THE DESIGN OF A NATIONWIDE SAMPLE FOR ROMANIAN YOUTHS LIVING IN RURAL AREAS: A MULTIPLE SOLUTION PROBLEM

Iulian STĂNESCU[1]

Vlad ACHIMESCU[2]

Abstract: *This paper looks into methodological issues of sample design for a specific target population: Romanian youths (aged 16-35) that reside in rural areas. The first part of the paper will provide an overview into general issues that come from demographic trends, such as the overall population decline and especially the wave of emigration to Western Europe, and the pros and cons of the most widely used sampling techniques in Romania. The stages of the survey design itself comprise the second part. The most challenging part was how to reduce undercoverage problem by putting forward a sample frame, namely fieldwork procedures to identify as many elements of the target population as possible under real-life conditions. Multiple solutions were taken into consideration, before choosing one after field testing. The final part features an analysis of the survey focusing on the outcome of the sample design.*

Key-words*: youths, sample design, random route, electoral register, systematic sampling*

1. Introduction

In real-life situations, designing a sample is not a straightforward task. Besides the demographical issues of our times that arise from mass migration towards Western Europe, there are other, more practical issues, such as the client's needs and requirements, cost containment, and the need to keep field operator error to a

[1] Researcher, The Research Institute for Quality of Life. Email: stanescu.iccv@gmail.com
[2] Researcher, The Research Institute for Quality of Life. Email: achimescu.vlad@gmail.com

minimum. This paper aims to bring forward a discussion on the practical problems that arise from such issues while designing a nationwide sample of rural youths. All things considered, sample design in real-life situations is part of category of problems that have multiple solutions. Unlike the classical situation where a problem has just one correct answer, all other being false, we are dealing with several solutions, each with a different degree of efficiency.

The survey was part of a larger human resources development project of the Romanian Ministry of Labour and Social Protection entitled "An inclusive labour market in rural areas", and with European Social Fund co-financing. BPI Management Group, a member of the project consortium, ran the research part of the project and commissioned a study on poverty, social exclusion and labour market opportunities for young people aged 16 to 35 that reside in rural areas. The overall research project was larger in scope, with the survey of rural youths being just one component. In addition, the project featured a policy and institutional framework analysis, secondary analysis on the social and economic status of rural youths, based on official statistics, a survey of employers and prospective employers, another survey of local and county officials with responsibilities concerning employment and social inclusion, as well as an ample and in-depth qualitative research on rural youths, employers and public officials. This paper will look only in the rural youth survey component of the research, more specifically the sample design.

The project covered seven out of the eight development regions of Romania. This meant 40 out of the 41 counties, with one county and the municipality of Bucharest forming the eighth development region. As this region was more than 90% urbanized, it made no sense to be included in the project as a whole and in the research component. Furthermore, the client's brief asked that all counties should be included in the survey. In addition, the interviews would be face to face based on a multi-thematic questionnaire. This is an important element. The usual practice for nationwide Romanian samples for surveys using face to face interviews does not entail going into all the counties, mainly due to costs. The target size of the sample was set at 2,000 individuals; with the sample being representative of all youths aged 16 to 35 from the 40 counties. The fieldwork period was set in August 2011.

2. Demographic trends in Romania: population decline and migration issues

The first issue concerns the overall number of rural youths in Romania. For the 40 counties comprising the area selected for the survey, the total rural youth population amounted to 2.61 million persons aged 16 to 35, as of July 1, 2010, according to the National Statistics Office. In relative terms, this represented almost 98% of the entire

rural youth population of Romania. However, this data is accompanied by a revealing disclaimer note on the National Statistics Office website (INS, 2012). The main source of the statistics is administrative data provided by the Interior Ministry. According to the disclaimer, "the source does not cover the entire migration phenomenon, especially emigration. As a result, there is a severe underrepresentation of this phenomenon that leads to an overestimate of Romania's population".

Taking into account the client's requirement to reach all 40 counties, the issue of the nationwide and county population total in rural areas becomes even more significant. Setting up a survey for the rural population might seem unusual for in a European society. Like other Central and Eastern European countries, Romania still has a large, significant minority of the population residing in rural areas (Cace C., Cace S., Nicolăescu V, 2011, p. 20). In fact, the industrialization and urbanization process was not that far away in time. After the World War Two, just two generations ago, more than three quarters of the population resided in rural areas (INS, 2002). The share of the population residing in rural areas decreased to around 45% by the end of the Communist regime in 1989. In absolute terms, the rural population declined from 12.16 million to less than 10.84 from the late 1940s to 1989.

After the 1989 Revolution, the share of rural the population remained around 45% (INS, 2012). On July 1, 2010, 9.63 million persons were registered in the official statics as residing in rural areas. Based on this data, rural youths would represent around 27.8% of the entire rural population and 42.4% of the entire youth population of Romania.

The two overlapping demographic trends, (1) a net loss in population, including rural population, due to failing birth rates and another (2) net loss in population due to emigration, are important for sample design. In the 40 counties included in the research there are 2,824 communes – the lowest level of administrative division comprising one or, in most cases, several villages. The population decline, especially due to emigration, might lead to situations in which some villages have very few youths. The decline in fertility occurred in the late 1980s, as a social feedback to the forced pronatalist policies of the old regime (Zamfir, 1999), and increased markedly during the 1990s, first as a result of the legalization of abortion, second due to the explosion of poverty and decrease of the living standards (Zamfir, 1999, Zamfir, 2004, p. 47-51, Gheţău, 2004).

While in the 1990s, the main source of population decline was the low fertility, it was superseded by emigration in the last decade, with immigration far too low to make an impact (Gheţău, 2007, p. 3). During the two decades since the 1989 Revolution, work migration and emigration were low, with a visa regime in place for Western Europe. Romania's process of accession to the EU changed this. First, visa for the then 15 EU member states and the Schengen area were waivered in 2001. On January 1,

2007, Romania joined the EU, which further eased travelling abroad. A revealing insight about the size of the emigration wave was revealed by the 2002 census. The final census data showed a difference close to 1 million persons between the administrative data of the Interior Ministry, which forms the basis for the official population statistics (INS, 2012).

