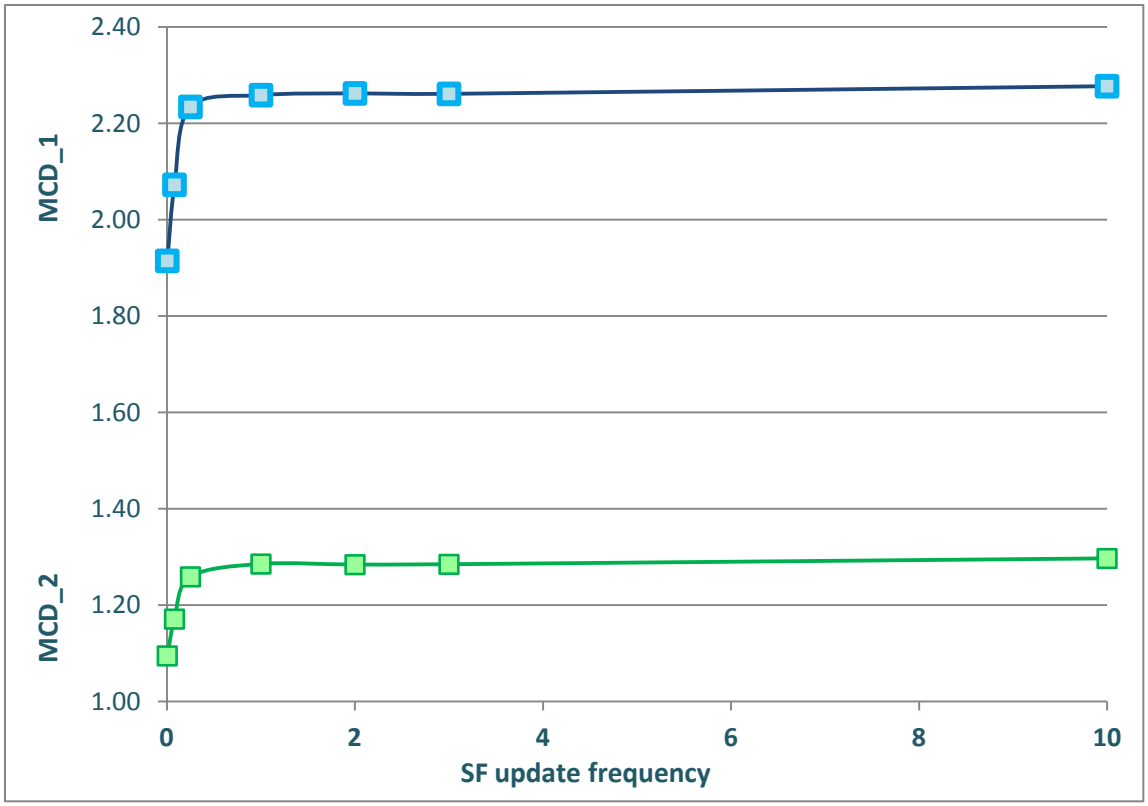

Chart 2.15. X-axis: grouped SFs with at least "x" update frequency. Y-axis: means of MCD_1 (upper line), MCD_2 (lower line), n=25

This point of view seems to be more helpful in reaching conclusion. Hence, one might accept hypothesis that more frequent updates of SFs favour better quality parameters.

2.c Sampling frames coverage - analysis of impact

SF coverage is new created variable based on information about SF size and population size. As it is not proper to compare both numbers among countries the simple ratio was calculated. Its formula is like below:

SF coverage = 100 * Size of the actual frame list / Population size

This ratio has character of scale variable however in this case these values are treated more like elements of an order.

