Kenya Education Statistical Disclosure Control Process

External Report

1. Disclosure risk and confidentiality protection

Microdata often contain confidential or sensitive information, which makes release of these datasets in their original form impossible. Release of the data could reveal this confidential information and lead to a breach of privacy of the respondents. This has ethical and in many cases legal objections. Furthermore, when confidentiality is not guaranteed, current and potential future respondents are less likely to be willing to respond in future surveys.

Statistical Disclosure Control (SDC) refers to (i) a set of methods to measure the risk in a dataset that confidential information may be disclosed when releasing a dataset and (ii) to the set of methods to treat the data in order to prevent the release of confidential information when releasing the dataset. SDC is used in many statistical offices to anonymize data before release. Templ et al. (2014) give a concise introduction to SDC. We refer to Hundepool et al. (2012) and the references therein for a complete overview of the SDC process and a detailed description of the methods.

In this report we describe the methods applied to mitigate this risk, with the emphasis on the implications for the user. Anonymization leads to changing the dataset.

2. SDI Kenya education data set

The SDI Kenya education dataset contains information on a sample of 306 schools from a total of about 27,900 schools in 7 regions. The statistical objects in this dataset are schools, teachers and pupils. The main concern for re-identification and confidentiality are the teachers, and, to a lesser degree, pupils. However, since the data are hierarchical, i.e. teachers and pupils belong to schools, the re-identification of a school might lead to the re-identification of a teacher or pupil too.

The datasets consists of the following modules that contain information on:

- Module 1: School information
- Module 2: Teacher roster
- Module 3: Financial information
- Module 4: Classroom observation
- Module 5: Pupil assessment
- Module 6: Teacher assessment

3. Risk before anonymization

Risk in the SDC context is the probability or likelihood that disclosure by an (hypothetical) intruder of a record occurs. Disclosure can be **identity disclosure**, when the identity of

an individual or entity in the dataset is correctly revealed, or **attribute disclosure**, when the intruder gains new (confidential) information from the dataset. Identity disclosure can imply attribute disclosure. The risk is dependent on several factors, amongst others the frequency of **keys** (i.e. combinations of values of key variables), sample size and sampling weights as well as the availability of external information to intruders to use for re-identification. The disclosure scenarios for a particular dataset describe these parameters and the way an intruder can use a dataset to gain new information.

The acceptable level of risk depends on the release type, e.g. scientific use file (SUF), public use file (PUF), or other ways of release and the sensitivity of the data. This file is released as PUF and hence needs a higher level of protection. Also, the potential harm caused by disclosure should be taken into account when determining the acceptable risk level.

In similar microdata releases with for instance business survey data, the geographical level is highly reduced, large companies are suppressed and the level of detail in the data is reduced to protect the records. Generally, the period between the survey and data release is also specified, e.g. 1 year. In case of the SDI education survey, the period between the survey and the data release introduces already uncertainty into several variables. It should be noted that a complete elimination of disclosure risk is not possible.

For categorical variables, risk measures are based on the frequency count of **keys**, the combination of values for categorical key variables. Key variables are the categorical variables that are known to the intruder or available in external datasets and can be used for re-identification. Disregarding the sample weight, the higher the frequency of a key, the lower the risk. This principle leads to the idea for the risk measure **k-anonymity**; it is a count of the number of records with keys that are not shared by at least k records. K-anonymity can also be used as a threshold when anonymizing a dataset. If 2-anonymity is violated we speak of a **sample unique**. A sample unique has a unique combination of values for the selected categorical key variables in the dataset and is at high risk of re-identification (depending on the sample weight).

Sample uniques can be further classified as **special uniques**: these are records that do not need the complete set of selected key variable to reach uniqueness in the dataset, but instead a subset of the key variables suffices to create uniqueness of the record. The **SUDA** (special uniques detection algorithm) score is based on this definition of special uniques. Special uniques are more likely to be re-identified by a possible intruder.

Based on the disclosure scenarios and the selected key variables, we can evaluate the disclosure risk for all scenarios. The risk measures are based on the prepared data.