After the completion of this survey, the preliminary results of the 2011 census released in early 2012 revealed that Romania's population decreased to just 18.3 million, with some 700,000 being temporary out of the country (INS 2012). Even with arguable inherent organizational deficiencies that result in some part of the population not being counted, the overall figure of the census means a population decrease close around 17% less than a quarter of a century. All this has an impact on sample design, as public figures on the county and commune population totals are according to the Interior Ministry figures, not census data. To this extent, the disclaimer of the National Statistics Office becomes even more relevant.

3. The practice of survey design in Romania: electoral register, random route or CATI?

The practice of designing nationwide samples in Romania features probability designs, with stratified multistage samples being by far the most common. The probability design "assigns to that each element in the frame a known and nonzero chance to be chosen" (Groves, 2004, p. 94). The main difference between these designs is the sample frame used in the last stage of the sampling. The three types of sample frames used in Romania are:

1. all adults, through the use of the electoral register, which includes all non-institutionalized adults and excludes those that have lost the right to vote because of a penal sentence (a tiny minority); this type of sample frame is called in practice "the electoral register";
2. households, through systematic sampling in sample areas, such as city blocks or housing units that comprise a polling station; in survey practice, this is called "the random route" option;
3. households, through lists of fixed and mobile telephony subscribers; this is usually named after the data collection method, that is the computer assisted telephone interview (CATI).

As we have seen above, the name of data collection method takes over the type of survey design. The data collection methods most widely used are: (1) face-to-face interviews with printed questionnaires (PAPI – paper and pencil interviewing), which is under a rather slow, especially in academic research, process of being replaced with (2) computer assisted personal interviews (CAPI). The latter eliminates the

possibility that fieldworkers would complete the questionnaires themselves, enables real-time monitoring of the fieldwork, and dramatically decreases the completion time of the database. Moreover, there is no cost with inserting the data from the paper questionnaire into a database, as it is done by computer software at the moment the field operator presses the key on the terminals (tablet PC or small laptop). Computer assisted telephone interview (CATI) is rapidly gaining ground especially in marketing, outside academic research.

The key issue is to what extent the theoretical probability design is carried through in practice. The importance is critical, because from the client's point of view a survey is worth the cost as long as the sample data is representative for the target population. The representativeness of a sample is quantifiable for probability designs only. Because of different sample designs, "representativeness is a relative notion" (Rotariu, 2006 p. 87) even for such samples are less or more representative, not just entirely representative or not at all. Probability designs, as the name suggests, rely on the mathematics of probability theory to infer that sample data describe characteristics of a population. The central limit theory, for instance, is valid for probability distributions only (Brase and Brase, 2009, p. 302). There are also non-probabilistic designs such as snowball sampling, or more recently responded driven sampling (Heckathorn, 1997). While sometimes successfully approximating rare and dispersed populations (such as drug addicts), these designs are more efficient on local samples than on a national level, where interview operators are hard to control. Let us review the above mentioned sample frames with their advantages and disadvantages.

The use of the *electoral register* presents the obvious unique advantage of include almost all adults. In European countries, including the UK where voter registration is made by the authorities by default, without the citizens having to register to an electoral authority, the electoral register is used extensively for random sampling at the last stage of respondent selection (Foster, 1993). Persons that have lost their right to vote because of sentencing or institutionalized persons are not included, but such persons make a negligible minority. Moreover, the use of an electoral register as a sampling frame is consistent with PAPI or CAPI data collections and with multi-thematic questionnaires, which are used in academic research. Another plus is the possibility to select a random sample of respondents from the electoral register of a precinct. For obtaining a probability sample, a simple random sample of individuals is always better than a systematic one of households. In short, the electoral register satisfies best both the undercoverage and the randomness issues. Last but not least, it is easier to monitor the progress and performance of field operators. The selection of respondents from the sample frame is done by those in charge of sampling, not by field operators, thus removing an important source of error.

In practice, the use of the electoral register as a sample base is severely limited. There are two causes: the unavailability or lack of access to the electoral register and, quite unexpectedly an issue with overcoverage instead of undercoverage. Due to changes in administrative law in the mid 2000s, access to the electoral register, even for research purposes only, has been severely restricted. There is just one place nationwide where access is granted to the electoral role in electronic form: the Permanent Electoral Authority. While access could be asked for, it is not the same thing with being granted. Moreover, the only way to access data is by the terminals at the Authority's premises and to a very limited number of persons. This seriously hampers the sampling work by taking much more time. For a nationwide sample of 1,200 or 2,000 persons, such as the case for the rural youths research project, plus a considerable number of reserve contacts for the field operators, the amount of time required would simply be unpractical. Another way of gaining access to the electoral register is by submitting an official request to the mayors in each of the communes or cities included in a sample. In this case, the electoral register is given in hardcopy only, while paying a tax for the cost of the photocopies. The cost tends to be significant per page. Additionally, acceptance or denial of the request is at the will of the mayor's office. In short, the logistics and costs of access to the electoral register deem this variant unpractical, too. In rural areas, the electoral register has one more flaw. The addresses are not updated, so that the house numbers (most villages have no street names) do not correspond with those on the list, making it very difficult for the field operator to find a person just by its name in a village of thousands or hundreds. Moreover, in some villages, most inhabitants have similar surnames, an addition source of difficulty.

The second issue with the electoral register comes from the size of the population included in it. For the past more than 20 years since the Revolution, the population total on the electoral register has increased, while the overall population, both the Ministry of Interior and census data, which differ, point to a decrease.

