Errors in Survey Data

Duplicates in Social Sciences

Przemek Powałko

May 14, 2015

International Workshop on Survey Methodology: Cross-national Survey Harmonization and Analysis: Weights, Data Quality and Multi-level Modeling

Mershon Center for International Security Studies

The Ohio State University

Columbus, OH

Topics

- Quality of data
- Fabricating survey data
- Duplicates. Terminology
- Importance of the problem
- Recognition of the problem
- Duplicate detection methods
- [Our method] Description of the project
- First results
- Uncovering duplicates
- Further results
- The Hamming diagram
- The likelihood of duplication
- Final remarks
- References

Quality of data

- What is quality of data?
- Factors compromising quality of data
  - Recognized and recoverable
  - Recognized and unrecoverable (noise)
  - Unrecognized
- Sources of variance in survey data
  - Substantive
  - Non-substantive
- The goal is to recognize non-substantive variance and to include it in models

Quality of data

- Sources of non-substantive variation
  - The mode effect
  - The interviewer effect
  - Response styles
  - Question misunderstanding
  - Wording/translation issues
  - The order effect
  - The scale effect
  - Missing data (item non-response)
  - Respondent fatigue

Quality of data

- Sources of non-substantive variation, continued
- Errors in data (non-intentional/intentional)
  - Empty/meaningless data entry
  - Typing errors
  - Reversed scales
  - Uncalibrated weights/wrongly used weights
  - Faked interviews
    - Improbable response patterns
    - Fully/partially fabricated data
    - Complete/near duplicates

Quality of data

- Indicator of high-quality data
  - High ratio of the substantive variance to the total variance
  - The goal: minimizing the non-substantive variance
- Is data quality characterized by a high signal-to-noise ratio?
  - Pitfall: duplicates can artificially boost the signal and take away noise

All data are dirty, but some are informative

Blasius & Thiessen 2012

Fabricating survey data

- Rich literature starting from the 1940's
- Folk knowledge: curbstoning
  - „[A] jargon for sitting on the curbstone and filling out forms with made-up information” (Mann 1993)
- Faked interviews
  - Partially faked interviews based on valid ones
  - Can significantly shorten the interview time
  - It is cheaper to copy and paste one interview and slightly change some values
  - Not all variables are equally likely to be modified

Duplicates. Terminology

- What is a duplicate?
  - Strictly: the additional instance of an item, indistinguishable from it
  - What is the item?
- Complete/near duplicates
- Problems with the correct understanding the term
  - Duplicate being a copy suggests that it is a copy of an original item; but *is* it?
  - It suggests we might drop the additional copy; but *can* we?
- Alternative formulation
  - Number of unique response patterns < Number of cases

Duplicates. Terminology

- If a response pattern occurs twice, then there is a duplicate, if it occurs three times, then there are two duplicates, and so on
- Multiplicity of response patterns
  - Expected is the unique
  - Duplicates, triplicates, … (alternatively: doublets, triplets, ... multiplets)
- „Duplicate” is a synonym of any repeated item

- Naming conventions: case, record, observation, response pattern

Importance of the problem

- Duplicates as a nuisance
- Duplicates as an insolent form of cheating
- Checks for duplicates are routine
  - *This must be obvious!*
- Folk knowledge, continued
  - Silence in the audience (*Everybody knows that!*) Nobody cares?
  - *This is for weighting!* (Perforated cards in „analog” times)

Importance of the problem

Przemek: *So, why are we puzzled by this discovery?*

Tad Krauze*: *The sheer number is the answer!*

* Hofstra University, Hempstead, NY

- Duplicates are common and universal

Importance of the problem

- Ethical issues
  - Does a single duplicate decrease confidence in survey data?
- Attribution issues
  - Who is to blame: interviewer, data entry person, or data supervisor?
- „[One duplicate] can be accidental, but when done on a large scale, it is more likely to be the result of an intentional effort to save time and money in the data collection process” (Kuriakose & Robbins 2015)

Importance of the problem

- Duplicates reveal severe deficiencies in institutional quality control despite codified and well-known good practices (see: AAPOR 2003)
- „[F]alsification through duplication of responses from a valid interview [is difficult to detect and can be dangerous] because it produces data that appear to be valid” (Kuriakose & Robbins 2015)

Importance of the problem

- Impact on statistical inference
- Sarracino 2014
  - Are duplicated observations in a data set a problem?
  - Do they bias the coefficients of a regression?
  - In which directions?
  - Can we control for this bias including an indicator of the duplicated observations?
  - To what extent would the inclusion of the control solve the eventual bias?
  - Does the severity of the bias depend on the position of the duplicates in the data set (i.e. close or far from the mean)?

Importance of the problem

- Sarracino 2014, continued
- Simulations: up to 50% cases were duplicated 2-5 times
  - Around the mean value of a target variable
  - At left and right tails of the variable distribution
  - At random
- The author was interested in the size, sign and significance of the coefficient in a linear regression model
- Results
  - The more duplicates the greater bias of the coefficient
  - Its significance increases
  - A dummy variable doesn't improve the model

Importance of the problem

- „[Heavy duplication] artificially increases statistical power and decreases variance, resulting in smaller estimated confidence intervals for point estimates” (Kuriakose & Robbins 2015)
- „[D]uplicates with a higher number of cases provide a larger base sample size for statistical significance testing and solidify the relationships between variables, all while reducing variance (…) boosting signal and taking away noise” (Kuriakose & Robbins 2015)

Recognition of the problem

- Review of literature
- Though duplication problem is known to scientific world, in social sciences it got little attention so far
- Total Survey Error (TSE) and Total Quality Measurement (TQM) frameworks mention duplicates as a nuisance and recommend to delete them
- There are three domains in which problem appears
  - Multiplicity problem
  - Linkage analysis
  - Empirical studies

Recognition of the problem

- Multiplicity problem
  - Arises when a sampling frame contains repeated entries that are associated with the same element of a target population (overcoverage)
  - A method that is sometimes used to deliberately increase the probability of polling rare respondents; cases obtained in this way need to be specifically treated during analysis, e.g. weighted inversely to multiplicity level
- Multiplicity is a problem with an unequal probability of being drawn to the sample

Recognition of the problem

- Linkage analysis
  - Attempts to solve the problem of matching data about the same item from heterogenous databases (as in business and medical sciences)
  - Provides techniques of comparing data entries based on fuzzy logic, i.e. the match of elements is not a 0-1 but a probabilistic game
  - Edit distance
- Linkage analysis assumes that duplicated records are *valid*

Recognition of the problem

- Empirical findings is what interests us most
- Sporadic reports; no systematic research
  - Usually, authors express their astonishment (as we did)

Duplicate detection methods

- Mushtaq 2014
- PAPI mode, 5000+ respondents, ~150 interviewers, 20+ supervisors
- Interested in techniques for detecting data falsification
  - Comparison of response pattern distributions
  - Correlation analysis of response patterns
- Method: Compare each record with all other records and measure the length of the duplicate sequence; then examine unusually long sequences of identical answers (interviews sharing half or more responses in sequence)
  - Clusters cases by interviewers and supervisors and finds interesting differences among their scores