4. Overview of anonymization methods

In this section we give an overview of commonly used anonymization methods. In case the disclosure risk is too high in the data set, there are several methods to reduce this risk before the release. The most common approach is to reduce the detail in the data by **recoding**, which is a method to reduce the number of categories in the data by combining

several categories. An example is to combine several regions or industry types in more aggregate categories. Often this can be realized by using a higher level geographical or more aggregate industry type without losing valuable information for data users. A variation to recoding is **top coding**, where high values of a certain variable, which are often outliers, are replaced by one common value. An example is to replace the age values over 60 with 60. Continuous variables can be transformed into bands, e.g. income bands.

In case recoding does not reduce the risk sufficiently, individual values can be removed by using **local suppression.** Local suppression algorithms seek to suppress values that cause uniqueness of a record. Suppressing or removing these values guarantees that these records cannot anymore be identified based in these values. Here also (rare) combinations of values are considered.

There are several other methods to introduce uncertainty in the microdata, which prevent an intruder knowing whether an identity disclosure is correct or not. These methods perturb the data and are called **perturbative** methods. One popular method for categorical variables is **PRAM**, which changes the values in a categorical variable at random. **Noise addition**, which is based on adding small distortions to continuous variables, is often used for creating uncertainty around values of continuous variables. Perturbative methods are generally only reverted to if non-perturbative methods do not provide sufficient protection. The reason is that perturbative methods might lead to a high level of information loss.

5. Anonymization measures applied

This dataset has been treated to protect confidentiality. Several methods have been applied to protect the confidentiality: removing variables from the original dataset, reducing detail in variables by recoding and top-coding and removing particular values of individual records at risk (local suppression) as well as randomization of the record order and record ids.

Removing variables

The released microdata set has only a selected number of variables contained in the initial survey. Not all variables could be released in this PUF without breaching confidentiality rules. Table 2 gives an overview of the removed variables.

Table 1 Overview of removed variables by module

Variable name	**Description**
Module 1	
m1siq1	Name of enumerator during first visit
m1siq2	Name of enumerator during second visit
m1saq4a	Does this school offer classes specifically for children with special needs/disabilities?
m1siq5a	School Name
m1siq5b	School Survey Code

m1siq6	School EMIS Code/Registration Number
m1siq7S	GPS Position S
m1siq7E	GPS Position E
m1saq0	Name of respondent
m1saq2	Respondents contact (Phone No.)
m1sdq9ac	For each term when did the school open and close?-End of term 1
m1sdq9bc	For each term when did the school open and close?-End of term 2
m1sdq9cc	For each term when did the school open and close?-End of term 3
m1sdq7_1e	Ending time of school day – First shift
m1sdq7_2e	Ending time of school day – Second shift
Module 2	
m2saq2	First and last names
m2sbq1	First and last names
Module 4	
m4siq1a	Enumerator name
m4siq3a	School Name
m4siq3b	School Code
m4siq4	School EMIS Code/Registration Number
m4siq6	Scheduled class time
m4siq8	Teacher name
Module 5	
m5siq1	Enumerator name
m5siq3	School Name
m5siq4	School Survey Code
m5siq8a	Teacher name
m5sa1q2	Pupil's first name
m5sa1q8a	Name of your English/Kiswahili teacher this year
m5sa1q9a	Name of your Math teacher this year
m5sa1q11a	Name of your English/Kiswahili teacher last year
m5sa1q12a	Name of your Math teacher last year
Module 6	
All variables with variable names ending in ab, bb, cb, db, eb, fb, gb, hb, ib, jb, kb	
All variables containing characters (answers to test questions)	

Reducing detail in variables by recoding and top-coding

In several variables, the detail is reduced by recoding values or top-coding. The following table gives an overview of the variables that have been changed.

Table 2 Recodes of key variables

Variable name	Description	Approach used	Before	After

school level				
m1saq3	School ownership type	Public/private, other to missing	1; 2, 3	1; 2
m1saq4	School type	Boarding and boarding/day school to boarding/day school	1; 2, 3	1; 3
m1saq5	School category	Group urban and semi-urban	1; 2, 3	1; 2
m1saq6	Begin operating year	Group by decennia, bottom-code at 1960, top-code at 2000		1960, 70, 80, 90, 2000
m2sbq19	Number of classrooms	0, 1-9, 10-19, 20+	0; 1-9; 10-14; 15-19, 20+	0; 5; 12; 17; 20
teacher level				
m2saq4	Position in school	Group to head/teacher	1, 4; 5-9	1;5

Local suppression by variable

Values of certain variables for particular schools and teachers were deleted. Table 5 gives an overview of the number of suppressed values per variable.

