Detection of duplicates in the Survey Data Recycling project

Przemek Powałko, IFiS PAN
ppowalko@ifispan.waw.pl

Building Multi-Source Databases for Comparative Analyses, Conference & Workshop
Warsaw , December 16-20, 2019

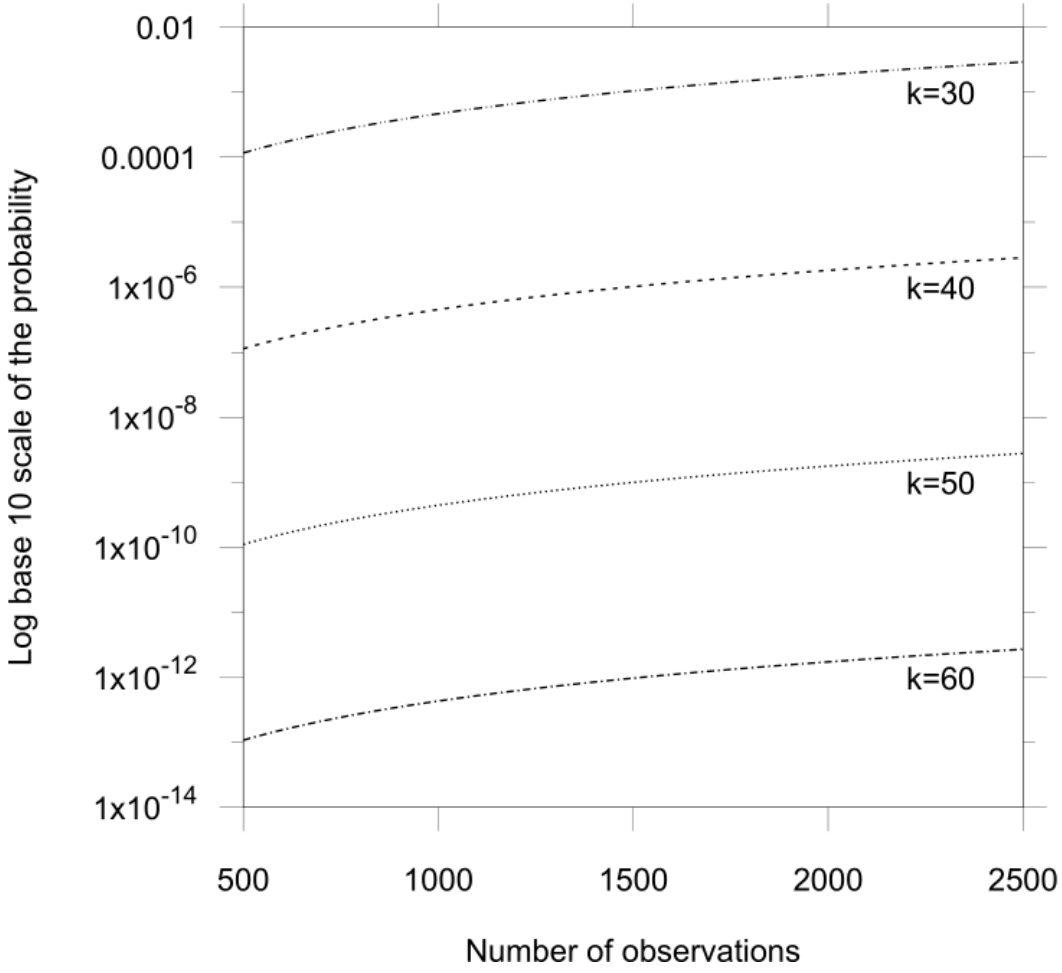
Outline

- Discovery
- Detection methods
  - Review
  - Hamming distance
  - Hamming diagram
- Results
  - Uncovering duplicates
  - Results: Duplicates in projects, surveys, and countries
  - Surveys with extreme duplication
- Antipodal cases
- References

Discovery

- Serendipity: screening SDR 1 data (2013)

- A simple probabilistic model of survey data
  - Dichotomous variables (a very conservative assumption)
- Results: a single duplicate with the probability 0.01 for
  - 30 independent variables needs 4,646 cases
  - 40 independent variables needs 148 thousands cases
  - 50 independent variables needs 475 millions cases



Detection method: Hamming distance

- Distance between cases (response patterns) can be measured in various ways → Hamming distance