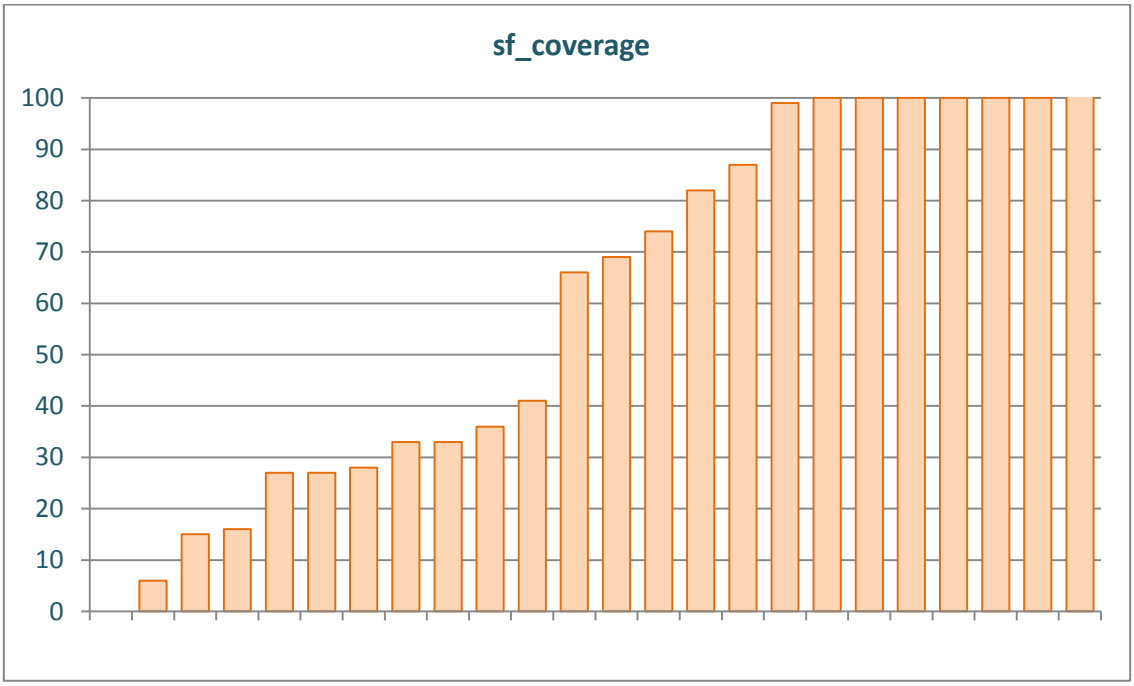

Chart 2.16. distribution of SF coverage, n=24

Looking for any interpretable pattern of relations between SF coverage and quality measures three scatter plot were produced (Charts 2.17 - 2.19).