Table 1

Population totals (adults only) in the electoral register, official statistics and census

Year	Electoral register	Adult population – official statistics	*Census 2002*
1990	17.200.722	16.608.208	
1992	16.380.663	16.288.133	
1996	17.218.654	17.146.982	
2000	17.699.727	17.437.135	
2004	18.449.344	17.088.071	*16.833.541*
2008	*18.464.274*	*17.257.981*	

Source: National Statistics Office

As we can see in table 1, the total number of adults registered in the electoral register has increased over time. During the same period, the Interior Ministry data shows a far lower increase, while the census data points to a decrease. This difference could not be became an obvious issue at the 2004 election. According to Comşa and Rotariu (2004, pp. 31-32), this phenomenon could not be explained by demographics, rather through low administrative capacity. The mayors do not update the electoral register. Therefore, people that moved on to another place, emigrated or are no longer alive still figure in the electoral register. In practice, this overcoverage of population means more that more contacts need to be given to field operators. Hedging for this and low response rates in urban areas, a field operator would require up to three times as many contacts for a certain target of interviews. This means more work and more time allotted for sampling from the electoral register, more time, frustration and travel costs for the field operators, especially in rural areas.

As a result of the abovementioned difficulties in accessing the electoral register, very few surveys use it as a sample frame. The most notable example of its usage in Romania is the Research Institute for Quality of Life's program of surveys, which began in 1990 (Mărginean and Precupețu, 2011). Taking into account the time and cost requirements of the client for the rural youths survey, it could not be considered as an option for sampling frame, in spite of its advantages.

The so-called *random route* is still the most widely used sample frame in survey design in Romania. It is used with the face-to-face interview data collection method. In fact, it is area sampling, with systematic sampling of households over an area, be it usually a precinct or just a part of a city or town. The field operator selects one in several households across his or hers progression through the sample area. Sometimes maps are provided. This would be an equal-probability method of selecting respondents, since every household seems to have the same theoretical chance of being included in the sample. It would make systematic sample function in a similar manner to simple random sampling, hence the random route name. In order to further "randomize" the approach, the procedure for selecting the individuals inside the household is a random one, usually the field operator looking for the person that is about to have or just had his or hers birthday. Notice that this is a hybrid, with a systematic sampling of households and a random one of individuals inside the households. Therefore, the random route is not really random. Developed in marketing surveys, this procedure is easy to use, cost effective, and benefits from field operators experienced at its use. However, marketing surveys are focused more on cost effectiveness at the expense of precision, while academic and other type of research require precision, even at a higher cost.

Its disadvantages lie in the risk of undercoverage and biased selection. The result is a non-probability sample, which is not representative of the population, regardless of

the sample size or the correct procedures used in selecting the sample areas. Of the three sample frames, it is the only one in risk of resulting in a non-probability sample. The main weakness lies in giving the field operator control over the selection of the households and, even more important, of the respondent from within the household. The field operator could easily not comply with the skip or sampling interval, through intended or unintended error.

However, the most serious problem is with respondent selection inside the household. If the selection probability among households is equal, the chance for a single adult to be selected is not, as it becomes inversely proportional to the number of adults in the household. There is little incentive for the field operator not to select the first person who comes across and report that person as the next that would or had his or hers birthday. This leads to underrepresentation of certain categories, such as males, young people or those with a job or a business and over-representation of other categories, such as retired people and women. After the sample data is collected, weighting becomes no longer just an option, but a requirement. Another source of undercoverage is the fact that the field operator does not progress through the entire sampling area. The selection process is completed when the interview target is achieved. In a precinct of 1,000 people this would understandably leave a significant part out of the selection process. Nevertheless, this is an unavoidable issue with area sampling.

The initial procedure called for a systematic sampling of households, followed by a systematic sampling of individuals. In effect, this meant a mini-census of the sample area, which is time consuming and not cost effective. In order to skip the birthday selection method, prone to fraud, and avoid its effect, the underrepresentation of certain population subcategories, Kish (1965) introduced the use of grids or contingency tables. The goal to achieve randomness and representation might have been reached in the mid 20th century America, but recent work in Hungary, a neighbouring Central European country, suggests that modifications are needed to fit the specifics of the survey (Nemeth, 2003).

In practice, the problem of lack of control over the field operators is partially resolved through a costly operation of field control. This requires another person to retrace, on the phone or on the field, the steps of the operator and check the compliance with the skip, the selection procedure inside the household, and the completion of the interview based on all the items on the questionnaire. Should fraud be detected, the same operator or another one would remake the interviews for the allotted sample area. While reducing the risk, the procedure is costly and reasonably effective with small samples that do not cover a large area or do not have so many sample areas. In many cases, while there is a procedure for control, it is not that extensive due to costs.

The so called *CATI* method is a sample frame comprising households reachable through fixed and mobile telephony. In this case, the frame consists of the list of telephony networks' clients that use subscriptions or pre-paid cards. As a sampling technique, CATI features several advantages. Firstly, it dramatically reduces costs. Phone operators are paid much less per questionnaire than field operators. There are no transport costs, which usually are a significant part of the survey budget. As with CAPI, the database is created by the software. Moreover, since the interviews could easily be recorded, the risk of operator fraud is completely eliminated. Secondly, it provides randomness in selecting the respondents. Unlike face to face interviews, the field operator no longer has to look for the respondents. The software selects the respondents from its database directly through random digit dialling (Groves, 2004, p. 74) or preceded by a stratification of sampling units (Groves, 2004, p. 127).

The main weakness of CATI as a sampling frame procedure stems from the risks to the probability design. The basic approach is a random selection from a list of households (for fixed telephony) and individuals (mobile telephony). There are three main sources of risk for that each individual in the target population would have the same known and nonzero chance of entering the frame: (1) ineligibility, (2) clustering, and (3) duplication. (1) Ineligibility means that some of the population has no telephony access, the numbers may be non-working or nonresidential (business, public services, etc.), or the numbers may be private, no longer listed in the phone list. In addition, the respondents for fixed telephony numbers tend to be older and female, according to the subset of population that spends more time at home. Poor people are less likely to be covered, while younger, more active people are more likely users of mobile than fixed telephony. (2) Clustering occurs when more users are assigned to the same number, usually fixed telephony. A person's chance to be selected is inversely proportional to household size. (3) Duplication means than one person could have in use more telephone numbers, being at the same a subscriber to one fixed telephony network and one or more mobile networks. This would increase that person's chances of entering the frame as compared with others. In addition to these three there is the issue of nonresponse due to refusals. Younger people or those living in large urban areas are less likely to accept the interview to lower social trust. The basic solution to these problems is the use of quotas sampling. In the sample design, the population is stratified according to age, sex, region etc. Targets of number of interviews are given to the phone operators' supervisor. As a technique, CATI is by and large avoided in academic research, especially because of the use of multi-thematic questionnaires that entail longer interviews.