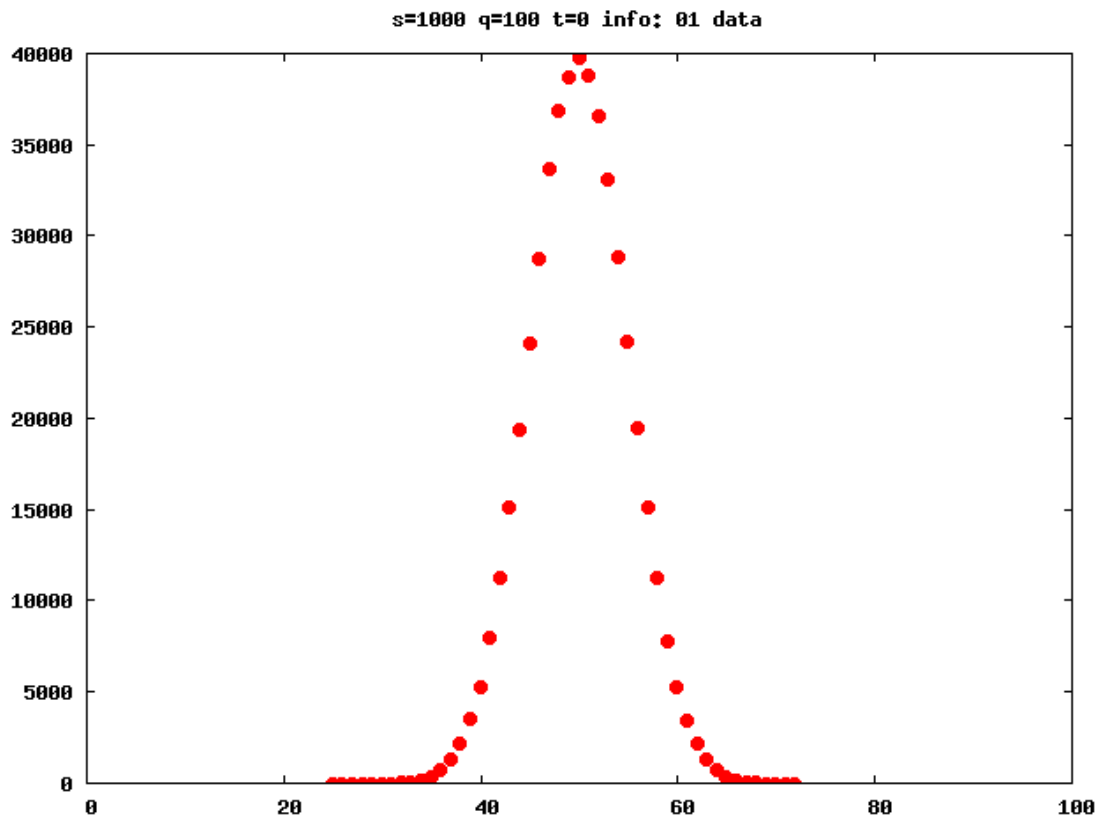
CASE#	VAR_1	VAR_2	VAR_3	VAR_4
A	3	4	2	1
B	3	5	5	3
C	3	5	5	3
D	3	5	3	2