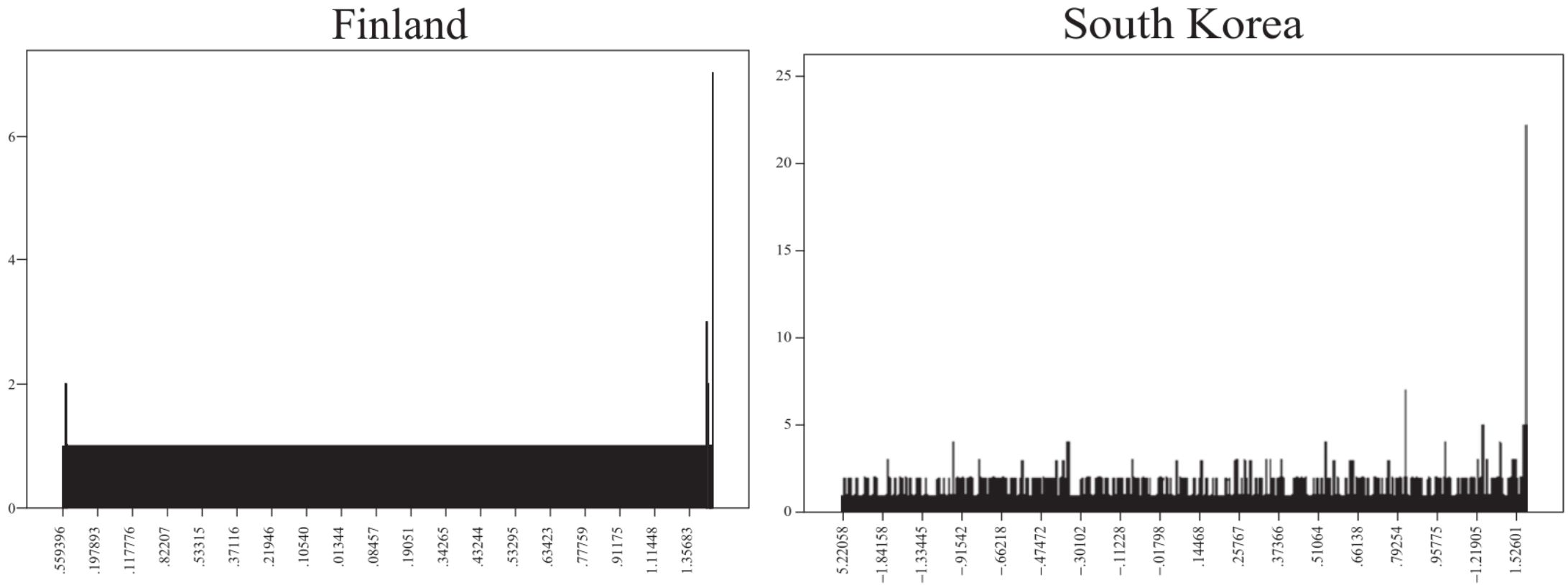
Duplicate detection methods

- Blasius & Thiessen 2012
- Interested in assessing survey data quality in general
- World Values Survey, wave 5
- Screening data: scaling methods
  - MCA (Multiple Correspondence Analysis)
  - PCA (Principal Component Analysis)
  - CatPCA (Categorical Principal Component Analysis)
- Advantages
  - MCA & CatPCA: relatively few assumptions
  - MCA: simple correspondence between distance and dissimilarity

Duplicate detection methods

- Blasius & Thiessen 2012, continued
- First take
  - 10 variables of 10-point scales (questions characterizing democracy)
  - Missing values excluded (listwise case deletion)
  - Method: PCA
- „Factor scales that occur more than once are anomalous from a strict probabilistic point of view, given the large number of possible patterns”

Duplicate detection methods



Duplicate detection methods

- Blasius & Thiessen 2012, continued
- Second take
  - 36 variables of various scales (questions on different topics)
  - Missing values included
  - Method: MCA
- „With such a large set of heterogeneous and largely uncorrelated items, we did not expect to find any duplicate values”
- Actually, they found many of them!

Duplicate detection methods

Table 4.3 Calculation of duplicates for selected countries

k	South Korea		Indonesia		Thailand		India		Moldova		Ethiopia	
	f_1	f_2	f_1	f_2	f_1	f_2	f_1	f_2	f_1	f_2	f_1	f_2
1	671	671	1922	1922	1454	1454	1659	1659	998	998	835	835
2	219	438	26	52	32	64	85	170	24	48	254	508
3	23	69	7	21	1	3	24	73			16	48
4	4	16	3	12			7	28			7	28
5					1	5	3	15			4	20
6	1	6					2	12			1	6
7							2	14			2	14
8			1	8	1	8	1	8				
10											2	20
11							2	22				
21											1	21
N	918	1200	1959	2015	1489	1534	1785	2001	1022	1046	1122	1500
Fakes		282		56		45		216		24		378

Key: *k* = frequency of occurrence; f_1 = number of instances; f_2 = number of cases represented. The number of duplicates or fakes is obtained by subtracting the *N* in f_1 from the *N* in f_2.

Duplicate detection methods

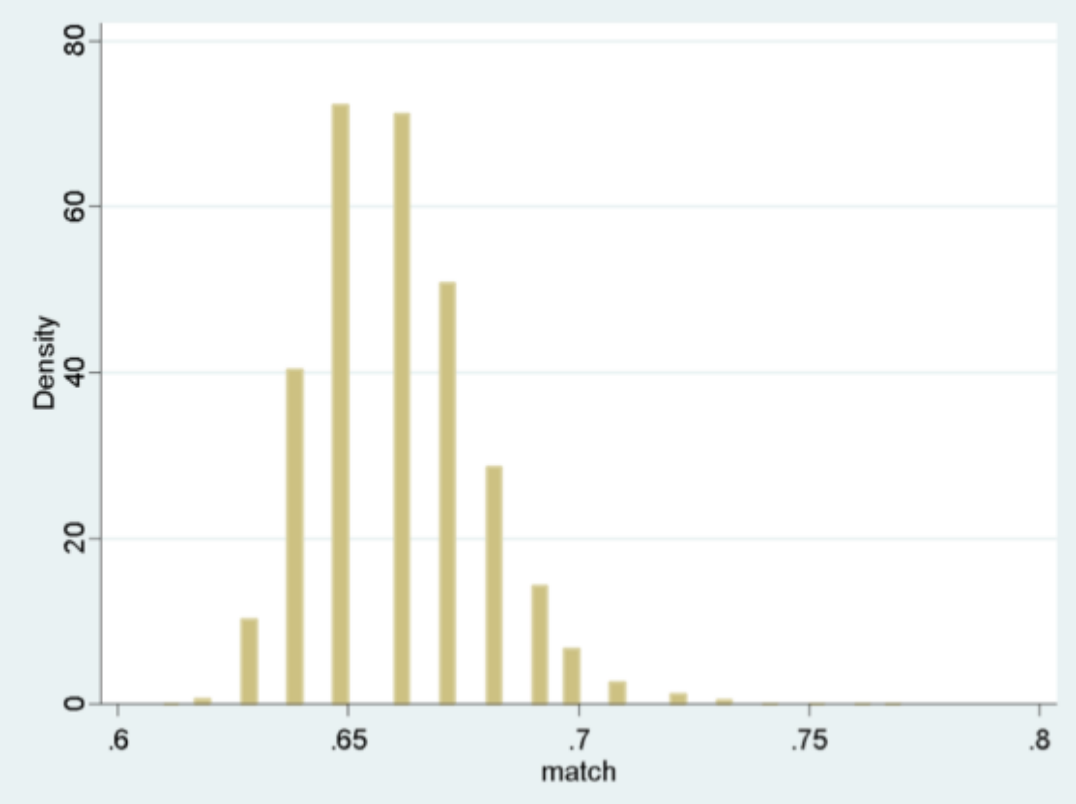
- Kuriakose & Robbins 2015
- Interested in near duplicates
- 11 survey projects, 604 national surveys, time span 35+ years
  - Afrobarometer, Arab Barometer, Asia Barometer, Asian Barometer, European Social Survey, Eurobarometer, Latin America Public Opinion Project (LAPOP, probably Americas Barometer), Pew Global Attitudes Project, Pew Religion Project, Sadat Chair at The University of Maryland, World Values Survey