4. Designing a sample for rural youths: stratification

This section will cover the stratification part of the design sample. In itself, it did not pose difficult challenges. The project's brief called for communes from 40 counties out of 41 to be included in the sample, which is seldom asked for by research clients. Since most clients do not have a detailed brief and because of cost containment, nationwide samples in Romania usually include a number of counties from the mid 20s to the mid 30s.

For the stratification procedure, the question was what kind of criteria to use for breaking down into categories the 2,824 communes in the 40 counties. Due to Romania's geography, plains, hills and mountains make up around one third of the area. Therefore, the communes vary not only in terms of population, but more importantly in terms of the layout.

Table 2

Sample strata for nationwide rural youths survey

No.	County	Communes by population size							
		Total number of inhabitants				No. of communes selected in the sample			
		Small	Medium	Large	Total	Small	Medium	Large	Total
1	Alba	12.074	16.939	12.177	41.190	0	1	1	2
2	Arad	8.478	20.394	28.625	57.497	1	1	1	3
3	Arges	48.371	25.882	12.987	87.240	2	1	1	4
4	Bacau	3.841	21.217	90.111	115.169	0	1	5	6
5	Bihor	12.211	30.204	39.735	82.150	1	1	2	4
6	Bistrita-N	5.958	16.150	37.094	59.202	0	1	2	3
7	Botosani	7.044	27.936	33.847	68.827	0	1	3	4
8	Braila	5.477	16.628	9.514	31.619	0	1	1	2
9	Brasov	5.240	14.054	28.385	47.679	0	1	1	2
10	Buzau	10.560	23.322	36.459	70.341	1	1	2	4
11	Calarasi	5.558	14.259	30.074	49.891	0	1	2	3
12	Caras-Severin	16.263	15.196	5.556	37.015	1	1	0	2
13	Cluj	12.709	12.834	36.651	62.194	1	1	1	3
14	Constanta	6.144	17.044	45.096	68.284	0	1	2	3
15	Covasna	7.234	6.028	19.577	32.839	0	0	1	1
16	Dambovita	3.821	16.188	85.917	105.926	0	1	4	5
17	Dolj	20.840	20.198	39.027	80.065	1	1	2	4
18	Galati	8.653	8.631	60.739	78.023	0	1	3	4
19	Giurgiu	6.910	15.983	27.170	50.063	0	1	2	3
20	Gorj	8.275	21.127	25.359	54.761	1	1	1	3
21	Harghita	7.495	17.973	27.669	53.137	1	1	1	3
22	Hunedoara	12.312	11.682	1.129	25.123	1	0	0	1
23	Ialomita	13.216	18.085	8.522	39.823	1	1	0	2
24	Iasi	4.265	18.379	110.569	133.213	0	1	6	7
25	Ilfov	0	0	0	0	0	0	0	0
26	Maramures	6.448	17.661	38.446	62.555	0	1	2	3

No.	County	Communes by population size							
		Total number of inhabitants				No. of communes selected in the sample			
		Small	Medium	Large	Total	Small	Medium	Large	Total
27	Mehedinti	12.760	15.156	9.685	37.601	1	1	0	2
28	Mures	13.881	19.846	45.049	78.776	1	1	2	4
29	Neamt	7.714	16.910	76.526	101.150	0	1	4	5
30	Olt	22.358	27.712	17.543	67.613	1	1	1	3
31	Prahova	8.805	14.918	88.416	112.139	1	1	4	6
32	Salaj	11.480	17.525	7.511	36.516	1	1	0	2
33	Satu Mare	6.411	16.679	35.597	58.687	0	1	2	3
34	Sibiu	7.363	16.496	19.269	43.128	0	1	1	2
35	Suceava	5.839	25.707	88.543	120.089	1	1	4	6
36	Teleorman	19.219	25.082	16.339	60.640	1	1	1	3
37	Timis	13.710	26.862	38.037	78.609	1	1	2	4
38	Tulcea	7.919	13.796	12.704	34.419	0	1	1	2
39	Valcea	15.423	26.260	13.391	55.074	1	1	1	3
40	Vaslui	11.730	25.832	33.188	70.750	1	1	2	4
41	Vrancea	9.470	15.446	41.292	66.208	0	1	2	3
	Total	433.479	748.221	1.433.525	2.615.225	22	38	73	133
	Share (%)	16,6	28,6	54,8	100	16,5	28,6	54,9	100,0

Where terrain is flat or hilly, the housing is usually clustered, either alongside around the major road through the commune or at least in compact major subunits. In mountainous area houses are more scattered, sometimes even making the mapping of the area very difficult. Another criterion would be the communes overall development level, which in turn is often strongly correlated with the proximity of a major city. Official data in terms of breaking down communes by predominant terrain type or level of development, though useful, is lacking. What is available is a population total for each commune, even the number of people aged 16-35.This data is provided by the Ministry of the Interior, so they should not be taken for granted. How much the emigration wave has changed the situation remains to be revealed by the 2012 census. However, we considered this data to be a useful criterion with the benefit of the doubt.

According to the number of inhabitants aged 16-35, the 2.824 communes from 40 counties feature the following breakdown (see table 2):

- up to 604 inhabitants; 959 communes totalling 433,479 people, 16,6% of total rural youth population;
- from 605 to 1.025 de inhabitants; 943 communes with 752,032 people, 28,6% out of total rural youths;
- over 1.025; 922 communes with 1,485,313 people, 54,8% out of total rural youths.