				Hamming distance
0	1	1	1	3
0	0	0	0	0
0	0	1	1	2

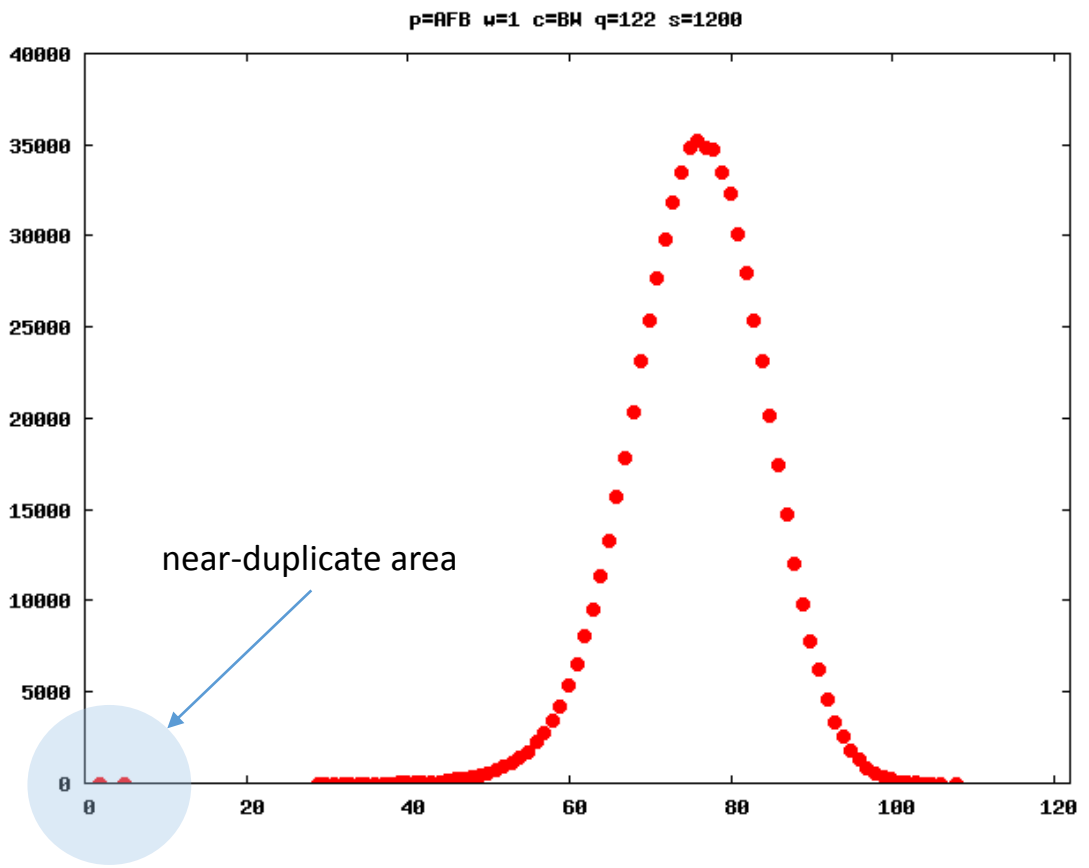
- Choose variables (ideally, covering all questionnaire items)
- Compare every case with all other cases
- Determine the distance between cases
- The existence of a duplicate is equivalent to the Hamming distance = 0 (see cases B and C above)

Detection method: Hamming diagram

Hamming diagram for a simulated data set



Hamming diagram for a real data set



Results: Uncovering duplicates

- From each survey data set sequentially remove the following blocks of variables:
  - Original case IDs (T.id)
  - Technical variables (T)
  - Interviewer's remarks (I)
  - Respondent's age and gender (R.a, R.g)
  - Urban/rural variables (R.u)
  - Information about household composition (R?, R??)
  - Other variables derived (RD) or calculated (RC) from the original responses
- At each step observe uncovering duplicates in remaining response patterns
- The final outcome: all variables the respondent is supposed to answer

Results: Uncovering duplicates

	case composition											number of infected							
	T.id	T	I	R.a	R.g	R.u	R?	R??	RC	RD	R	#vars	srvs	Δ	cntrs	Δ	#dups	Δ	~vars
V1	1	1	1	1	1	1	1	1	1	1	1	32428	2		2		**72**		203
V2	0	1	1	1	1	1	1	1	1	1	1	32332	90	88	60	58	**1342**	1270	245
V3	0	0	1	1	1	1	1	1	1	1	1	30397	116	26	71	11	**2371**	1029	237
V4	0	0	0	1	1	1	1	1	1	1	1	29056	143	27	76	5	**2868**	497	242
V5	0	0	0	0	1	1	1	1	1	1	1	28886	153	10	79	3	**2993**	125	238
V6	0	0	0	0	0	1	1	1	1	1	1	28794	156	3	79	0	**3060**	67	234
V7	0	0	0	0	0	0	1	1	1	1	1	28779	156	0	79	0	**3064**	4	234
V8	0	0	0	0	0	0	0	1	1	1	1	28159	158	2	79	0	**3068**	4	227
V9	0	0	0	0	0	0	0	0	1	1	1	27919	159	1	80	1	**3069**	1	229
V10	0	0	0	0	0	0	0	0	0	1	1	27435	161	2	80	0	**3086**	17	224
V11	0	0	0	0	0	0	0	0	0	0	1	26963	162	1	80	0	**3088**	2	222

