RURAL PLACE OF LIVING - General Target Variable Report (GVR)

1. General Information

The dichotomous target variable T_RURAL indicates whether the respondent lives in a rural (value 1) or non-rural (value 0) place (cf. Table 1).

SDR2 employs the following basic operational definition of rural:

a) respondent's place of living has up to 5,000 inhabitants;

b) respondent's place of living is in a rural area;

c) any combination of (a) and (b).

Given the variation in source variables that capture the type of respondent's place of living, we extend this basic definition to include additional information (as discussed in Section 3.2). Whenever we use the extended definition of "rural" to assign respondents the value 1 on T_RURAL, we code the harmonization control variable C_RURAL_EXTEND as 1 (see below, and Section 3.3).

T_RURAL is accompanied by three harmonization control variables that capture specific features of the "place of residence" source variables (see Table 1, and Section 3.3):

1. C_RURAL_EXTEND is a nominal variable that identifies when source variable values go beyond the SDR2 basic definition of rural area, yet we code such values as 1 on T_RURAL. This methodological indicator ranges from 0 to 3 .
2. C_RURAL_SOURCE indicates whether the target variable T_RURAL is based on a country-specific or standardized (pre-harmonized) source variable whose answer categories are common across countries of a given survey wave.
3. C_RURAL_CATEGORY_NUM provides the total number of response categories in a source schema about place of residence.

The target variable report for T_RURAL is accompanied by the following Excel documents:

- The Detailed Variable Report (DVR): T_RURAL_DVR_SDR2.xlsx. DVR Excel files in SDR2 systemize all information about source variables that were used for harmonization into a given target variable of the SDR2 database;
- Three Crosswalk Tables (CWT): T_RURAL_CWT_SDR2.xlsx, which contains details about mapping of source values to target values and codes for all control variables despite T_RURAL_EXTEND, for which we provide and additional CWT file: CWT file - T_RURAL_EXTEND_CWT_SDR2.xlsx.

Table 1.1. RURAL PLACE OF LIVING: Description of the target, source, and control variables

	Variable description	**Variable name**	**Variable values a**
Target variable	Respondent's place of residence	T_RURAL	1 = Rural 0 = Non-rural
Source variables			See: T_RURAL_DVR_SDR2.xlsx T_RURAL_CWT_SDR2.xlsx

			T_RURAL_EXTEND_CWT_SDR2.xlsx
Control variables	Extended definition of "rural"	C_RURAL_EXTEND	0 = Basic definition of "rural" 1 = Source category joins 'rural and small-sized non-rural' [textual description] 2 = Lowest category lists places as max. 30,000 inhabitants [numeric description] 3 = Source category joins 'rural, and max. 30,000 inhabitants' [both textual and numeric description]
	Type of source schema	C_RURAL_SOURCE	1 = Target variable based on country-specific classification of place of residence 2 = Target variable based on standardized classification of place of residence
	Total number of response categories in source schema of place of residence	C_RURAL_CATEGORY_NUM	1 = One category in a source variable … 241 = Two-hundred forty-one categories in a source variable

[a] Missing values are assigned according to the SDR2 missing codes schema, provided in the Appendix.

2. Survey Projects

Source variables that we used for T_RURAL appear in 23 international survey projects: ABS, AFB, AMB, ARB, ASES, CB, CDCEE, CNEP, EB, EQLS, ESS, EVS, ISJP, ISSP, LB, LITS, NBB, NEB, PA1, PA2, PPE7N, VPCPCE, WVS, 168 waves and 3082 national surveys. The data cover 154 countries and years from 1966 to 2017.

3. General Rules and Procedures

3.1. Source data description

To construct T_RURAL, we rely on three types of multi-categorical source variables:

(1) Source variables that record place of residence as ordered numerical intervals, for example NEB/1 source variable `s2` where 1 = < 5000; 2 = 5000-19,999; 3 = 20,000-100,000; 4 = 100,000 - 1,000,000 and 5 = > 1,000,000;

(2) Source variables that record place of residence as ordered textual qualifiers, for example ABS/1 source variable `level3` where 1 = urban; 2 = rural and 3 = no registration); and

(3) Source variables that record place of residence as a combination of numbers and textual qualifiers, for example ISJP/1-2 source variable `v706` where 1 = urban (> 5000) and 2 = rural (< 5000).

Hence, in SDR2 the operational definition of "rural" encompasses a basic component and an extended component, for special cases discussed in Section 4 and marked via C_RURAL_EXTEND.

The main differences in source variables describing respondents' place of living can be summed up as follows:

- *Number of response categories and type of information stored in response values*: In some source datasets the variable is dichotomous (rural/urban), while other datasets have more detailed information on locality type, which is stored (i) in numeric form (locality size), (ii) as textual description, or (iii) a mix of both.
- *Availability of source coding schemas harmonized ex-ante:* Some survey projects offer source variables that use only country-specific codes to describe respondents' place of living. In other projects, data producers provide additional variables with harmonized categories that can be compared between countries. For example, the World Value Survey provides two 'locality type' variables. The first variable has detailed country-specific categories (e.g. Japan – 1. 12 major large cities i.e., Tokyo, Osaka, etc.; 2. 150,000 more residents in cities; 3. 50,000 to 150,000 residents in cities; 4. Up to 50,000 residents in cities; 5. Rural districts), while in the second variable this information is recoded into three categories (1. rural/village; 2. small/medium town; 3. large town). There are also project waves with country-specific categories only for some countries.
- *Response origin:* in some survey projects respondents answered the question on locality type; in others, the interviewer/coordinator coded the locality type of respondent's place of living.

3.2. Rules of transformation of source variables into target variable

First, to decide which variable to harmonize into T_RURAL when a source data file provides more than one variable describing respondent's place of living, we apply the following rules:

1. **Prioritize country-specific variables.** When both pre-harmonized and country-specific measures of respondents' place of living are available, we choose the country-specific variable.
2. **Prioritize the country-specific variable that originates from respondents' self-assessment of their place of living.** If a dataset provides several country-specific measures of respondents' place of living, we select for harmonization the source variable that measures respondents' own assessments, rather than the variable that originates from survey administrators.
3. **Prioritize the source variable with more detailed information about respondents' place of living.** If a source dataset contains more than one pre-harmonized 'locality type' variable and no country-specific variables, we select for harmonization the variable with more detailed and clear information about rurality.
4. **Prioritize textual over numeric information**. If all source country-specific or pre-harmonized variables have the same number of answer categories, we prefer the ones

whose answer categories are textual (e.g. “rural”, “village”) than numerical (e.g. “less than 5000 inhabitants”)

Once we select the source variables, we apply the following harmonization rules to create T_RURAL:

1. **Code T_RURAL = 1** if:

- The basic SDR2 definition of “rural” is met:
  (a) size of locality is smaller than 5,000 inhabitants;
  (b) textual response in source variable states: *rural*, *village*, *country*, *countryside*, *remote area*, *farm*, *smaller areas* or *tribal*;
  (c) any combination of (a) and (b).

- Extended SDR2 definition of “rural” (flagged via C_RURAL_EXTEND) is met:

  (d) the smallest category of a source variable codes respondent’s place of living in a category described, textually, as ‘rural, small-sized non-rural (e.g. “*village and small town*”);

  (e) the smallest category of a source variable is listed as max. 30,000 inhabitants, and there is no other category for “rural”;

  (f) the smallest category of a source variable is described as rural, max. 30,000 inhabitants using both textual and numerical information.

2. **Code T_RURAL = 0** when source values indicate non-rural locality.

3. **Code T_RURAL as INSUF,** the SDR2 missing code for source variables that correspond to target concept (here, place of living) but whose response options as a whole cannot be mapped to target values, for example:
   1. source category indicates more than 30 000 inhabitants, rural
   2. the lowest response category ranges from over 30 000 inhabitants to ‘*and below’* or *‘less than’* a higher value (e.g., *less than 150 000* in ASES `V0348`)
4. **Code T_RURAL as UNFIT**, the SDR2 missing code for a single response option which does not fit to the rurality concept, for example *refugee camp* (e.g. ARB/2 `v13`), *hostel* (in a few waves of ISSP in South Africa), *no settlement* (CNEP/3/HU `Z.Hu.L.urban`), *other answer* (ISSP/2010 `urbrural`), *non self-representing PSU’s* and *all locations in non-SMSA* (PA8NS `v297`), and *other country* (PPE7N/NL `v453`).

5. **Standardize missing codes** (for SDR2 missing codes schema, see Appendix).

3.3. Methodological variables that accompany T_RURAL

T_RURAL is accompanied by three harmonization control variables.

