Workshop

Building Multi-Source Databases for Comparative Analyses

Survey Data Recycling as an Analytic Framework for Survey Data Reprocessing (I)

The SDR Team

SDR Project, 2017-2021, SMA-1738502

PIs: Craig Jenkins, Irina Tomescu-Dubrow, Maciek Slomczynski – Sociology, OSU, PAN

Han-Wei Shen, Spyros Blanas – Computer Science and Engineering, OSU

Ilona Wysmulek, Sociology, PAN

Joonghyun Kwak, OSU

Computer analyst: Przemek Powalko, PAN

Graduate students: Nika Palaguta, Weronika Boruc, Denys Lavryk - PAN; Yamei Tu, OSU.

Researchers: Marcin W. Zieliński, Marcin Ślarzyński - PAN.

Advisory Board:

Claire Durand, University of Montreal, WAPOR

Peter Granda, University of Michigan, ICPSR

Dean Lillard, Department of Human Sciences, OSU

Malgorzata Mikucka, Mannheim University, MZES

Pamela Paxton, UT Austin

Markus Quandt, GESIS

SDR Project – what we do

Create the **SDR database**, a multi-country multi-years data structure for comparative, cross-national research on three main areas:

- Democracy and Political Participation
- Social Capital and Political Participation
- Social Capital and Wellbeing

Develop **methodology** for reprocessing extant survey data into multi-dimensional data structures, such as the SDR database

Develop **analytic tools** for using and analyzing the SDR database

Why is so?

Methodology dimension: agency as antidote to alienation

Part I: inter-survey variability in the context of data reprocessing in general

Part II (after coffee break): ex-post harmonization & inter-survey variability

The SDR Database V.2

Selected international survey projects

Asian Barometer
Afrobarometer
Americas Barometer
Arab Barometer
Asia Europe Survey
Caucasus Barometer
Consolidation of Democracy
Comparative National Elections Project
Eurobarometer
European Quality of Life Survey
European Social Survey
European Values Study
International Social Justice Project
International Social Survey Programme
Latinobarometro
Life in Transition Survey
New Baltic Barometer
Political Action II
Political Action - An Eight Nation Study
Political Participation and Equality
Values and Political Change
World Values Survey
New Europe Barometer

Selection criteria:

(1) contain measures of *political attitudes and behaviors, social capital,* and *wellbeing* + main correlates; (2) non-commercial; (3) designed as cross-national, preferably, multi-wave; (4) national samples are intended as representative of the adult population; (5) English language documentation (study description, codebook, questionnaire); (6) freely available in the public domain.

SDR Database

	SDR 1	**SDR 2**
# survey projects	22	23
# waves	89	174
# national surveys	1,721	3,485
# respondents	2,300,000	4,400,000
# countries/territories	142	169
Time span	1966-2013	1966-2017
# source data files	81	215

SDR framework of analysis

Define, measure and store, as indicators, variability due to:

- source survey quality

- data reprocessing (i.e. ex-post harmonization)

Methodological biases and errors understood as consequences of:

(a) deviations from standards of documenting and preparing survey data suggested in the specialized literature (e.g. Biemer and Lyberg 2003)

(b) inter-survey variability of items measuring the same issue

(c) harmonization procedures

SDR analytic framework – source data quality

Define

Measure

Store as indicators

variability due to deviations from standards of documenting and preparing survey data suggested in the specialized literature (e.g. Biemer and Lyberg 2003)

SDR framework: defining source survey quality

Total Survey Error (TSE) + Survey Process Quality Management (SQM)

3 dimensions of survey quality

a) Quality of the **data records** in national datasets (i.e. computer files)

- errors can lead to distortion of empirical results.

b) Quality of surveys as reflected in the **survey documentation**

- inadequate information in documentation reduces confidence in the data

c) Degree of consistency **documentation <-> data records** in the computer file

- processing errors can affect the overall usability of the survey

Operationalization:

a) Data Records in the Computer File: are data records formally correct?

Summary index on the basis of 4 variables:

Are survey weights free of formal errors (not inflating sample size)?	Yes = 1, No = 0
Do survey cases (respondents) have unique identification numbers (IDs)?	Yes = 1, No = 0
Is the proportion of missing values for gender and age within the standard limits (< 5%)?	Yes = 1, No = 0
Is the data file free from repeated cases (duplicates)?	Yes = 1, No = 0
Effect of positive answers (Yes = 1): Less distortion of research results based on the data	

Operationalization: (b) **Survey Documentation: How were the data collected?** (for SDR2)

Var. name	Var. label	Value
rsp_rate_info	Value of response rate is available	0/1
rsp_rate_value	Value of response rate	value
rsp_rate_approx	Value of response rate in the documentation is approximated	0/1
rsp_rate_calc	Response rate calculated based on numbers provided in the documentation	0/1
int_mode_info	Information about interview mode is available	0/1
transl_info	Information on translation method is available	0/1
transl_value	Information indicates that a professional translation method was employed	0/1
pret_info	Information about whether or not pretesting was performed is available	0/1
pret_value	Information indicates that pretesting was performed	0/1
fctrl_info	Information about whether or not fieldwork control was performed is available	0/1
fctrl_value	Information indicates that fieldwork control was performed	0/1
univ_info	Information about the universe is available	0/1
sample_info	Information about sampling scheme is available	0/1

Operationalization **b) Survey Documentation: How were the data collected?** (for SDR2)

Info on the type of the sampling scheme

Info on the type of interview mode

Summary index of binary indicators

Operationalization:

c) Consistency Documentation <->Data: processing errors categories

Based on analysis of variable values for 7 variables:

gender, age, year of birth, education levels, years of schooling, trust in parliament, participation in demonstrations

Do variable values in the codebook correspond to the values in the data file?	Yes=1 No=0
Illegitimate Variable Values	
Misleading Variable Values	
Contradictory Variable Values	
Variable Values Discrepancy	
Lack of Variable Value Labels	
Presence of errors, Yes = 1: decreased interpretability of the data	

Operationalization: Data Records in the Computer File

Survey weights

Survey weights assign an adjustment number to each respondent. Persons in under-represented get a weight larger than 1, and those in over-represented groups get a weight smaller than 1.

However, some numbers >1 and some numbers < 1 seems suspicious.

MIN&MAX (weight)

Ranges of MIN(wght):

exactely=0 in 42 surveys!
1.91 Philippines (ISSP 1991)

Ranges of MAX(wght):

0.92 Lithuania (NBB 2001)
90.32 New Zealand (ISSP 2007)

These findings prompted us to study weights carefully.

Main work on weights in SDR1 and SDR2 is done by Marcin Zieliński

Frequency of weighting procedures

43.4 % poststratification type of weighting only

8.5 % design type of weighting only

22.9 % combined

25.2 % no information on the type of weighting

Components of wght. factors

Gender (62.4 %)
Age (61.5)
Region (39.3)
Urbanity level (24.8)
Education (18.7)
Economical factors (1.4)

Corrections for HH samples (13.8)
Corrections due to the stratified sampling (21.8)

Quality of weights

Technically “good weight”

MIN(wght) > 0 and MIN(wght) < 1

MAX(wght) > 1 but small

mean(wght) = 1

sd(wght) as small as possible

Consequences

MIN(wght) = 0 : excluding cases

high MAX(wght) : possible bias

mean(wght) <> 1 : inflation or deflation
of the net sample size (stnd errors, potential bias)

high sd(wght) : high variance introduced into the data

mean(weight)

70 % mean(wght) != 1

Less strict: 0.999=< weight <= 1.001
12.7 % bad

e.g.:
Philippines (ASB 2010) = 0.83
Philippines (ISSP 1996) = 3.29

Cross-project perspective

No evident errors:

Americas Barometer (AMB)
Comparative National Elections Project (CNEP)
European Quality of Life (EQLS)
European Social Survey (ESS)
World Values Survey (WVS)

Summing up (SDR 1)

National surveys differ in:

- the use of data weighting
- weighting procedurę (post-stratification, design, combined)
- composition of weighting factors (gender, age, region, urban, education, economic factors)
- quality of weights (errors in min/max, mean, sd)

Correcting errors and provided recalculated weights

Strategy for SDR2

1. Preserve as much information on weights as possible – coding:

- the use of data weighting (yes/no)
- weighting procedurę (post-stratification, design, combined)
- composition of weighting factors (gender, age, region, urban, other)
- quality of weights (errors in min/max, mean, sd)

2. Providing new weights (Re-weighting as needed due to errors)

Advantages:

re 1 – maximum information for users

re 2 – elimimnation of errors and standardizing impact on the data

Biblio

Marcin W. Zieliński, Przemek Powalko and Marta Kołczyńska (2018). “The Past, Present and Future of Statistical weights in International Survey Projects: Implications for Survey Data Harmonization.”In: Timothy P. Johnson, Beth-Ellen Pennell, Ineke I. Stoop, Brita Dorer (Eds.) Advances in Comparative Survey Methodology (Wiley Series in Survey Methodology), John Wiley & Sons, Inc.

Recommended:

Dominique Joye, Marlène Sapin, and Christof Wolf. 2019. Weights in Comparative Surveys? A Call for Opening the Black Box. In *Harmonization*: Newsletter on Survey Data Harmonization in the Social Sciences Vol. 5, No. 2.