Results: Uncovering duplicates

	case composition											number of infected							
	T.id	T	I	R.a	R.g	R.u	R?	R??	RC	RD	R	#vars	srvs	Δ	cntrs	Δ	#dups	Δ	~vars
V1	1	1	1	1	1	1	1	1	1	1	1	32428	2		2		72		203
V2	0	1	1	1	1	1	1	1	1	1	1	32332	90	88	60	58	1342	1270	245
V3	0	0	1	1	1	1	1	1	1	1	1	30397	116	26	71	11	2371	1029	237
V4	0	0	0	1	1	1	1	1	1	1	1	29056	143	27	76	5	2868	497	242
V5	0	0	0	0	1	1	1	1	1	1	1	28886	153	10	79	3	2993	125	238
V6	0	0	0	0	0	1	1	1	1	1	1	28794	156	3	79	0	3060	67	234
V7	0	0	0	0	0	0	1	1	1	1	1	28779	156	0	79	0	3064	4	234
V8	0	0	0	0	0	0	0	1	1	1	1	28159	158	2	79	0	3068	4	227
V9	0	0	0	0	0	0	0	0	1	1	1	27919	159	1	80	1	3069	1	229
V10	0	0	0	0	0	0	0	0	0	1	1	27435	161	2	80	0	3086	17	224
V11	0	0	0	0	0	0	0	0	0	0	1	26963	162	1	80	0	3088	2	222

Results: Uncovering duplicates

	case composition											number of infected							
	T.id	T	I	R.a	R.g	R.u	R?	R??	RC	RD	R	#vars	srvs	Δ	cntrs	Δ	#dups	Δ	~vars
V1	1	1	1	1	1	1	1	1	1	1	1	32428	2		2		**72**		203
V2	0	1	1	1	1	1	1	1	1	1	1	32332	90	88	60	58	**1342**	1270	245
V3	0	0	1	1	1	1	1	1	1	1	1	30397	116	26	71	11	**2371**	1029	237
V4	0	0	0	1	1	1	1	1	1	1	1	29056	143	27	76	5	**2868**	497	242
V5	0	0	0	0	1	1	1	1	1	1	1	28886	153	10	79	3	**2993**	125	238
V6	0	0	0	0	0	1	1	1	1	1	1	28794	156	3	79	0	**3060**	67	234
V7	0	0	0	0	0	0	1	1	1	1	1	28779	156	0	79	0	**3064**	4	234
V8	0	0	0	0	0	0	0	1	1	1	1	28159	158	2	79	0	**3068**	4	227
V9	0	0	0	0	0	0	0	0	1	1	1	27919	159	1	80	1	**3069**	1	229
V10	0	0	0	0	0	0	0	0	0	1	1	27435	161	2	80	0	**3086**	17	224
V11	0	0	0	0	0	0	0	0	0	0	1	26963	162	1	80	0	**3088**	2	222

Results: Uncovering duplicates

	case composition											number of infected							
	T.id	T	I	R.a	R.g	R.u	R?	R??	RC	RD	R	#vars	srvs	Δ	cntrs	Δ	#dups	Δ	~vars
V1	1	1	1	1	1	1	1	1	1	1	1	32428	2		2		**72**		203
V2	0	1	1	1	1	1	1	1	1	1	1	32332	90	88	60	58	**1342**	1270	245
V3	0	0	1	1	1	1	1	1	1	1	1	30397	116	26	71	11	**2371**	1029	237
V4	0	0	0	1	1	1	1	1	1	1	1	29056	143	27	76	5	**2868**	497	242
V5	0	0	0	0	1	1	1	1	1	1	1	28886	153	10	79	3	**2993**	125	238
V6	0	0	0	0	0	1	1	1	1	1	1	28794	156	3	79	0	**3060**	67	234
V7	0	0	0	0	0	0	1	1	1	1	1	28779	156	0	79	0	**3064**	4	234
V8	0	0	0	0	0	0	0	1	1	1	1	28159	158	2	79	0	**3068**	4	227
V9	0	0	0	0	0	0	0	0	1	1	1	27919	159	1	80	1	**3069**	1	229
V10	0	0	0	0	0	0	0	0	0	1	1	27435	161	2	80	0	**3086**	17	224
V11	0	0	0	0	0	0	0	0	0	0	1	26963	162	1	80	0	**3088**	2	222