A matrix was designed using the two stratification criteria, the county and the commune type by population. Each commune had allotted to it 15 questionnaires, resulting in 133 communes in the sample (a handful had 16 due to rounding up to 2,000), selected with a probability proportional to commune size in terms of youth population. The result was a matrix with 120 cells (40 x 3), all eligible. For each cell selection of one or more communes, according to the matrix, by a simple random selection the total number of communes ranked alphabetically. As a result, the sample of communes was as follows:

- 22 communes in the first category, totalling 330 questionnaires, amounting to 16,5% of sample size;
- 37 communes in the second category, totalling 557 questionnaires, 27,85% of sample size;
- 74 communes in the third category, totalling 1,113 questionnaires, 55.65% of sample size.

Once the communes where identified, each one was checked up with the local field operators, looking for communes in the first category, which would be seriously depopulated or for other issues. In Hunedoara County, the single commune from the first category was replaced, aptly called "Bătrâna" (Romanian for Old Lady), after it was reported that the total number of youths in the commune is significantly lower than official data: less than 20 instead of around 50. The main challenge in the next step of sample design was the sample frame, namely devising the procedure for respondent selection.

5. Designing a sample for rural youths: the sample frame

After selecting the 133 communes where the fieldwork would take place, the last stage posed the challenge of what kind of sample frame to use. As we have seen above, all three major types of sampling frames – electoral register, random route and CATI – have their own pros and cons. Under the research brief, the use of the electoral register was not a realistic option due to time constraints. Added to the difficulty of access to the electoral register was the required additional effort to sort out the youths from all the entries in the electoral register. CATI was not really on the table since the brief called for a multi-thematic questionnaire. The interview's duration meant a very long phone call, at or above the limit of what is practical. The only option left standing was to devise a procedure of systematic sampling in the manner of the random route, using as sample areas parts of villages. As noted by Kalton (1983, p. 62), "the difficult problem of sampling rare populations arises when the survey population comprises only a small fraction of the frame, and when the frame does not provide the means to identify elements in the survey population".

This section of the paper deals with the ideas that were put forward, the presumptive solution, its field work and the final set of procedures for systematic sampling that made up the sampling frame for the survey of rural youths in Romania. As we have discussed above, the main disadvantage of the random route is the biased selection of the respondent households member by the field operator. Intuitively, the path to the solution meant limiting the control of the field operator over which household and which person within the household to visit.

The first option was the use of Kish tables. The field operators lacked experience in using them, which was a drawback. Moreover, it would not guarantee in a reasonable manner the reduction of household or respondent selection bias. The operators could add fictive household members so that the interviewee would be the person at home at the time of visit. Besides, having a households list as sample frame was risky due to demographics. Field operators could spend a lot of time before finding households with youths in the overall context of population decline and emigration.

A mini-census of youths in the village was another idea. It would involve asking for assistance from the local authorities for drawing up a list of young residing in the village, followed by random or systematic sampling. This option implied full co-operation from commune mayors, which was far from certain. Besides, filed operators would again be in some control, as could change the order on the list to fit the persons they found at home.

The mini-census idea had the merit of bringing forward a change from a household list to individuals. The frame would not contain households from a sample area, but individuals from the survey's target population. Addressing the issue of surveys that have as target population only a part of the sample frame, Kalton (1983, p. 61-62) suggests a two-phase design. In the first phase a cluster of households is selected to provide a list of at adults several times over the target of interviews. This is followed in the second phase with conducting screening interviews to select from all the adults in sample area only those that fall in the target population. A systematic sample is then used, with no more than one adult from a household.

At this point we developed a procedure for the selection of youths from sampling areas to be tested in a real-life situation. In order to get a list of all persons aged 16-35 from a village area, the field operator would need to cover the entire area, namely housing, small businesses, shops etc, in effect giving a picture of how the villages look like. All this information would be entered din a field card. For each household, just the youths, but every single one, would be entered, in descending order by age, with the first name, gender and age. The selection by age was introduced to two reasons. First, so that the field operator would not interview the first youth found in

the household; second, to circumvent the bias of non-selecting those around and over 30, which spend less time at home due to work.

The overall selection procedure required the following:

- the field operator visits each households in his or hers progress in the sampling area;
- a sampling interval of three, meaning a skip of two youths; the sampling interval covers youths only, not households;
- no more than one questionnaire per household;
- all people aged 16-35 are to be entered in the field card, but those that have left the household abroad or in other parts of the country would not be counted in the sampling interval; those that commute to work count for the sampling interval;
- if there are more than 3 youths in a household, they are entered in the field card, but do not count for the sampling interval;
- the first interview will be conducted with the eldest youth from the first household with youths;
- in case of refusal, the sampling interval is maintained with the skip of the following two;
- if the selected youth is not at home, the field operator would return later in the day; under no conditions should an interview be conducted with the next youth in the household.

Other detailed instructions were about the progress through the sampling area (see the *Typical village* figure). Operators would continue on the same part of the street from where the first interview was conducted. Rules were set about continuing on adjacent streets. Each field operator was provided with detailed instructions, including the attached map with an example for respondent selection in a typical village. The skip was set at two because of the population decrease trend. With the average household around 2.8-2.9 in rural areas, according to the 2002 census, the probability that a household would have more than 3 youths ages 16-35 would be low.

Another issue was where the operator would begin once in the village. The standard procedure was used during field testing, that is to choose a reference point on the main road like the commune hall or the church and start opposite. The field testing was done at the same time with the one for the questionnaire in Dâmbovița County, which due to its geography features communes in plains, hills and mountains. Field testing revealed a flaw in the procedure to select respondents, in that the reference point could take the field operators towards an area with few households. Therefore,

another procedure was put forward, whereby the mayors would be asked to recommend a starting point, preferably on the main road, in an area with enough households to reach the target of interviews. The target was set to 8 in the most populated village of the commune and 7 in one other randomly chosen village, but excluding from the selection hamlets, i.e. severely depopulated, small settlements.