Duplicate detection methods

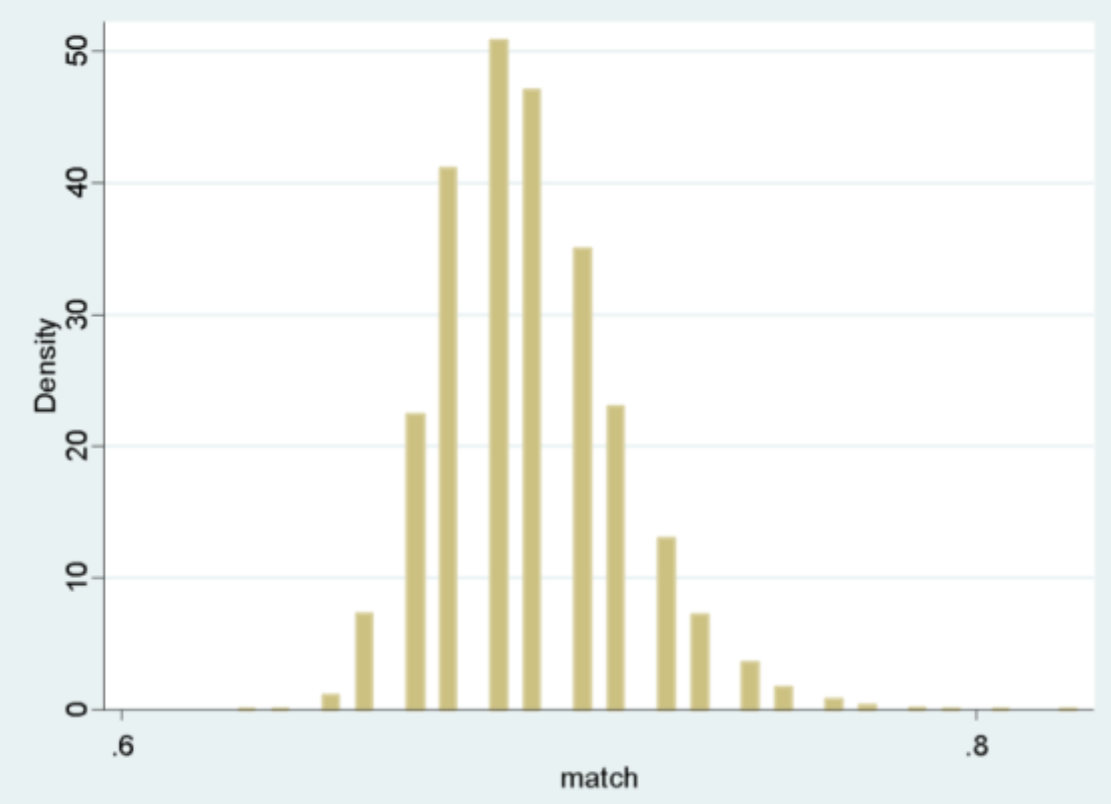
- Kuriakose & Robbins 2015, continued
- Method
  - Choose only substantive/attitudinal variables
  - Compare each case with all other cases
  - Determine the maximum percentage of variables that the case shares with any other case in the data set
- Mass function of the maximum percentage match yields a Gumbel distribution