Results: Uncovering duplicates

	case composition											number of infected							
	T.id	T	I	R.a	R.g	R.u	R?	R??	RC	RD	R	#vars	srvs	Δ	cntrs	Δ	#dups	Δ	~vars
V1	1	1	1	1	1	1	1	1	1	1	1	32428	2		2		72		203
V2	0	1	1	1	1	1	1	1	1	1	1	32332	90	88	60	58	1342	1270	245
V3	0	0	1	1	1	1	1	1	1	1	1	30397	116	26	71	11	2371	1029	237
V4	0	0	0	1	1	1	1	1	1	1	1	29056	143	27	76	5	2868	497	242
V5	0	0	0	0	1	1	1	1	1	1	1	28886	153	10	79	3	2993	125	238
V6	0	0	0	0	0	1	1	1	1	1	1	28794	156	3	79	0	3060	67	234
V7	0	0	0	0	0	0	1	1	1	1	1	28779	156	0	79	0	3064	4	234
V8	0	0	0	0	0	0	0	1	1	1	1	28159	158	2	79	0	3068	4	227
V9	0	0	0	0	0	0	0	0	1	1	1	27919	159	1	80	1	3069	1	229
V10	0	0	0	0	0	0	0	0	0	1	1	27435	161	2	80	0	3086	17	224
V11	0	0	0	0	0	0	0	0	0	0	1	26963	162	1	80	0	3088	2	222

Results: Uncovering duplicates

	case composition											number of infected							
	T.id	T	I	R.a	R.g	R.u	R?	R??	RC	RD	R	#vars	srvs	Δ	cntrs	Δ	#dups	Δ	~vars
V1	1	1	1	1	1	1	1	1	1	1	1	32428	2		2		72		203
V2	0	1	1	1	1	1	1	1	1	1	1	32332	90	88	60	58	1342	1270	245
V3	0	0	1	1	1	1	1	1	1	1	1	30397	116	26	71	11	2371	1029	237
V4	0	0	0	1	1	1	1	1	1	1	1	29056	143	27	76	5	2868	497	242
V5	0	0	0	0	1	1	1	1	1	1	1	28886	153	10	79	3	2993	125	238
V6	0	0	0	0	0	1	1	1	1	1	1	28794	156	3	79	0	3060	67	234
V7	0	0	0	0	0	0	1	1	1	1	1	28779	156	0	79	0	3064	4	234
V8	0	0	0	0	0	0	0	1	1	1	1	28159	158	2	79	0	3068	4	227
V9	0	0	0	0	0	0	0	0	1	1	1	27919	159	1	80	1	3069	1	229
V10	0	0	0	0	0	0	0	0	0	1	1	27435	161	2	80	0	3086	17	224
V11	0	0	0	0	0	0	0	0	0	0	1	26963	162	1	80	0	3088	2	222

Results: Uncovering duplicates

	case composition											number of infected							
	T.id	T	I	R.a	R.g	R.u	R?	R??	RC	RD	R	#vars	srvs	Δ	cntrs	Δ	#dups	Δ	~vars
V1	1	1	1	1	1	1	1	1	1	1	1	32428	2		2		72		203
V2	0	1	1	1	1	1	1	1	1	1	1	32332	90	88	60	58	1342	1270	245
V3	0	0	1	1	1	1	1	1	1	1	1	30397	116	26	71	11	2371	1029	237
V4	0	0	0	1	1	1	1	1	1	1	1	29056	143	27	76	5	2868	497	242
V5	0	0	0	0	1	1	1	1	1	1	1	28886	153	10	79	3	2993	125	238
V6	0	0	0	0	0	1	1	1	1	1	1	28794	156	3	79	0	3060	67	234
V7	0	0	0	0	0	0	1	1	1	1	1	28779	156	0	79	0	3064	4	234
V8	0	0	0	0	0	0	0	1	1	1	1	28159	158	2	79	0	3068	4	227
V9	0	0	0	0	0	0	0	0	1	1	1	27919	159	1	80	1	3069	1	229
V10	0	0	0	0	0	0	0	0	0	1	1	27435	161	2	80	0	3086	17	224
V11	0	0	0	0	0	0	0	0	0	0	1	26963	162	1	80	0	3088	2	222

