Duplicates in Social-Science Surveys

Przemek Powałko

ESRA 2015, Reykjavik (July 14, 2015)

Research supported by the Polish National Science Centre (2012/06/M/HS6/00322) as part of the study *Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling*

Topics

- Introduction. The Harmonization Project.
- Duplicates. Terminology. What is being duplicated?
- Importance of the problem. Statistical effects. A probabilistic model.
- Recognition of the problem. Total Survey Error.
- Duplicate detection methods used in social sciences.
- A new method. The Hamming distance. The Hamming diagram.
- Results. Uncovering duplicates. Surveys with extreme duplication.
- Final remarks. Sources and supplements.
- References.

Introduction

- The Harmonization Project, 2013-2015+
  - A joint venture of The Polish Academy of Sciences and The Ohio State University: *Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling*
- 22 survey projects, 142 countries, 1721 national surveys, a time span of 47 years, over 2.2 milion respondents
  - Afrobarometer, Americas Barometer, Arab Barometer, Asian Barometer, Asia Europe Survey, Caucasus Barometer, Consolidation of Democracy in Central and Eastern Europe, 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 - An Eight Nation Study, Political Action II, Political Participation and Equality in Seven Nations, Values and Political Change in Postcommunist Europe, World Values Survey

†only selected waves

Duplicates. Terminology

- What is a duplicate?
  - Strictly: the additional instance of an item, indistinguishable from it
  - What is the item? What is being duplicated?
    - A whole case
    - A response pattern
- Problems with the correct understanding of the term
  - A duplicate being a copy suggests that it is a copy of the original; but *is* it?
  - The term „duplicate” suggests that we might drop the additional copy; but *can* we?
- Complete vs. near duplicates
- Counting duplicates

Importance of the problem

- Duplicates as a nuisance
  - Folk knowledge: who cares?
- Duplicates as a procedural mistake
- Duplicates as a brazen form of cheating
  - Fabricating survey data
  - Who is to blame: the interviewer, data entry person, or data supervisor?
- Confidence in survey data
  - Duplicates reveal severe deficiencies in institutional quality control despite codified and well-known good practices

Importance of the problem: Statistical effects

- Simulations of heavy duplication (Sarracino 2014)
  - Up to 50% of cases were duplicated 2-5 times around the mean value, at the distribution tails, or at random
  - The more duplicates, the greater bias of the regression coefficient
  - The significance increases
  - A dummy variable seems not to improve the model
- Real survey data can also be heavily duplicated
  - Example: WVS 5 South Korea, linear regression, *Interest in politics* vs *Interest in interview (*along with *Sex, Age,* and *Education)*
  - A comparison between models: all cases included vs. all suspected cases excluded: beta coeff. **0.11 → 0.06** while significance coeff. **0.00 → 0.08**

Importance of the problem: A probabilistic model

- The likelihood of duplication
  - The Birthday paradox: How many persons are needed in order to find two persons having an identical birthday with the probability of 0.5? (The answer: 23)
- 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

Recognition of the problem

- Though the duplication problem is known to the scientific world, in social sciences it has received little attention thus far
  - No systematic research has been done in social-science surveys
  - Infrequent reports in the literature
- The Total Survey Error (TSE) framework mentions the multiplicity problem in the sampling frame and suggests
  - Removal of duplicates in advance, or
  - Weighting by the reciprocal of the case's multiplicity
- Linkage analysis develops techniques for matching data coming from diverse sources

Duplicate detection methods

- Blasius & Thiessen 2012
  - Interest in techniques of data screening
  - Small subsets of substantive variables
  - Principal Component Analysis, Multiple Correspondence Analysis
- Mushtaq 2014
  - Interest in techniques of detecting the falsification of data
  - Large sets of variables
  - Searching for long matching sequences
- Kuriakose & Robbins 2015
  - Interest in estimating the risk of a data set containing duplicates
  - Substantive/attitudinal variables
  - The Gumbel distribution

A new method: The Hamming distance

- Distance between records/response patterns can be measured in various ways; we have chosen the **Hamming distance**

CASE#	VAR1	VAR2	VAR3	VAR4
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)

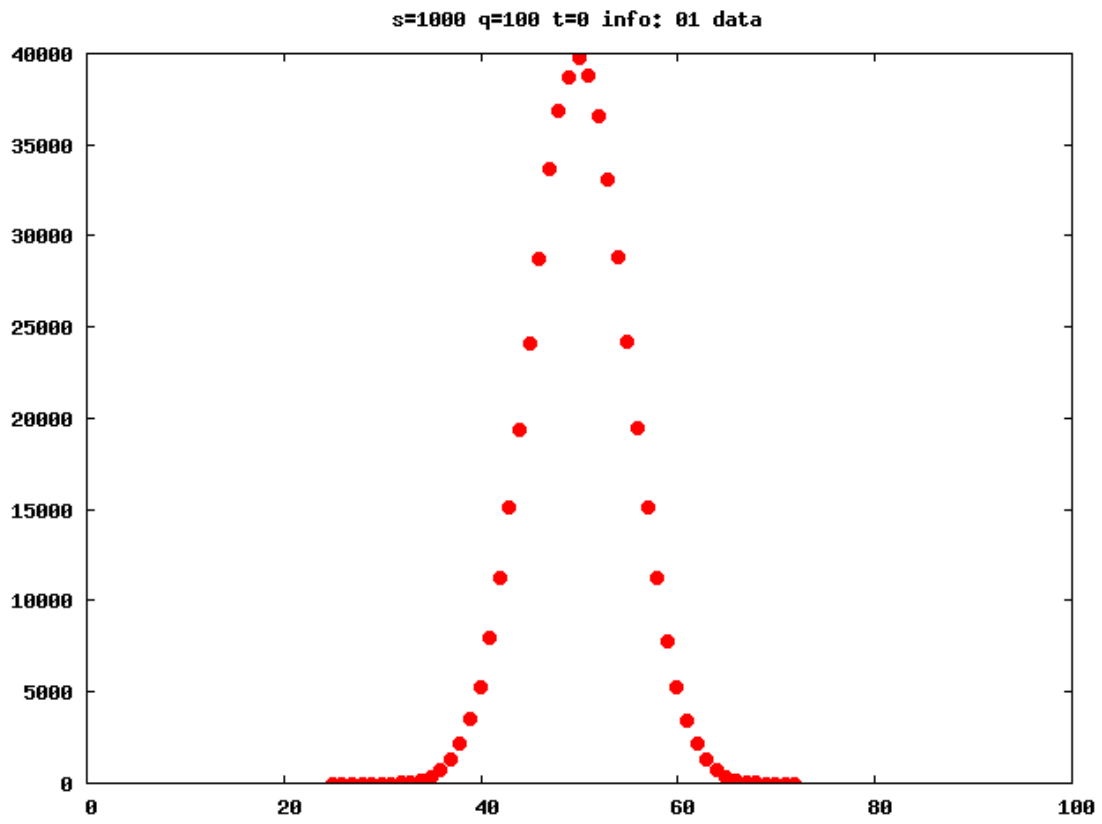
A new method: The Hamming diagram

- The distribution of all pairs of records sharing a Hamming distance
- Probability density function → the **Hamming diagram**
- The graphical presentation shows the overall diversity of data for each survey and thus facilitates the detection of unusual/improbable cases

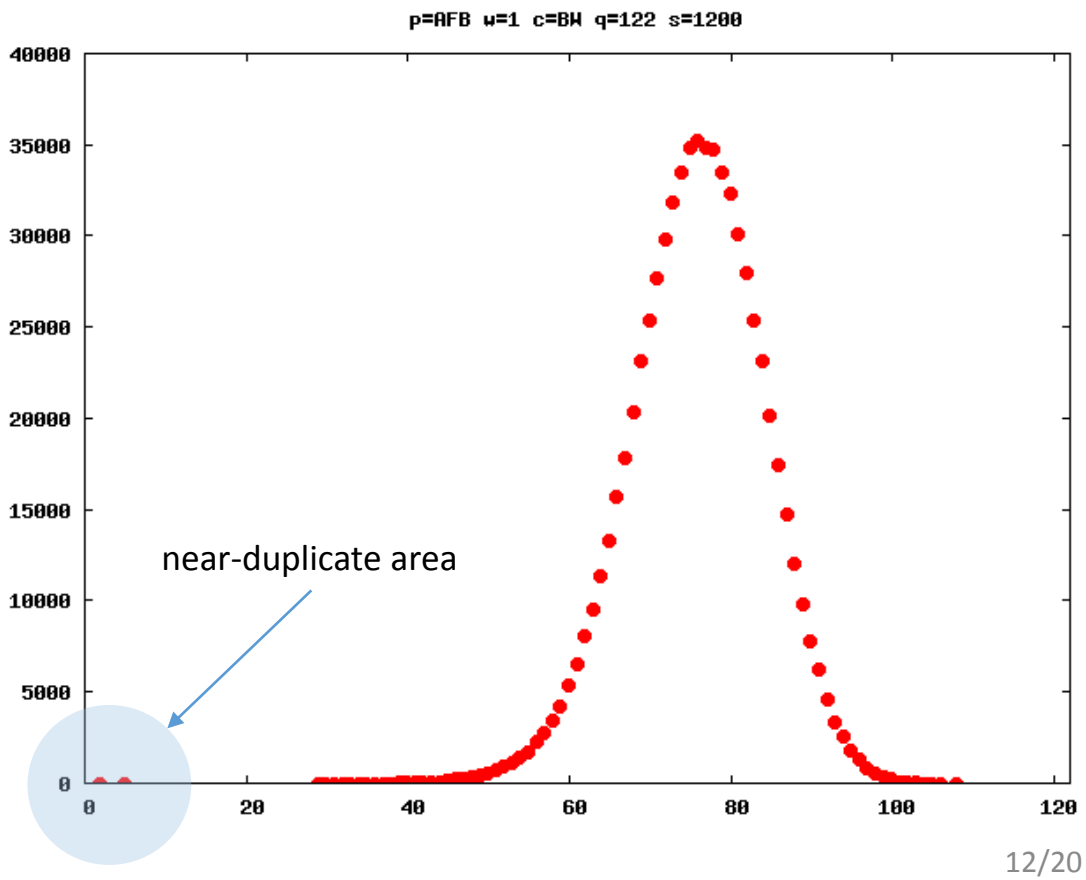
- We constructed a Hamming diagram for each of 1721 surveys
- The binomial density function approximates the Hamming diagram for almost all surveys

A new method: Visualization

The Hamming diagram for a simulated data set



The Hamming diagram for a real data set



Results: Sequential uncovering of duplicates

- From each survey data set sequentially remove the following blocks of variables:
  - Original respondent/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.h1, R.h2)
  - Other variables derived (R.d) or calculated (R.c) 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: Sequential uncovering of duplicates

	blocks of variables†													
	T.id	T	I	R.a	R.g	R.u	R.h1	R.h2	RC	RD	R	#surveys	#patterns	#duplicates
V1	1	1	1	1	1	1	1	1	1	1	1	2	63	72
V2	0	1	1	1	1	1	1	1	1	1	1	90	1243	1342
V3	0	0	1	1	1	1	1	1	1	1	1	116	2238	2371
V4	0	0	0	1	1	1	1	1	1	1	1	143	2659	2868
V5	0	0	0	0	1	1	1	1	1	1	1	153	2751	2993
V6	0	0	0	0	0	1	1	1	1	1	1	156	2781	3060
V7	0	0	0	0	0	0	1	1	1	1	1	156	2783	3064
V8	0	0	0	0	0	0	0	1	1	1	1	158	2787	3068
V9	0	0	0	0	0	0	0	0	1	1	1	159	2788	3069
V10	0	0	0	0	0	0	0	0	0	1	1	161	2803	3086
V11	0	0	0	0	0	0	0	0	0	0	1	162	2805	3088

† The block is included=1 or excluded=0 from the set of variables

Results: Duplicates in projects, surveys, and countries

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**

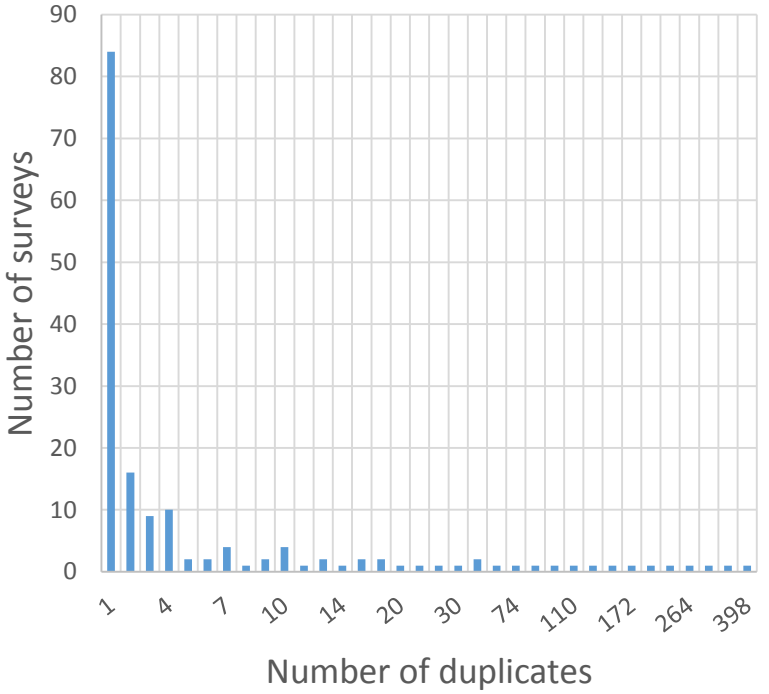
*Survey projects without detected duplicates at V11 level: ARB, CNEP, PA2, PA8NS, VPCPCE. †Only selected waves have been analyzed.

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

Results: A typology of surveys in terms of duplicates

- 84 surveys with a single duplicate, 16 surveys with two duplicates, etc. (see the diagram)
- 67 occurrences of „empty” cases (i.e. patterns containing only missing values)
- Sometimes (rarely) whole cases are not unique
- Respondent/case IDs provided in some data files are not unique



Final remarks: Questions and suggestions

- What to do with „bad” duplicates: delete or retain?
  - If delete, *which* case? Especially, if gender/age are different in „copies”
  - In the case of the worst surveys we *can* delete duplicates; however, should we trust the remaining data?
  - The impact on post-stratification weights: shall we recalculate them?
- Duplicates as a component of error measurement
- Notify the principal investigators, insisting that the published data be preserved for possible future replication
  - Alerts and patches as a recommended solution
- Lesson learned: Screen your data before starting a substantial analysis

Final remarks: Sources and supplements

- Papers and research results from The Harmonization Project are disseminated through
  - http://dataharmonization.org
  - http://dataharmonization.org/newsletter
  - http://consirt.osu.edu/working-papers-series
- Documentation and other materials needed for replication of the presented study are shared on Dataverse
  - https://dataverse.harvard.edu/dataverse/duprecords

References

- Blasius & Thiessen (2012) *Assessing the Quality of Survey Data*
- Feller (1968) *An Introduction to Probability Theory and Its Applications*
- Kuriakose & Robbins (2015) „Falsification in Surveys: Detecting Near Duplicate Observations”
- Mushtaq (2014, presentation) „Detection Techniques Applied”
- Sarracino (2014, private communication) „Estimation Bias Due to Duplicated Observations”
- Slomczynski, Powałko, Krauze (2015, working paper) „The Large Number of Duplicate Records in International Survey Projects: The Need for Data Quality Control”
- Weisberg (2005) *The Total Survey Error Approach*