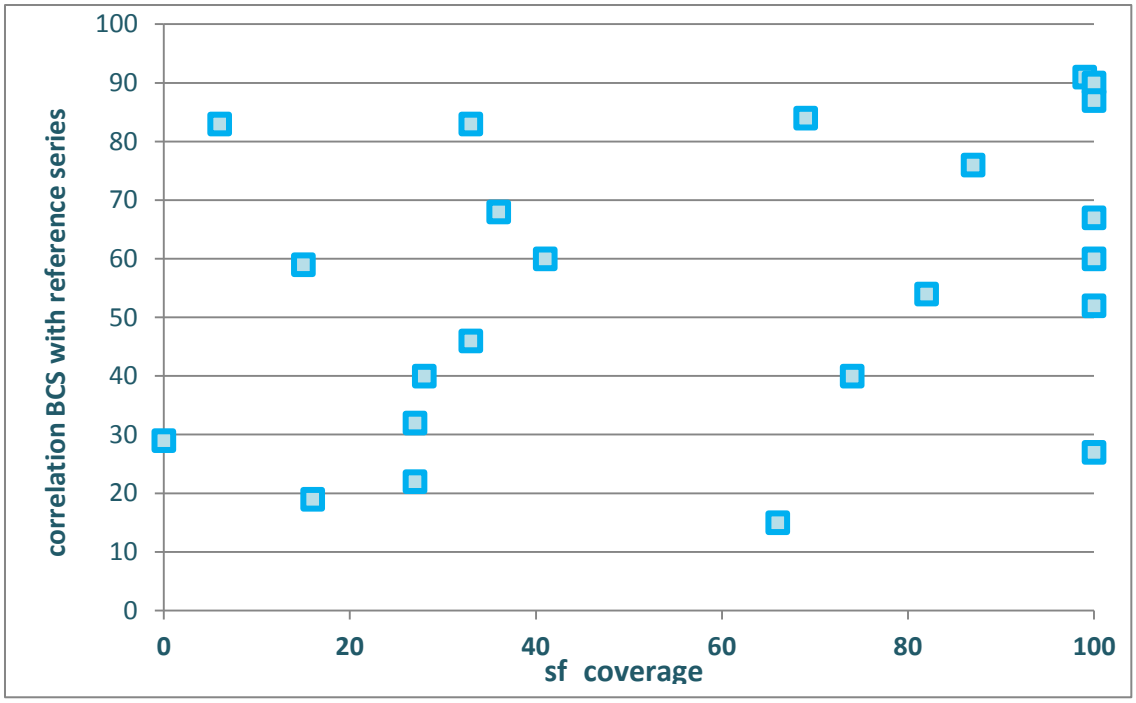

Chart 2.17.scatter plot for SF coverage and correlation BCS with reference series , n=24.

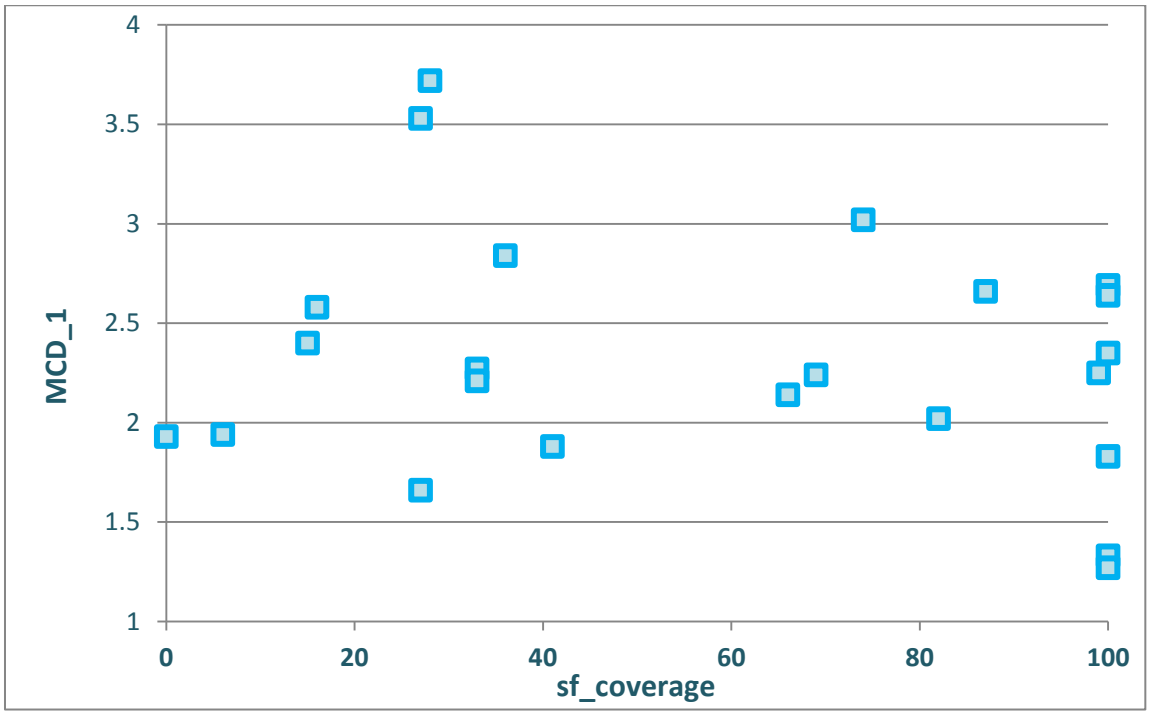

Chart 2.18.scatter plot for SF coverage and MCD_1, n=24.

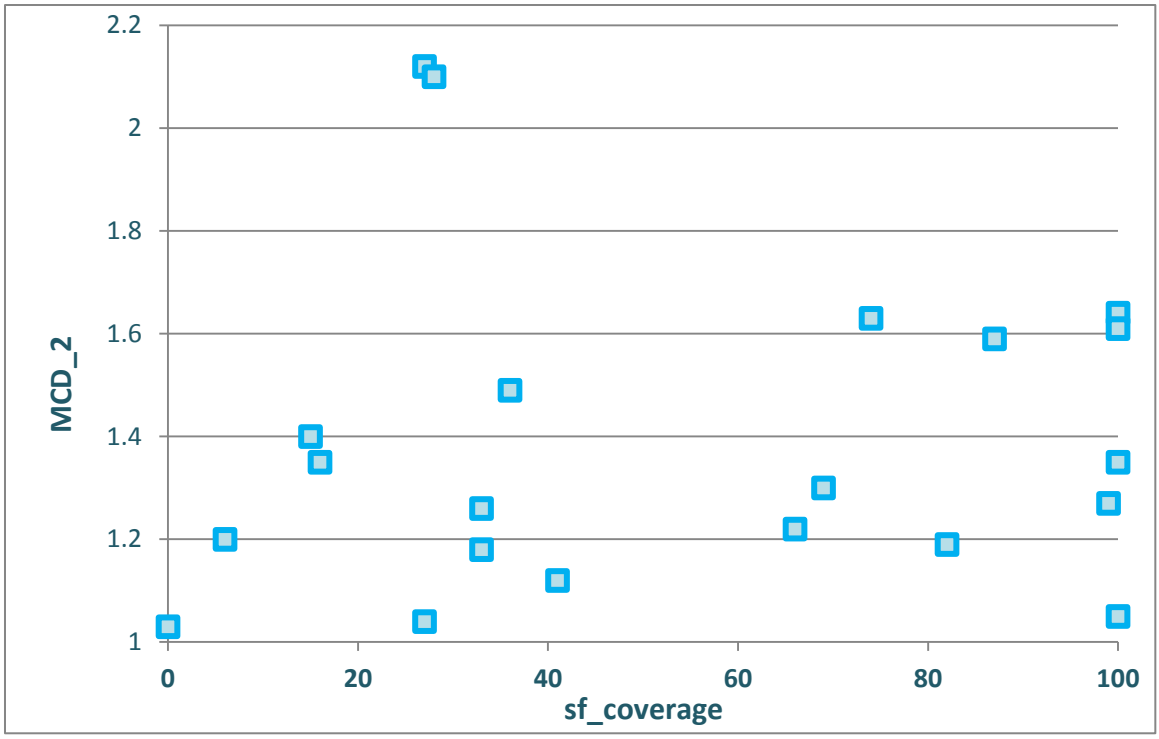

Chart 2.19.scatter plot for SF coverage and MCD_2, n=24.

Again, changing of perspective appeared to be necessary. Consequently groups of SFs were ordered by coverage level from 100 down to 0 including ones with higher coverage. For each group average quality parameters were calculated.