Duplicate detection methods

Result of a simulation
for randomized data



Result of a simulation
for correlated data

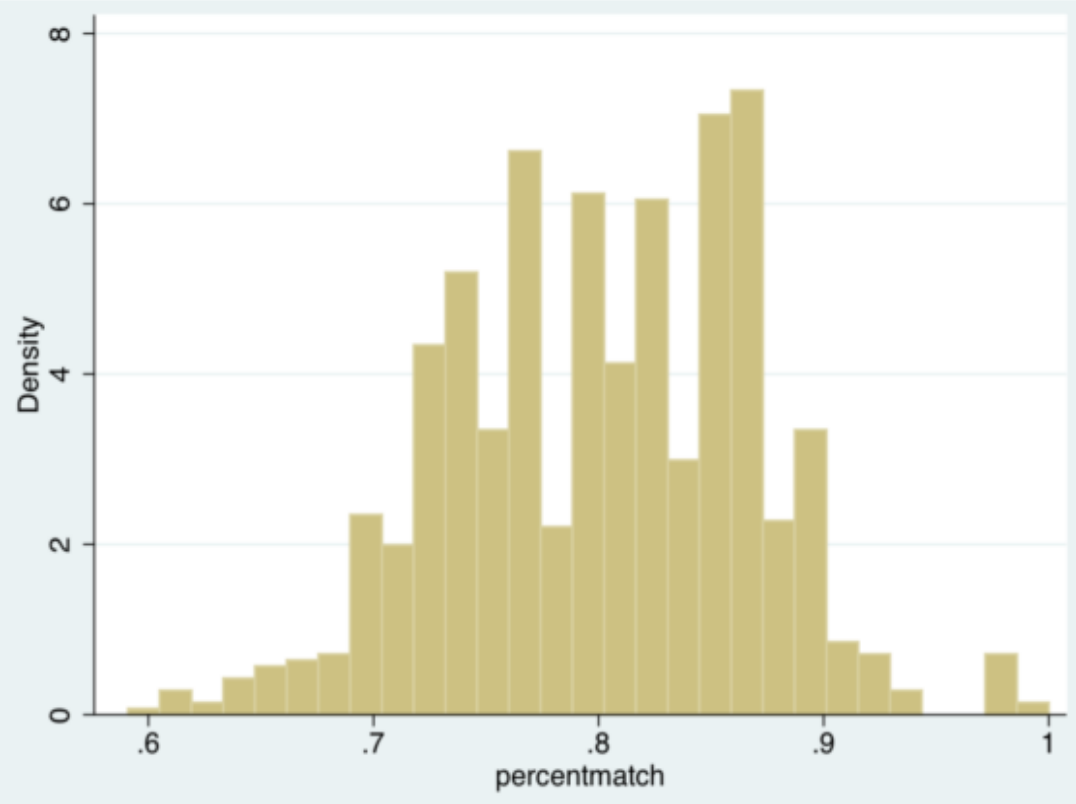


Duplicate detection methods

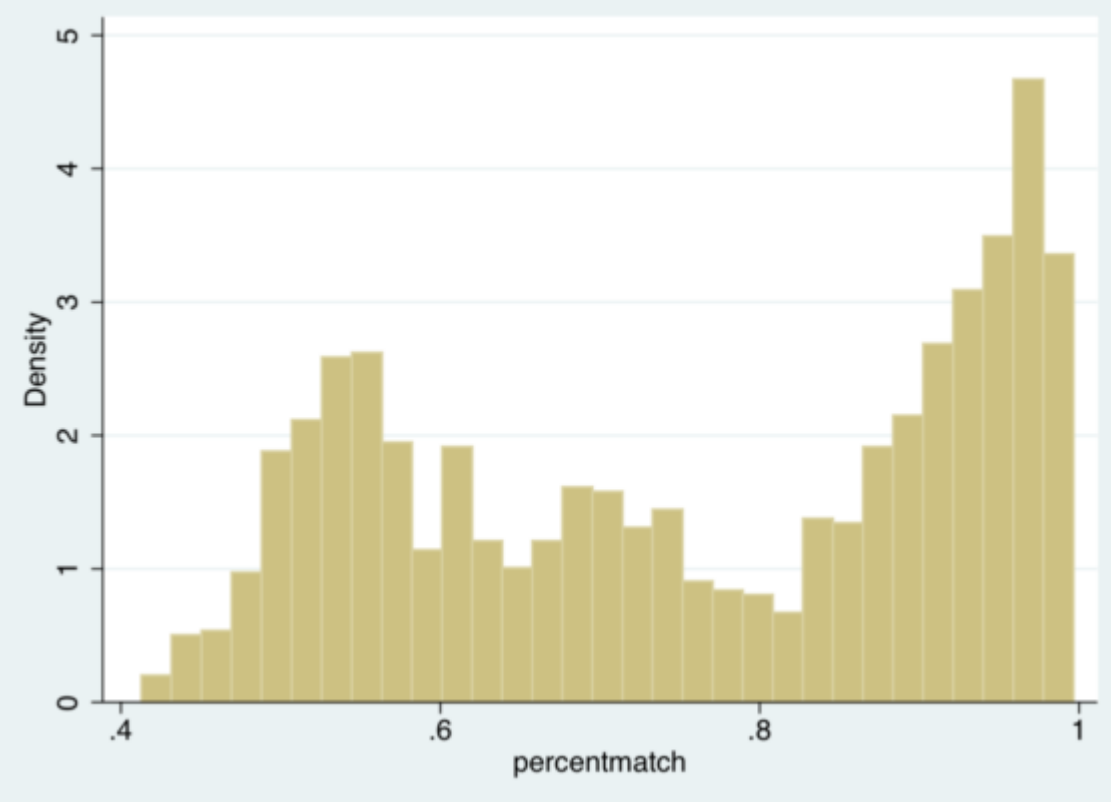
- Kuriakose & Robbins 2015, continued
- Authors assess distributions using two criteria
  - Is the distribution monotonic on each side of the mode?
  - Are there fewer than 5% of observations where the maximum percent match exceeds 0.9?
- Risk of data set containing duplicates
  - Moderate – if one of the criteria fail
  - High – if both criteria fail

Duplicate detection methods

Result for a data set with a moderate risk of having duplications



Result for a data set with a high risk of having duplications



Duplicate detection methods

- Kuriakose & Robbins 2015, continued
- Results
  - Moderate risk – 19% surveys
  - High risk – 16% surveys

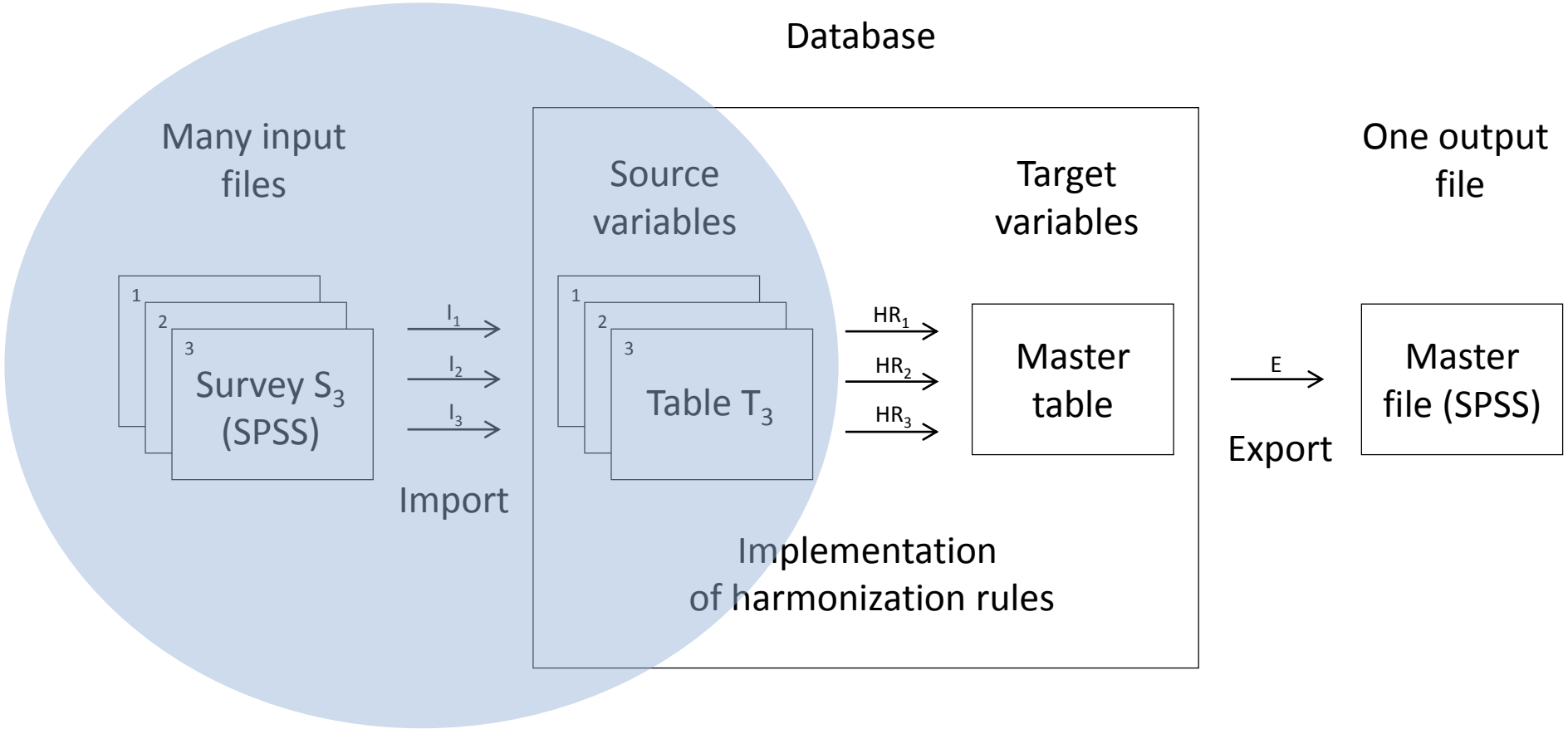
- STATA program (*percentmatch*) is offered for quickly estimating the risk the survey data contain duplicates