6. Analysis of the field cards database

The data from the field cards was entered in a database of its own, different from the survey database. While the survey database includes 1,943 valid questionnaires, the other one is considerably larger with a total of 6,028 entries. It covers all the persons from the target population included in the target frame as well as those excluded because of long-term absence from their residence, like emigrants. This section of the paper features an analysis of the data from the field cards database. Using statistical tests, we will look into whether or not differences exist between the distributions of the youths that entered the survey sample, the ones in the sample frame, as well as other relevant categories, such as those unavailable or out of the country.

Out of the total 6,028 youths included in the field card database:

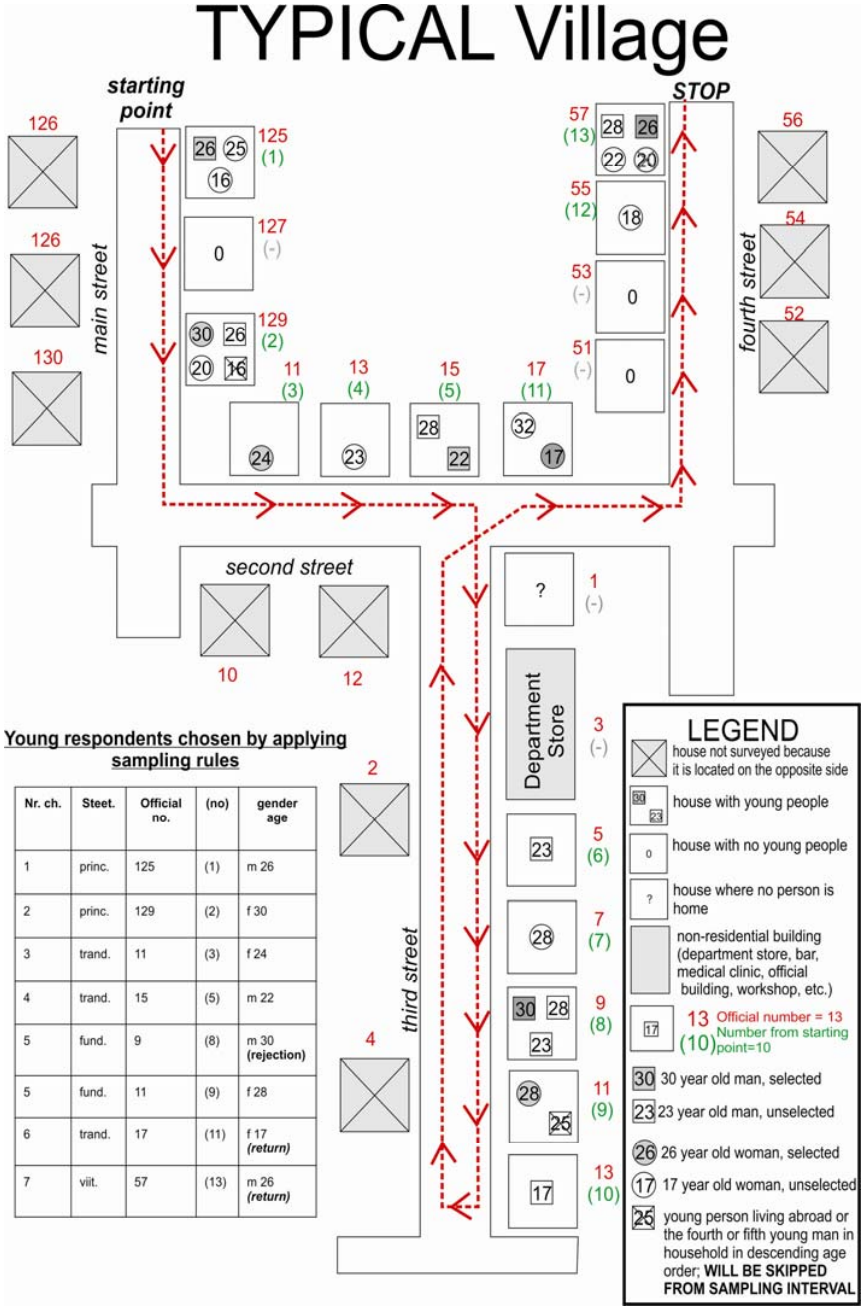
- 32% are youths that were interviewed, they formed the source for the survey database;
- 53% were screened but skipped, including those residing in Romania that were not at home at the time of the survey;
- 13% were not available, approximately half of which were out of the country;
- 2% were non responses due to refusal to take part in the survey.

The number of persons that were unable or unwilling to be interviewed is quite low. The former might be underestimated because of the difficulty to obtain socio-demographics data for youths where there was no relative at home at the time of the field operator's visit. These situations were in part remedied by using the neighbours as informants.

Table 3

Gender distribution: percentage of the survey database youths, skipped, unavailable, emigrants, and field card database total

Category	Survey database	Skip	Unavailable	Emigrants	Total
Men (%)	51	53	62	61	53
Women (%)	49	47	38	39	47



Nr. ch.	Steet.	Official no.	(no)	gender age
1	princ.	125	(1)	m 26
2	princ.	129	(2)	f 30
3	trand.	11	(3)	f 24
4	trand.	15	(5)	m 22
5	fund.	9	(8)	m 30 **(rejection)**
5	fund.	11	(9)	f 28
6	trand.	17	(11)	f 17 ***(return)***
7	viit.	57	(13)	m 26 ***(return)***

The gender profile of the survey database youths does not differ significantly (chi square test P-value is above α at 5% level of significance) compared to the field card database total: 51% percent men compared to 53%. However, for those that were unavailable the share of men was much higher than that of women (62% vs. 38%). This might be due to the fact that a disproportionately higher share of rural men is employed compared to women, thus increasing their chances of not being available.

Table 4

Age group distribution: percentage of the survey database youths, skipped, unavailable, emigrants, and field card database total

Category	<18	18-20	21-25	26-30	31-35
Survey database	13	18	25	22	22
Skipped	13	19	27	22	19
Unavailable	6	14	30	27	23
Emigrants	3	11	31	30	25
Total	13	18	27	22	20

The age group distribution for the survey database youths is very close to the one of those that were skipped and, more importantly, to the field card database total. There is a statistically significant difference between the survey youths and those skipped for the 31 to 35 age group: 22% compared with almost 19% (chi square test P-value is above α at 5% level of significance, but not at 1% level of significance).