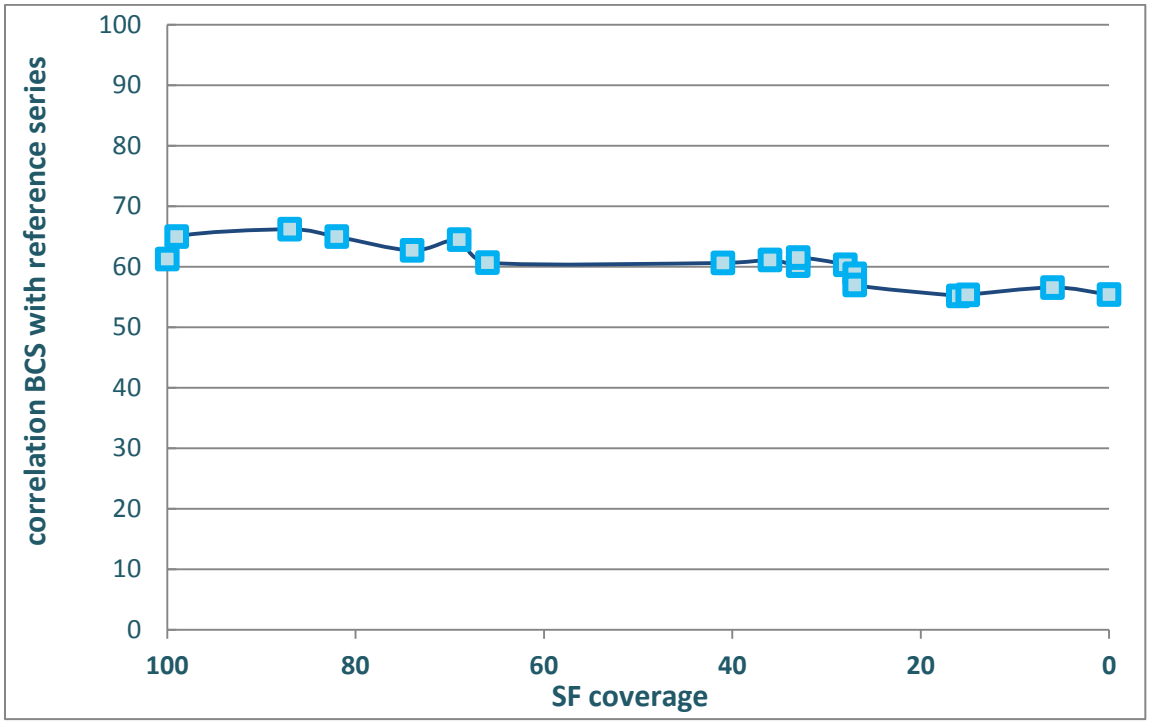

Chart 2.20. X-axis: grouped SFs with at least "x" SF coverage. Y-axis: means of correlation BCS with reference series for grouped SFs, n=24

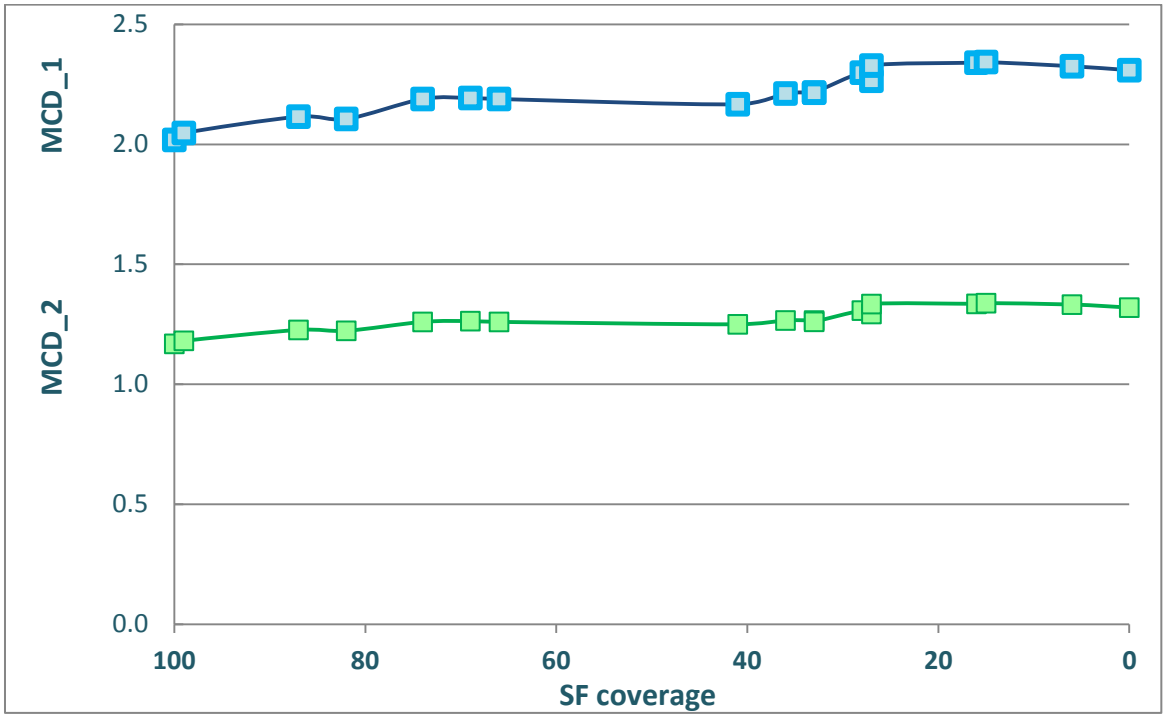

Chart 2.21. X-axis: grouped SFs with at least "x" SF coverage. Y-axis: means of MCD_1 (upper line) and MCD_2 (lower line) for grouped SFs, n=24

These charts do not bring any clear message. However fairly slight tendency is noticeable. Along with joining SFs with lower and lower coverage all quality measures are somewhat worse.

2.d Additional analysis

Revising all results one additional checking was proposed. The aim was to test in a more complex way what is the impact of update frequency and SF coverage together on quality measures. Therefore SF coverage was aggregated to three classes:

- The lowest - 0 to 50
- Medium - 51 to 90
- The highest - 91 to 100

The idea was to compare average quality parameters for SFs defined by combination of categorized SF coverage and update frequency.

		correlation BCS with reference series			
		categorized SF coverage			
		0 - 50	51 - 90	91 - 100	Total
SF update frequency	,003	60.0		74.0	71.2
	,080		40.0		40.0
	,250	42.7		46.0	43.5
	1,000	55.7	15.0	59.3	52.7
	2,000	19.0	54.0		36.5
	3,000		84.0		84.0
	10,000		76.0		76.0
	Total	49.2	53.8	65.0	55.4

Table 2.22. mean correlation BCS with reference series for respective level of SF update frequency and category of SF coverage, n=24

		MCD_1			
		categorized SF coverage			
		0 - 50	51 - 90	91 - 100	Total
SF update frequency	,003	1.88		2.02	1.99
	,080		3.02		3.02
	,250	2.68		2.02	2.52
	1,000	2.41	2.14	2.09	2.29
	2,000	2.58	2.02		2.30
	3,000		2.24		2.24
	10,000		2.66		2.66
	Total	2.45	2.42	2.05	2.31

Table 2.23. mean MCD_1 for respective level of SF update frequency and category of SF coverage, n=24

		MCD_2			
		categorized SF coverage			
		0 - 50	51 - 90	91 - 100	Total
SF update frequency	,003	1.12		1.17	1.16
	,080		1.63		1.63
	,250	1.51		1.12	1.41
	1,000	1.38	1.22	1.21	1.32
	2,000	1.35	1.19		1.27
	3,000		1.30		1.30
	10,000		1.59		1.59
	Total	1.39	1.39	1.18	1.32

Table 2.24. mean MCD_2 for respective level of SF update frequency and category of SF coverage, n=24

Looking at differences in averages it could be assumed that tenuous connection between quality measures from the one side and update frequency and SF coverage from the other side is probable.

3. Analyses summary and conclusions

Assuming sufficient differentiation among diverse sampling frames in terms of quality of measurement some analyses were conducted.

All features that describe sampling frames seem to have a small impact on quality measures. What is important, although tendencies are slight they are practically always coherent for all quality of measures.

Having no arguments for strong statistical conclusions some general findings can be given.

Better quality measures are rather observed when:

- Individuals are sampling frame units
- If telephone numbers are sampling frame units – fixed together with mobile rather than fixed alone
- The more frequent update of sampling frame list
- The higher sampling frame coverage

All conclusions rather support our intuition than bring striking findings. Features that describe sampling frames in principle can be indications of general research standards present on a particular market. Sampling frames are usually result of accessible official resources together with state regulations. Therefore size and type of sampling frames, frequency of available updates could be the result of local possibilities and consequently limitations of sample management. In a situation where different sampling frames are available it is reasonable to give recommendation based on presented outcome.

For more theoretical purposes there is a possible way to go deeper in the analytical process. In order to do that some more information could be completed and some new data collected. Being aware of limited access to local sampling frames there are probably elements worth changing on this stage of surveys' conducting.

References

1. Glossary of Statistical Terms OECD
2. Särndal, Carl-Erik; Swensson, Bengt; Wretman, Jan (2003). *Model assisted survey sampling*. Springer. pp. 9–12. ISBN 978-0-387-40620-6. Retrieved 2 January 2011
3. Turner, Anthony G. "Sampling frames and master samples". United Nations Secretariat. Retrieved 12/11/2012.
4. Data source: *Metadata_checked_by_partners - Consumers.xlsx*
5. Cochran, W. G. (1963). *Sampling Techniques*, 2nd Ed., New York: John Wiley and Sons, Inc.
6. Kish, Leslie. (1965). *Survey Sampling*. New York: John Wiley and Sons, Inc.