[Our method] Description of the project

- The Harmonization Project, 2013-2015+
  - 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, 1721 national surveys, time span 47 years, over 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 II, Political Action - An Eight Nation Study, Political Participation and Equality in Seven Nations, Values and Political Change in Postcommunist Europe, World Values Survey

†only selected waves

[Our method] Description of the project



[Our method] First results

- Serendipitous discovery
  - Verifying candidate variables in data files to be primary keys in corresponding database tables
- Respondent/case IDs provided in some data files are not unique
- Infrequently, whole cases are not unique

- We have found
  - 3088 duplicates
  - 162 national surveys (*common*)
  - 80 countries (*universal*)

[Our method] First results

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**

[†]only selected waves

Uncovering duplicates

- What exactly have we done?
  - Compared each case with all others
  - Chose variables (ideally, covering all questionnaire items)
  - Determined (the Hamming) distance between cases
  - Duplicates are cases with zero-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

Uncovering duplicates

- From each survey data set remove:
  - 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 repondent's responses
- At each step observe uncovering duplicates in remaining response patterns

Uncovering duplicates

	blocks of variables†												non-unique					
	T.id	T	I	R.a	R.g	R.u	R.h1	R.h2	RC	RD	R	srvs	pttrns	Δ	#dups	Δ	occrs	Δ
V1	1	1	1	1	1	1	1	1	1	1	1	2	63		72		135	
V2	0	1	1	1	1	1	1	1	1	1	1	90	1243	1180	1342	1270	2585	2450
V3	0	0	1	1	1	1	1	1	1	1	1	116	2238	995	2371	1029	4609	2024
V4	0	0	0	1	1	1	1	1	1	1	1	143	2659	421	2868	497	5527	918
V5	0	0	0	0	1	1	1	1	1	1	1	153	2751	92	2993	125	5744	217
V6	0	0	0	0	0	1	1	1	1	1	1	156	2781	30	3060	67	5841	97
V7	0	0	0	0	0	0	1	1	1	1	1	156	2783	2	3064	4	5847	6
V8	0	0	0	0	0	0	0	1	1	1	1	158	2787	4	3068	4	5855	8
V9	0	0	0	0	0	0	0	0	1	1	1	159	2788	1	3069	1	5857	2
V10	0	0	0	0	0	0	0	0	0	1	1	161	2803	15	3086	17	5889	32
V11	0	0	0	0	0	0	0	0	0	0	1	162	2805	2	3088	2	5893	4

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

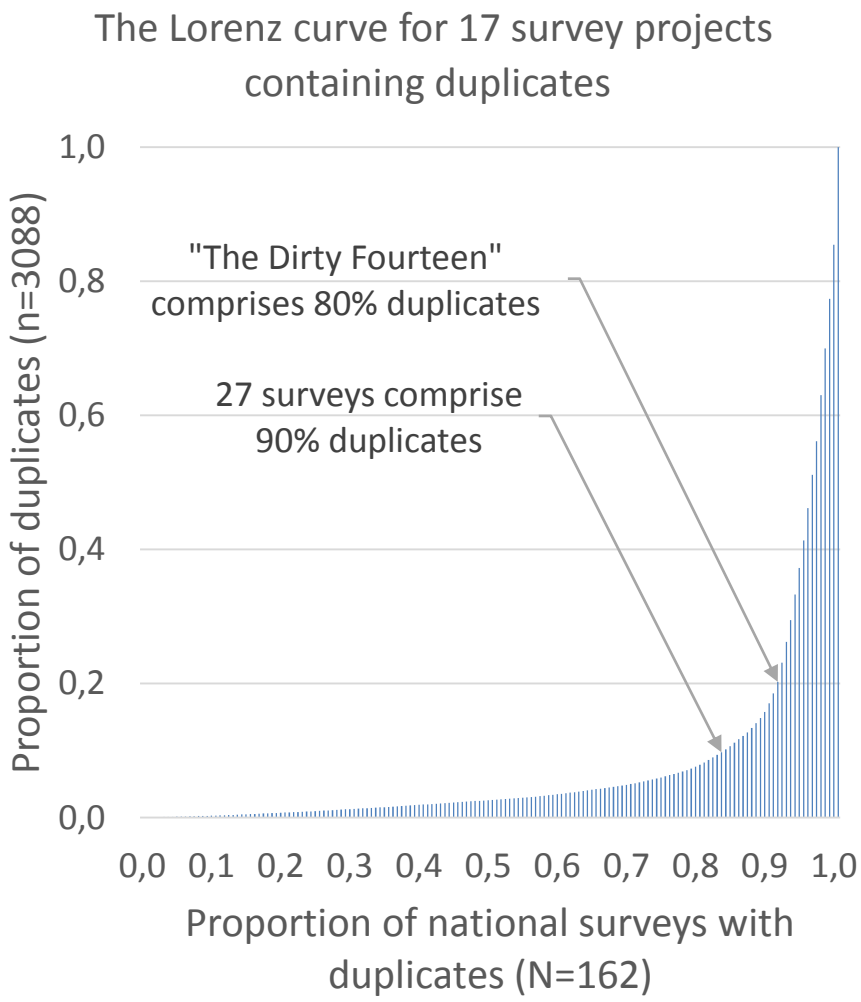
Number of non-unique patterns + Number of duplicates = Number of occurrences

Further results [„The Dirty Fourteen”]