First, C_RURAL_EXTEND identifies when source variable values go beyond the SDR2 basic definition of rural area, and we code such values as 1 on T_RURAL. This harmonization control takes values:

- 0 when the source value meets the SDR2 basic definition of rural;
- 1 when the source value exceeds the SDR2 basic definition of rural by textually describing a place as both rural and small-sized non-rural, for example "village and small town", and this is the smallest category among source answer options;
- 2 when the source value indicates a place of living of max. 30,000 inhabitants; this is the smallest category among answer options, and there is no other category for "rural";
- 3 when the source value indicates a place of living of max. 30,000 inhabitants and includes also a textual description that identifies locality as rural, for example 'rural, max. 30,000 inhabitants'; this is the smallest category of a source variable, and there is no other category for "rural."

Second, C_RURAL_SOURCE indicates whether the schema of the source variable used for T_RURAL is country-specific (value 1) or was pre-harmonized by data producers (value 2).

Third, C_RURAL_CATEGORY_NUM indicates the total number of categories of place of residence in a given source schema. The total number of categories of a given source variable in SDR2 ranges from 1 (ABS/1-3 Hong Kong, ABS/2,4 Singapore, WVS/5 Zambia) to 241 (ISSP/2008 Philippines) response options per one source variable (based on realized values).

Special Cases

Table 4.1 provides information about special cases and decisions we made.

Table 4.1. Information about variable name, values and codes for special cases

Dataset, variable name, country (if country level)	Case	Coding		Comment
		T_RURAL	C_RURAL_EXTEND	
EB				For all Eurobarometer waves we systematically chose only pre-harmonized variables. We do so because country-specific variables display problems, such as large departures from World Bank data on urban/rural population for the respective country-year.
EB/42 v529 Norway	*5 = Scattered populated*	1	0	Coded as rural with C_RURAL_EXTEND =0 since it constitutes the lowest response category, although it does not clearly fit the target description of rural. Same EB/52.1 v666 (Sparsely populated area)

EB/44 v816 Belgium	*4 = Autres localites (Other localities)*	1	1	Coded as rural because this is the smallest category for Belgium. Same for EB/49 v627, EB/52.1 v666, and EB/54.1 v403.
EB/44 v618 Finland	*4 = Rural center*	1	1	Documentation does not explain what rural center means. Because this is the smallest category, we code it as rural. Same situation appeared in EB/49 v627, EB/52.1 v666, EB/54.1 v403, and in EB/49 v627 (although in the latter case the category is the second lowest).
ISJP/1-2 V11008	*10*	0	0	Undocumented source value. We assume that documentation for 9,11,12 refers actually to 10,11,12.
ISSP/1985 V119 GB-GBN	*4 = Country village or town*	1	1	Although it is not the smallest category, we decided to code it as rural. Same holds for ISSP/1987 v84 and ISSP/1990 v104.
ISSP/1987 v84 Austria	*6 = Town in rural area*	1	1	Type of area drives our coding decisions: We code 'town in rural area' as rural; vice versa, village in urban area is coded as non-rural. Same holds for ISSP/1988 v108 and ISSP/1989 v125.
	7 = Village in urban area	0	0	
ISSP/1987 v84 USA	*8 = Town or village (2.500 to 9.999)*	1	3	Coded as rural following the definition of the US Census Bureau. Same for ISSP/1988 v108, ISSP/1989 v125, ISSP/1990 v104, and ISSP/1991 v122.
	9 = Incorporated area less than 2.500 or unincorporated area of 1.000 to 2.499	1	0	
ISSP/1985 v119 Australia	*4 = Country town, 1.000 – 20.000*	1	3	Because this category captures settlements located in remote non-coastal areas of Australia, we coded it as rural. Same for ISSP/1987 v84, ISSP/1990 v104, and ISSP/1991 v122.
ISSP/1992 aus128 Australia	*4 = Country town, 1.000*	1	3	Because this category captures settlements located in remote non-coastal areas of Australia, we coded it as rural.
ISSP/2004 au_size Australia	*5 = A small country town (under 10,000 people)*	1	3	Because this category captures settlements located in remote non-coastal areas of Australia, we coded it as rural. Same for ISSP/2004 v341, ISSP/2005 AU_SIZE, ISSP/2006 AU_SIZE, ISSP/2007 AU_SIZE, ISSP/2008 AU_SIZE and ISSP/2009 AU_SIZE.
ISSP/1989 v126 Norway	*5 = Rural, 3.000 - 29.999*	1	3	Because of the presence of scattered and remote rural areas in Norway, we coded this category it as rural. Same for ISSP/1990 v105, ISSP/1991