Those that were unavailable have a higher share for the 21-30 age group (57% compared to 47% in the survey database), most likely due to commuters. For those out of the country, the age group distribution reveals a very low share for those under 21, which is explainable through the difficulty of finding work in another country at such an age. However, this does not affect the overall sample, as emigrants are around 7% of the field card database.

Table 5

Age group distribution, women: percentage of the survey database youths, skipped, unavailable, emigrants, and field card database total

Category	<18	18-20	21-25	26-30	31-35
Survey database	13	16	24	22	25
Skipped	15	20	27	21	17
Unavailable	7	14	34	21	24
Emigrants	6	10	37	23	24
Total	15	18	26	21	20

The largest difference of percentage points for all socio-demographical categories was found at the age group distribution for women. In the survey database the share of women aged 31-35 group is 8 percentage points higher compared with the same group in the skipped category and 5 percentage points higher than in the field card database. Such a difference in percentage points is statistically significant as well (chi square test P-value is below α at 1% level of significance). The higher response for housewives compared with employed women might be a cause. For women, the chance of employment is much higher around 30 and over than for in the lower or mid 20s.

Table 6

Age group distribution, men: percentage of the survey database youths, skipped, unavailable, emigrants, and field card database total

Category	<18	18-20	21-25	26-30	31-35
Survey database	12	19	27	22	20
Skipped	11	19	27	23	20
Unavailable	5	13	28	31	23
Emigrants	2	11	27	34	26
Total	11	18	27	24	20

For men, there are no statistically significant differences between the survey database, the ones that were skipped and the field card database. Therefore, in the overall database, when both genders are included, the difference for the 31-35 age group between the former and the latter databases climbs down to just 2 percentage points, which is reasonable.

Table 7

Employment status distribution: percentage of the survey database youths, skipped, and field card database total

Category	Survey database	Skipped	Total
Not employed	71	71	72
Employed in the village	16	14	15
Employed, commuter	11	14	11
Employed, unspecified	2	1	2

Regarding the employment status, between 70-72% are not employed, 14-16% are employed in their village, 11-14% are employed outside the village, having to commute, and for the rest of 1-2% that are employed there was no available

information. There are no significant differences concerning the not employed category between the three main categories, survey database, skipped, and field card database. The differences between employed in the village and commuters are statistically significant (chi square test P-value is below α at 5% level of significance), but their share in the overall population is low. Further differences on employment status in terms of gender and age group are very low and not statistically significant.

The analysis of the field cards database and its components on those that were not selected for interviews and thus not included in the survey database – skipped, unavailable, emigrants – suggests that the field operators complied with the respondent selection procedures. Further systematic control of the field workers confirmed that the sample interval was by and large complied with. There are no major differences in the distributions between those included in the survey database and those skipped in terms of gender, age group or employment status.

The youths that were to be selected, but did not carry out the interview, were mainly unavailable, not at home during the field operator's visit. Some of them were commuter workers. Men aged 25-30 and women aged 21-25 were unavailable more often than other categories. The profile of those that are out of the country is similar to the others that were unavailable: likely males, aged around 30 and over. The overall share of those that were to be selected, but did not carry out the interview is reasonably low (15%). Since their profile is not very different compared to the rest of rural youths, the survey database could be used without weighting.

Table 8

Comparison between field card database, the survey database and the National Statistics Institute's (INS) population statistics

Total	Field card database	Survey database	Ins	Percentage point difference
Men	51	52	53	-2
Women	49	48	47	+2
16-18	13	13	9	+4
19-20	18	19	14	+4
21-25	25	26	26	+1
26-30	22	22	24	-2
31-35	22	20	27	-5

A comparison between the sample data and the official data from the National Institute of Statistics shows only minor differences in terms of gender (2% less females in sample) and age (4% more youths aged 16-20, 2-5% more youths aged

26-35 in sample). Keeping in mind that the official data does not fully take into consideration the emigration phenomenon, these differences can be accounted for.

7. Discussion

Sample design, including the issue of sample frames is a problem with multiple solutions. When dealing with large populations, there is no such thing as on the one hand a single correct, full-proof way of designing a sample, representative of the target population and on the other hand an infinite number of incorrect sample designs. Each sample design has its own degree of representativeness, of cost efficiency and data quality. The client's needs and goals play an important role, ranging from the highest potential precision to as cost effective as possible at the expense of precision. Therefore, the basic approach to sample design should set its starting point in the project's brief and not in a universal sample design applicable by sociologists to all briefs.

Dealing with the unknowns about the target population and the undercoverage issues of any major sample frame type leads to a situation where decisions need to be taken under persistent uncertainty. Open cognitive systems tend to make better decisions when faced with these challenges. Such systems for decision feature several traits, such as: continuous exploration for alternatives, sensitivity to feedback, keeping decisions open to revision, admitting relative truths, and the logic of problems with more than one solution, but with different levels of effectiveness (Zamfir, 1990, pp. 90-92).

The challenge of designing a representative sample for the young population of Romania falls into this category, as many possible options are available, but none seems to avoid all known biases. Our solution was developed from both theoretical and empirical points of view, always keeping the issue of costs in mind. In the end, it was an elimination process that led to the final two-step design presented in this article. *CATI* and *random route* were excluded mainly because of inflated sample frame (most households do not contain youths, most fixed line telephony subscribers are people over 35) but also undercoverage, since using a random route would have meant skipping over households with many youths, and the fact that not all young people use phones. Non-probabilistic methods, like *responded driven sampling* were also excluded, due to not being representative for such a large population such as the 2.6 million of rural youths. *Sampling with electoral lists* was disregarded because of their unreliability and the legal impediments to use them. In conclusion, after eliminating these methods and checking the available literature (Kalton, 1983, p.61), all that remained was to use a two-step design, the first step being a mini-census of an area of the village from the sample (insufficient funding prevented making a census of the whole village).