Results: Uncovering duplicates

	case composition											number of infected							
	T.id	T	I	R.a	R.g	R.u	R?	R??	RC	RD	R	#vars	srvs	Δ	cntrs	Δ	#dups	Δ	~vars
V1	1	1	1	1	1	1	1	1	1	1	1	32428	2		2		**72**		203
V2	0	1	1	1	1	1	1	1	1	1	1	32332	90	88	60	58	**1342**	1270	245
V3	0	0	1	1	1	1	1	1	1	1	1	30397	116	26	71	11	**2371**	1029	237
V4	0	0	0	1	1	1	1	1	1	1	1	29056	143	27	76	5	**2868**	497	242
V5	0	0	0	0	1	1	1	1	1	1	1	28886	153	10	79	3	**2993**	125	238
V6	0	0	0	0	0	1	1	1	1	1	1	28794	156	3	79	0	**3060**	67	234
V7	0	0	0	0	0	0	1	1	1	1	1	28779	156	0	79	0	**3064**	4	234
V8	0	0	0	0	0	0	0	1	1	1	1	28159	158	2	79	0	**3068**	4	227
V9	0	0	0	0	0	0	0	0	1	1	1	27919	159	1	80	1	**3069**	1	229
V10	0	0	0	0	0	0	0	0	0	1	1	27435	161	2	80	0	**3086**	17	224
V11	0	0	0	0	0	0	0	0	0	0	1	26963	162	1	80	0	**3088**	2	222

Results: Uncovering duplicates

	case composition											number of infected							
	T.id	T	I	R.a	R.g	R.u	R?	R??	RC	RD	R	#vars	srvs	Δ	cntrs	Δ	#dups	Δ	~vars
V1	1	1	1	1	1	1	1	1	1	1	1	32428	2		2		72		203
V2	0	1	1	1	1	1	1	1	1	1	1	32332	90	88	60	58	1342	1270	245
V3	0	0	1	1	1	1	1	1	1	1	1	30397	116	26	71	11	2371	1029	237
V4	0	0	0	1	1	1	1	1	1	1	1	29056	143	27	76	5	2868	497	242
V5	0	0	0	0	1	1	1	1	1	1	1	28886	153	10	79	3	2993	125	238
V6	0	0	0	0	0	1	1	1	1	1	1	28794	156	3	79	0	3060	67	234
V7	0	0	0	0	0	0	1	1	1	1	1	28779	156	0	79	0	3064	4	234
V8	0	0	0	0	0	0	0	1	1	1	1	28159	158	2	79	0	3068	4	227
V9	0	0	0	0	0	0	0	0	1	1	1	27919	159	1	80	1	3069	1	229
V10	0	0	0	0	0	0	0	0	0	1	1	27435	161	2	80	0	3086	17	224
V11	0	0	0	0	0	0	0	0	0	0	1	26963	162	1	80	0	3088	2	222

Results: Uncovering duplicates

	case composition											number of infected							
	T.id	T	I	R.a	R.g	R.u	R?	R??	RC	RD	R	#vars	srvs	Δ	cntrs	Δ	#dups	Δ	~vars
V1	1	1	1	1	1	1	1	1	1	1	1	32428	2		2		**72**		203
V2	0	1	1	1	1	1	1	1	1	1	1	32332	90	88	60	58	**1342**	1270	245
V3	0	0	1	1	1	1	1	1	1	1	1	30397	116	26	71	11	**2371**	1029	237
V4	0	0	0	1	1	1	1	1	1	1	1	29056	143	27	76	5	**2868**	497	242
V5	0	0	0	0	1	1	1	1	1	1	1	28886	153	10	79	3	**2993**	125	238
V6	0	0	0	0	0	1	1	1	1	1	1	28794	156	3	79	0	**3060**	67	234
V7	0	0	0	0	0	0	1	1	1	1	1	28779	156	0	79	0	**3064**	4	234
V8	0	0	0	0	0	0	0	1	1	1	1	28159	158	2	79	0	**3068**	4	227
V9	0	0	0	0	0	0	0	0	1	1	1	27919	159	1	80	1	**3069**	1	229
V10	0	0	0	0	0	0	0	0	0	1	1	27435	161	2	80	0	**3086**	17	224
V11	0	0	0	0	0	0	0	0	0	0	1	26963	162	1	80	0	**3088**	2	222