				v123, ISSP/1992 v127, ISSP/1995 v309, and ISSP/1996 v292.
ISSP/1992 i128 Italy	*5 = Open country with township*	1	1	Based on the ordering of categories in the source variable (1=Large city, 2=Suburb of a large city, 3=Small city, 4=Village, 5=Open country with township), we assume that category 5 is "less" than '4=Village' and code it as rural.
ISSP/1999 x_size Hungary	*804 = 20010-10000*	1	2	Lower value of range (20010) is most probably a typo: with other categories going in descending order ("2 mill ing, Budapest; 100000-250000; 10001-100000, …"), this category likely corresponds to "up to 10000".
ISSP/2002 V333 Denmark	*5 = Town,less than 5 000*	1	3	Coded as rural because this is the smallest category for Denmark. Same coding decision holds for ISSP/2004 v350, ISSP/2005 DK_SIZE, ISSP/2006 DK_SIZE, ISSP/2008 DK_SIZE, and ISSP/2009 DK_SIZE,
ISSP/2006 JP_SIZE Japan	*5 = Less than 50.000, town, village*	1	3	Coded as rural because this is the smallest category for Japan. Same holds for ISSP/2008 JP_SIZE, and ISSP/2009 JP_SIZE.
WVS_1_6 X049CS Uganda, South Africa, Zimbabwe	*710002, 716002, 800002 = 500-7999 (Village)*	1	3	Based on textual (village) and numerical (500-7999) information, we coded this source category as rural, although it is not the smallest category for these countries.

In addition, we note that:

- In three cases - CNEP/3 Spain (Z.Sp.L.Habitat), ISJP/1-2 (v25025) and LB/1996 (th) - we found source variables whose names and labels suggest they refer to the locality type of respondent, but whose response categories are not documented. This precluded their harmonization into T_RURAL.
- ISJP wave 1 (1991) source variable on place of residence provided information only for Germany East and Germany West, which is also reflected in T_RURAL and its accompanying control variables.
- NBB has a pre-harmonized variable ss18 with one exception for Estonia, wave 4, where we additionally used the country-specific variable ss18b_e. Despite this one case, we decided to code C_RURAL_SOURCE = 2, meaning pre-harmonized source classification, as all other countries of NBB have the pre-harmonized response categories.
- In CNEP/3 some source variables used for T_RURAL are pre-harmonized, but most are country-specific. Thus, we code C_RURAL_SOURCE as 1 (country-specific).

To assist decisions on rural/urban cuts, when needed we also used information from the World Bank: Percentage of Population in Rural Areas (in % of Total Population)

http://data.worldbank.org/indicator/SP.RUR.TOTL.ZS and the US Census Bureau (www2.census.gov/geo/pdfs/reference/GARM/Ch12GARM.pdf)

Appendix: Codes for missing values in SDR2

In the SDR database v.2 we identify different situations that warrant to be treated as missing data. Table A.1 lists all SDR2 missing value codes:

Table A.1. Codes for missing values in SDR2

SDR tag [a]	SPSS (STATA) codes	Label
		Standardized source codes for missing values
DK	-1 (.a)	Don't know
NA	-2 (.b)	No answer
REF	-3 (.c)	Refusal
DU	-4 (.d)	Don't understand the question
DNR	-5 (.e)	Any combination of DK, NA, REF, DU
INAP	-6 (.f)	Inapplicable
NEC	-7 (.g)	Not elsewhere classified
		SDR created codes for missing values
UNFIT	-8 (.h)	Source value does not fit to target
ERR	-9 (.i)	Errors in source data and undocumented source values
COMBI	-10 (.j)	Different missing codes on multiple sources taken for a target
CINAP	-11 (.k)	For control variables only: inapplicable
INSUF	-12 (.l)	For survey: Insufficiently defined response categories
QNA	-20 (.t)	For survey: Question not available

[a] Abbreviations for the labels corresponding to the SDR2 codes for missing values. These tags are used in the Crosswalk Table (CWT) files (Excel) that accompany documentation of SDR2 target variables.

In exceptional situations when codes for missing data listed in Table A.1 cannot be used, we apply a system missing <null> value.