There remained the problem of selecting the respondent. The lack of experience of interview operators and the difficulty of verifying them disfavoured the use of Kish tables and led us to choose the systematic sampling of individuals and not households in the second step. Field testing before starting allowed for some more fine tuning (such as setting the starting point and revising the statistical step), and through this iterative process a method was put forward. The data collected through field cards allowed for an analysis that validated the choices made in the sample design for the rural youths survey, recommending it for future national studies where the target is only a fraction of the entire population and is not found in every household.

On a more general note regarding the three main types of sample frames used in Romania, there should be more use of the electoral register. The problem of undercoverage is best resolved with this type of sample frame. Access to the electoral systems should be revised, returned to the more open manner of the 1990s. Since the electoral register is maintained by taxpayer money, including the public institution that has the software and the database for its centralized maintenance, there is no sensible reason for making access so difficult, even for commercial research purposes. On the other hand, the overcoverage of the electoral register is a serious issue, stemming from poor administrative capacity. It is unacceptable for a country in the 21st century that the total of registered voters is higher by the order of millions than the current population statistics and census data.

In Romania, the list of fixed and mobile telephony users is arguably the most used sample frame of the (near) future. Its use will become even more widespread for marketing, due to its cost efficiency. Some, but not all, non-commercial surveys, will use CATI as well, while in some special circumstances, such as academic research projects, face-to-face interviews will remain and call for the use of the other two sampling frames. Without electronic access to the electoral register, even with a modest tax, CATI is the only practical, cost-effective alternative. Internet based polls could become an option for target populations such as urban residents, upward mobile, younger people.

Last but not least there is the random route. When the sampling interval uses households, the random route is not really random. A screening process should be followed by systematic sampling of individuals. For both sampling frames that use face to face interviews as data collection methods, it is essential to reduce the risk of questionnaire fraud and selection bias by keeping key decisions, such as the respondent selection, away from the discretion of the field operators. On the other hand, this should not lead to the other extreme of control freakery or demanding the all but the impossible from the field operators. Field testing is essential both for the sampling frame and questionnaire design.

As for data collection methods, PAPI is to be replaced by CAPI in Romania as well. The process has already begun with major market research organizations. While the initial cost is substantial, there are long term benefits in costs per research reduction, less risks of questionnaire fraud, fieldwork monitoring and decrease of database completion times. Not wasting paper is an additional point for consideration. Regarding sample size, the higher the better. Although sometimes challenging logistically, it is better for a nationwide sample to be much higher than a 900-strong minimum. Even if the overall margin of error does not decrease by that much, advanced analysis becomes more robust and overall precision is gained, reducing the occurrence of risky stratification choices. Also in terms of precision, randomness is a strong point, from the selection of sampling areas during stratification to selection of respondents in the sample areas by use of the sampling frame.

The data collected through field cards allowed for an analysis that validated the choices made in the sample design for the rural youths survey. As a final note on the sample design process, let us remind ourselves of C. Wright Mill's (2000 [1959]) call for keeping an open mind, using the sociological imagination, while trying to avoid abstract empiricism and the bureaucratic ethos.

References

Brase, C. H. & Brase, C. P. (2009). *Understandable Statistics. Concepts and Methods,* New York, Houghton Mifflin Company.

Cace C., Cace S., Nicolăescu V. (2011). "*Economic competences in rural entrepreneurship – current approaches and perspectives*, Logos - Universalitate - Mentalitate - Educație - Noutate: conferință internațională, Vol. 4: Secțiunea Ştiințe economice şi administrative, Lumen, Iaşi, pp. 13-28

Comşa, M. & Rotariu, T. (2004). *Alegerile generale 2004: o perspectivă sociologică,* Cluj, Eikon.

Foster, K. (1993). The electoral register as sampling frame. *Survey methodology bulletin,* 33**,** 1-42.

Gheţău, V. (2004). Declinul demografic al României: ce perspective? *Sociologie Românească,* II**,** 5-41.

Gheţău, V. (2007). *Declinul demografic şi viitorul populației României,* Bucureşti, Alpha MDN.

Groves, R. M. (Ed.) (2004). *Survey Methodology,* Hoboken, Wiley.

Heckathorn, D. D. (1997). Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems,* 44**,** 174-199.

Kalton, G. (1983). *Introduction to Survey Sampling,* Newbury, Sage.

Kish, L. (1965). *Survey Sampling,* New York, John Wiley and Sons.

Mărginean, I. & Precupeţu, I. (Eds.) (2011). *The Paradigm of Quality of Life,* Bucharest, Expert.

Mills, C. W. (2000 [1959]) *The Sociological Imagination,* Oxford, Oxford University Press.

Nemeth, R. (2003). Sampling design of health surveys: household as a sampling unit. *Luxembourg Income Study Working Paper Series.* CEPS/INSTEAD.

Rotariu, T. (Ed.) (2006). *Metode statistice aplicate in ştiinţele sociale,* Iaşi, Polirom.

Zamfir, C. (1990). *Incertitudinea. O perspectivă sociologică,* Bucureşti, Editura Ştiinţifică.

Zamfir, C. (1999). Tranziţia demografică şi problemele sociale asociate. IN ZAMFIR, C. (Ed.) *Politici sociale în România: 1990-1998.* Bucureşti, Expert.

Zamfir, C. (2004). *O analiză critică a tranziţiei,* Iaşi, Polirom.

Population statistics data

INS. "Tempo: population and demographic structure". Available at https://statistici. insse.ro/shop/index.jsp?page=tempo3&lang=ro&ind=POP102A, accessed on June 2, 2012

INS. (2003) The 2002 Population and Housing Census. Volume I

INS. (2011) Romanian Statistical Yearbook 2001

INS. (2012) Romanian Statistical Yearbook 2011

INS. (2012) "Press release on the provisional results of the 2011 census". Available at www.insse.ro/cms/files%5Cstatistici%5Ccomunicate%5Calte%5C2012%5C Comunicat%20DATE%20PROVIZORII%20RPL%202011.pdf, accessed on February 2, 2012