Results: Uncovering duplicates

	case composition											number of infected							
	T.id	T	I	R.a	R.g	R.u	R?	R??	RC	RD	R	#vars	srvs	Δ	cntrs	Δ	#dups	Δ	~vars
V1	1	1	1	1	1	1	1	1	1	1	1	32428	2		2		**72**		203
V2	0	1	1	1	1	1	1	1	1	1	1	32332	90	88	60	58	**1342**	1270	245
V3	0	0	1	1	1	1	1	1	1	1	1	30397	116	26	71	11	**2371**	1029	237
V4	0	0	0	1	1	1	1	1	1	1	1	29056	143	27	76	5	**2868**	497	242
V5	0	0	0	0	1	1	1	1	1	1	1	28886	153	10	79	3	**2993**	125	238
V6	0	0	0	0	0	1	1	1	1	1	1	28794	156	3	79	0	**3060**	67	234
V7	0	0	0	0	0	0	1	1	1	1	1	28779	156	0	79	0	**3064**	4	234
V8	0	0	0	0	0	0	0	1	1	1	1	28159	158	2	79	0	**3068**	4	227
V9	0	0	0	0	0	0	0	0	1	1	1	27919	159	1	80	1	**3069**	1	229
V10	0	0	0	0	0	0	0	0	0	1	1	27435	161	2	80	0	**3086**	17	224
V11	0	0	0	0	0	0	0	0	0	0	1	26963	162	1	80	0	**3088**	2	222

Results: Duplicates in SDR 1 survey projects

Survey project*	Number of surveys	Number of countries	Average number of questions	Average sample size	Number of cases	Number of duplicates	Number of affected	
							surveys	countries
ABS	30	13	174	1456	43691	7	3	3
AFB	66	20	210	1499	98942	14	4	4
AMB	92	24	178	1645	151341	24	12	10
ASES	18	18	193	1014	18253	4	1	1
CB	12	3	275	2052	24621	1	1	1
CDCEE	27	16	299	1071	28926	118	3	3
EB†	152	37	342	913	138753	399	11	8
EQLS	93	35	167	1135	105527	20	8	7
ESS	146	32	223	1928	281496	7	5	5
EVS	128	50	347	1301	166502	285	5	5
ISJP	21	14	205	1229	25805	1	1	1
ISSP†	363	53	88	1359	493243	507	31	19
LB	260	19	251	1134	294965	644	32	13
LITS	64	35	636	1060	67866	16	7	7
NBB	18	3	172	1200	21601	1	1	1
PPE7N	7	7	299	2360	16522	26	1	1
WVS	184	89	221	1394	256582	1014	36	31
All projects	**1681**	**137**	**228**	**1329**	**2234636**	**3088**	**162**	**80**

*Survey projects without detected duplicates at V11 level: ARB, CNEP, PA2, PA8NS, VPCPCE. †Only selected waves have been analyzed.

Results: Surveys with extreme duplication

Project/year	Country	Number of cases	Number of variables	Number of duplicates	Proportion of duplicates (%)
ISSP 1998	Bulgaria	1102	88	71	6
EB 19	Belgium	1038	249	74	7
ISSP 2009	Norway	1456	84	107	7
WVS 1	Japan	1204	119	105	9
CDCEE 1	Romania	1234	262	111	9
ISSP 1989	Austria	1997	109	187	9
EB 31	Belgium	1002	377	110	11
EVS 1	United States	2325	328	264	11
WVS 3	Mexico	2364	230	269	11
LB 1996	Panama	1005	253	158	16
WVS 5	South Korea	1200	238	190	16
EB 21	Belgium	1018	138	172	17
WVS 5	Ethiopia	1500	247	275	18
LB 2000	Ecuador	1200	186	398	33

Detection methods: Review

- Blasius & Thiessen 2012
  - Principal Component Analysis
  - Multiple Correspondence Analysis
- Mushtaq 2014
  - Searching for long matching sequences
- Kuriakose & Robbins 2015
  - Estimating the risk of a data set containing duplicates (Gumbel distribution)

References

- Blasius & Thiessen (2012) *Assessing the Quality of Survey Data*. London: Sage
- Kuriakose & Robbins (2015) „Falsification in Surveys: Detecting Near Duplicate Observations" [working paper]
- Mushtaq (2014) „Detection Techniques Applied". Conference on Curbstoning at Washington Statistical Society

- Slomczynski, Powałko, Krauze (2017) „Non-unique Records in International Survey Projects: The Need for Extending Data Quality Control", Survey Research Methods, 11:1, 1-16

Antipodal cases