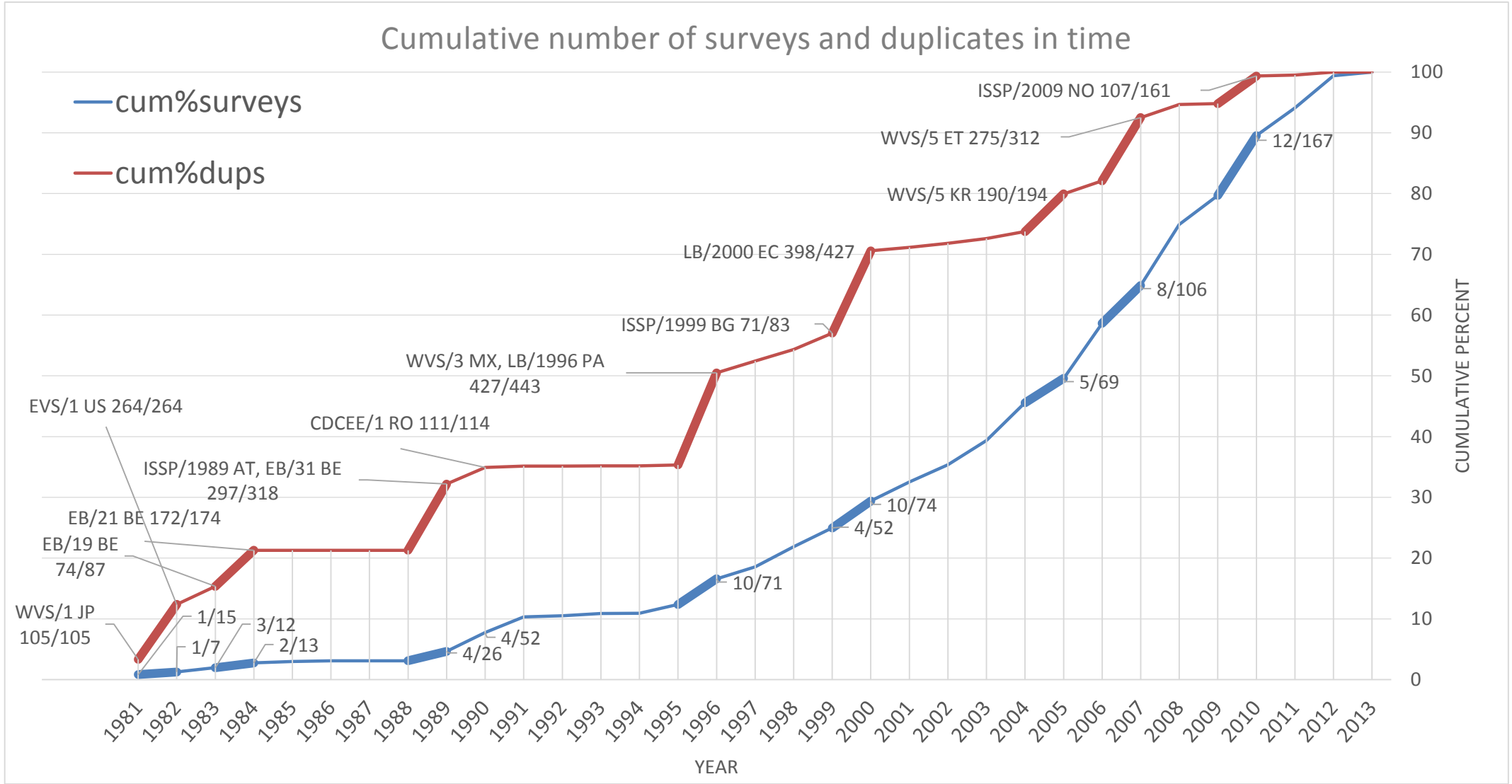
Project/year	Country	Number of cases	Number of variables	Number of duplicates
ISSP 1998	Bulgaria	1102	88	71
EB 19	Belgium	1038	249	74
ISSP 2009	Norway	1456	84	107
WVS 1	Japan	1204	119	105
CDCEE 1	Romania	1234	262	111
ISSP 1989	Austria	1997	109	187
EB 31	Belgium	1002	377	110
EVS 1	United States	2325	328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

Ordered by percentage of duplicates in a sample

Further results [The Lorenz curve]

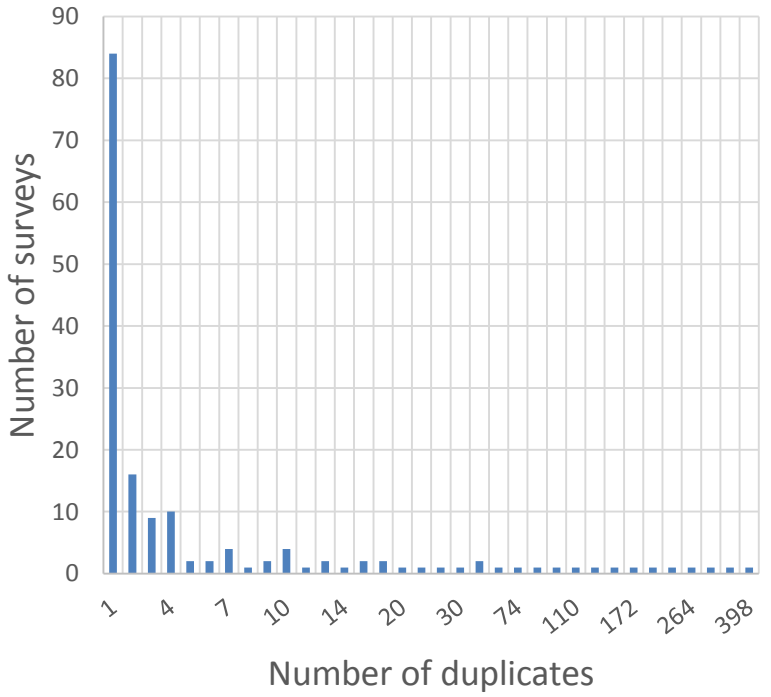


Further results [Cumulative]



Further results [Quantitative]

- 67 empty non-unique response patterns (i.e. patterns containing only missing values)
- Typology of surveys based on the number of duplicates: 84 surveys with a single duplicate, 16 surveys with two duplicates, etc



Further results [Quantitative]

- Excess number of duplicates in modulo-100 samples: 26% for all surveys vs 32% for those containing duplicates
- No dependence between the number of duplicates and fieldwork control (test chi2)

- We can provide IDs for all problematic cases

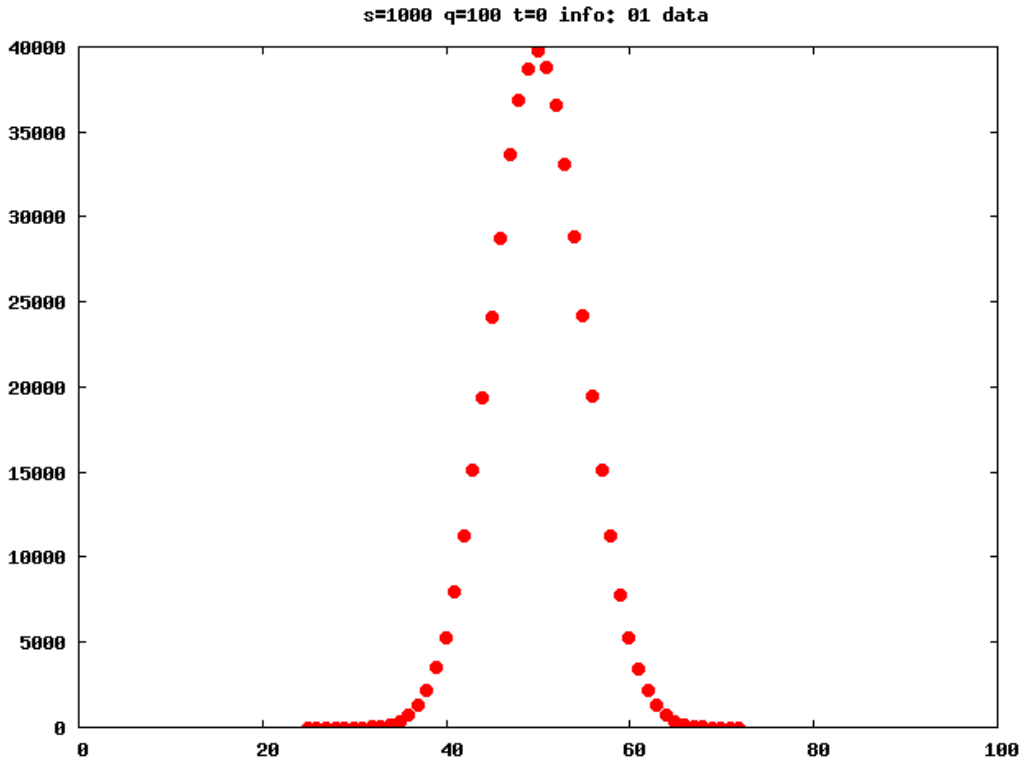
The Hamming diagram

- Distance between records can be measured in various ways, we have chosen the Hamming distance
- Distribution of all pairs of records sharing a Hamming distance
  - Probability mass function → The Hamming diagram
- The graphical presentation shows the overall quality of data for each survey and thus facilitates the detection of dirty data

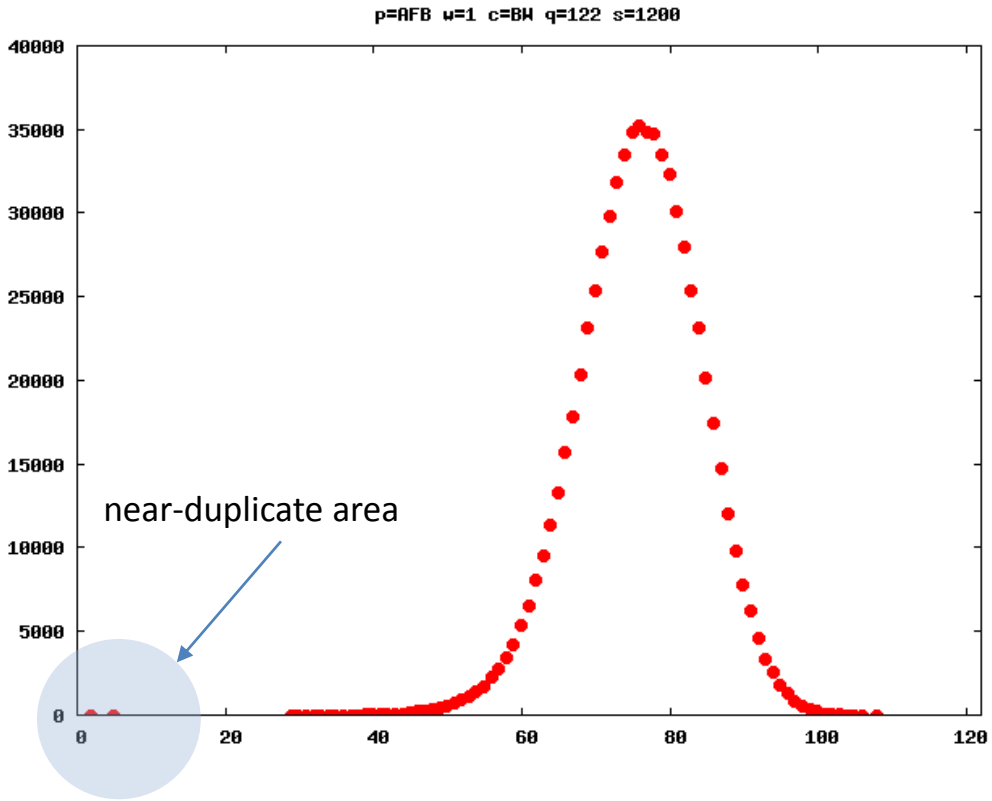
- We constructed the Hamming diagram for each of 1721 surveys

The Hamming diagram

Simulated data set

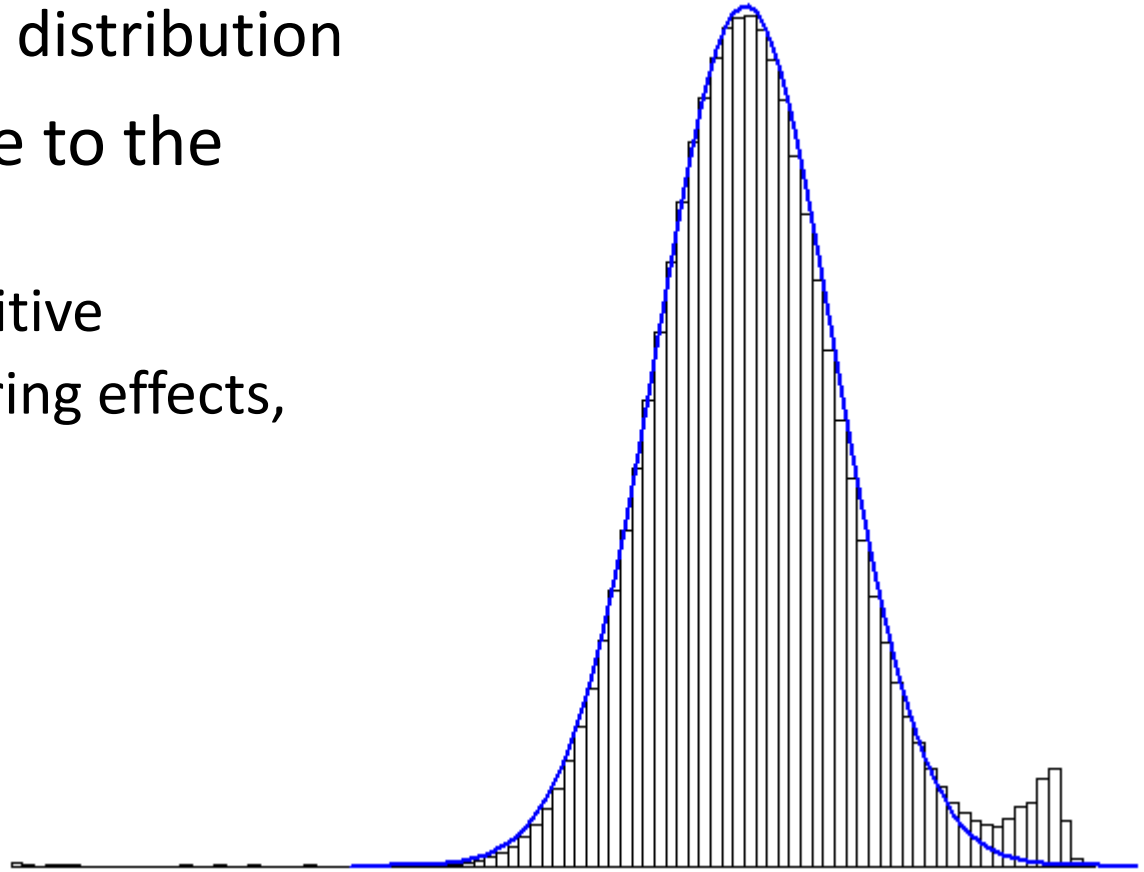


Real data set, example



The Hamming diagram

- One can prove that Hamming diagram for independent variables is described by a binomial distribution
- Real survey data usually converge to the binomial distribution
  - Kolmogorov-Smirnov tests are positive
  - Any discrepancies come from filtering effects, missing values, etc



The likelihood of duplication

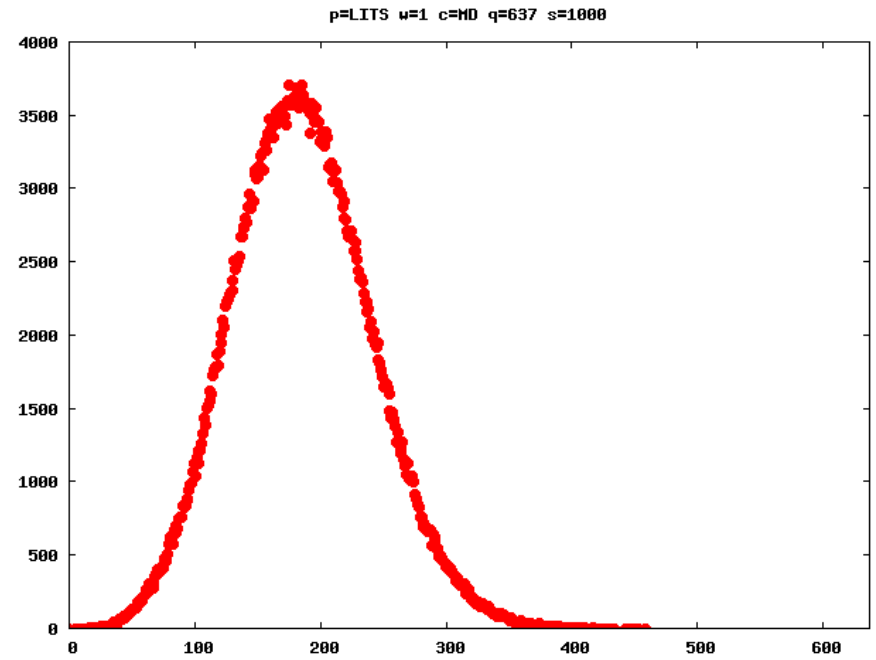
- In the literature we haven't found any attempts to estimate the likelihood of a duplicate occurrence, just expressions such as „it has to be very small”
- Common-sense argument: uniqueness of individuals
- Empirical argument (after Converse 1964 and Zaller 1992): „It is very unlikely that even if the same person took a survey twice that he or she would provide the identical set of answers to all questions; by extension, even if two individuals are highly similar in backgrounds and view, the likelihood of them providing exactly the same responses to a lengthy survey is infinitesimally small” (Kuriakose & Robbins 2015)

The likelihood of duplication

- Combinatorial method
  - Birthday paradox: how many persons are needed in order to find two persons having an identical birthday with the probability 50%? (The answer: 23)
- Probabilistic model
  - Dichotomous variables (a very conservative assumption)
  - 1/3 variables are mutually independent
- Results for probability 1%
  - 90 variables (30 mutually independent) → 4,646 respondents needed
  - 120 variables (40 mutually independent) → 148 thousands respondents needed
  - 150 variables (50 mutually independent) → 475 millions respondents needed
  - 240 variables (80 mutually independent) → 155 billions respondents needed

The likelihood of duplication

- Plausible reasoning: since the majority of Hamming diagrams fit the binomial distribution, this suggests the independence of the majority of variables
- Pitfalls
  - The low number of questions in surveys
  - The idiosyncrasies of some surveys
- An accurate model should be built for each survey independently



Final remarks

- Three possibilities
  - Both duplicated cases are real (coincidence)
  - One is real and another is faked
  - Both are fakes - no real respondent exists
- One may try to distinguish falsifications done by individual interviewers in the field or by the firm while compiling data
  - Clustering by interviewer ID
  - Use of paradata

Final remarks

- What to do with duplicates: delete or retain?
  - If delete, which ones (especially if gender/age are different)?
  - In the worst surveys we can delete duplicates; however, do we trust the remaining data?
  - Impact on post-stratification weights: shall we recalculate them?
- Notify the principal investigators, insisting that the published data should be preserved for possible future replication
  - Alerts and patches as a recommended solution
- Publish your data!
  - Transparency as the key to data quality
  - Dataverse

Final remarks

- Possible impact on scientific results published so far
- Estimates for all 22 survey projects
  - 11,000+ papers based on information from the projects' web pages
  - 25,000+ papers at Google Scholar
  - 2,000 papers and ~20,000 citations at Web of Science Core Collection
- Lesson learned: screen your data before starting a substantial analysis! (Look before you jump)

References

- AAPOR (2003) Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection and Repair of Its Effects
- Blasius & Thiessen (2012) „Assessing the Quality of Survey Data”
- Bredl, Storfinger, Menold (2011) A Literature Review of Methods to Detect Fabricated Survey Data
- Converse (1964) „The Nature of Belief Systems in Mass Publics”
- Elmagarmid, Ipeirotis, Verykios (2007) Duplicate Record Detection. A Survey
- Feller (1968) „An Introduction to Probability Theory and Its Applications”, vol.1

References

- Groves (1989) „Survey Errors and Survey Costs”
- King (1995) Replication, Replication
- Kreuter (2013) „Improving Surveys with Paradata”
- Kuriakose & Robbins (2015) Falsification in Surveys: Detecting Near Duplicate Observations
- Lessler & Kalsbeek (1992) „Nonsampling Error in Surveys”
- Mann (1993) How Many Is Too Many?
- Mushtaq (2014) Detection Techniques Applied
- Sarracino (2014) Estimation Bias Due to Duplicated Observations

References

- Schräpler & Wagner (2005) Characteristics and Impact of Faked Interviews in Surveys
- Slomczynski, Powałko, Krauze (2015, in preparation) The Surprising Number of Duplicate Records in International Survey Projects
- Smith (2011) Refining the Total Survey Error Perspective
- Waller (2013) Interviewing the Surveyors: Factors which Contribute to Questionnaire Falsification (Curbstoning) among Jamaican Field Surveyors
- Weisberg (2005) The Total Survey Error Approach
- Zaller (1992) „The Nature and Origins of Mass Opinion”