Table 3 Overview of number of suppressions per key variable

Variable name	**Description**	**Number of supressions**	**Total number of observations**
School level			
m1siq2a	Region	7	306
m1siq4	Rural/urban	1	306
m1saq3	School ownership type	3	306
m1saq4	School type	4	306
m1saq5	School category	10	306
m1saq6	Begin operating year	30	306
m2sbq19	Number of classrooms	10	306
Teacher level			
schid	School id	13	4,413

Other measures

To guarantee confidentiality, the following measures are implemented in the R code:

- recode and randomize the district ids within the regions.
- randomize order and ID of schools within the regions. The initial order of the file would allow re-identification based on the order: two schools that are close to one another in the file are also geographically close, since the file was ordered by lower level geographical variables. The mapping of the old to the new IDs is available.
- Randomize teacher IDs within schools. The mapping of the old to the new IDs is available.
- Randomize enumerator codes. The mapping of the old to the new IDs is available.

- Dates are recoded relative to the day of the survey (variables m1siq8, m1siq9, m1siq10d, m1siq10e, m4siq7, m4siq10d, m4siq10e, m5siq5, m5siq9d, m6siq4)
- m1sbq8 (when was the last visit of the official government quality assurance officer or inspector?) is recoded to months past until survey date
- Start and end times are recoded to duration and AM/PM (variables m1siq11, m1siq12, m1siq13, m1siq14, m5sa1q14, m5sa1q15)
- m4sbq1 (number of pupils in the room) is rounded to the nearest 10 and top-coded at 60.
- Recode the values in m1sdq2a-c to proportions of the totals an suppress the totals
- m4sdq2 (number of pupils registered in class) is rounded to the nearest 10
- m5sa1q3 (age of pupils) is bottom-coded at 9 and top-coded at 13
- Number of toilets (m1scq3 and m1scq4) is top-coded at 12.
- Recode absolute number to proportions. This is important when the aggregate is recoded. The proportions are computed from the original values (m2sbq20, m2sbq21 from m2sbq19 (number of classrooms), m4sbq2, m4sbq3, m4sbq4, m4sbq4a, m4sbq4b, m4sbq5, m4sbq5a, m4sbq5b, m4sbq6, m4sbq6a, m4sbq6b, m4scq2, m4scq2a, m4scq2b, m4scq5, m4scq5a, m4scq5b, m4scq6, m4scq6a, m4scq6b, m4scq10, m4scq10a, m4scq10b, m4scq12 and m4saq5a – m4saq60a from m4sbq1 (number of pupils in the room), m4sdq3 from m4sdq2 (number of pupils registered in class)).
- Set all values in m4saqc3 and m4saqc4 to NA (count of proportion doesn't make sense)
- Recode m5sa1q5 (mother tongue) to Kikuyu, Kisii, Kimeru, Luhya, Luo and Other
- Recode m1sdq7_1s and m1sdq7_2s to length of school day in hours
- Age of teachers (m2sbq14, m4sdq8, m6siq8) is recoded as follows: <25, 25-34, 35-44, 45-54, 55+
- Year of beginning teaching (m2sbq12, m4sdq11) is recoded <1990, 1990-1999, 2000-2010, 2010+.
- Several variables are recoded according to the recodes of the key variables

Randomization

To further anonymize the data, the order of the records was randomized. Also teacher and school IDs were randomized. The randomized IDs still allow to match the files. Randomization allows using these variables for the evaluation of fixed effects.

6. Note on information loss

SDC methods lead to a loss of information or utility in the data. To check the validity of the data for research purposes, several indicators were computed from the original data and from treated data. The indicators values computed from the released dataset do not significantly differ from the values in the original dataset and hence the dataset is valid for research purposes.

7. Bibiliography

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., et al. (2012). *Statistical Disclosure Control.* Chichester, UK: John Wiley & Sons Ltd.

Templ, M., Meindl, B., Kowarik, A., & Chen, S. (2014, August 1). *Introduction to Statistical Disclosure Control (SDC).* Retrieved July 14, 2015 from http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf