

Center for Enterprise Dissemination
Disclosure Avoidance

Working Paper Series
(Disclosure Avoidance CED-DA-FY20-001)

Re-Identification Primer using Four Metrics - Updated

Phyllis Singer and Aref Dajani, CED-DA

March 19, 2020

Center for Disclosure Avoidance Research
U.S. Census Bureau
Washington DC 20233

Report Originally Issued: September 2019. No titled nor sensitive data in this report.
Consequently, this updated report did not require vetting through the Disclosure
Review Board.

Disclaimer: This report is released to inform interested parties of ongoing research and
to encourage discussion of work in progress. The views expressed are those of the
authors and not necessarily those of the U.S. Census Bureau.

DRB Approval Number CBDRB-FY19-545

ABSTRACT

Re-identification studies look for vulnerabilities in an entire database to attacks from an external source or collection of sources. This working paper updates our methodology for re-identification studies. We introduce the concept of ranking by value to two of the three methods, as used by our external stakeholder from HUD for his *adhoc* custom scoring method. We also introduce thresholds to enhance the ranking procedure in the *adhoc* method.

1. INTRODUCTION

Re-identification studies conducted by CED-DA assume that we know as much as external intruders. Re-identification studies look for vulnerabilities in an entire database to attacks from an external source or collection of sources. They do not determine whether someone with intimate knowledge of a specific respondent can find that respondent in the database. The only way to protect a single individual in a database perceived to be at high risk of re-identification is through data perturbation, such as noise injection, or information reduction, such as removing the observation altogether.

1.1 Files Required for a Re-Identification Study

There are three files analyzed when conducting re-identification studies in CED-DA:

1. The Public Use File (PUF) contains, at a minimum, a set of linking variables and a key variable: **pufid**.
 - The **pufid** can be a combination of variables.
2. The External Intruder File (EIF) contains, at a minimum, a set of linking variables and a key variable: **eifid**.
 - The **eifid** can be a combination of variables.
3. The Internal Use File (IUF) contains, at a minimum, the **pufid** and the **eifid**.
 - The IUF may also contain the non-perturbed values of the linking variables.

1.2 General Course of a Re-Identification Study

When conducting re-identification studies:

1. One links a PUF with an EIF using the linking variables.
2. The number of unduplicated suspected re-identification pairs divided by the number of observations in the PUF, expressed as a percentage, is the suspected re-identification rate. This is also referred to as the putative re-identification rate.
3. These suspected re-identification pairs are checked against the IUF to determine which are confirmed re-identifications and which are false positives. The number of confirmed matches divided by the number of observations in the PUF, expressed as a percentage, is the confirmed re-identification rate.
4. The conditional re-identification rate is expressed as a percentage, calculated either by:
 - a. The number of confirmed re-identifications divided by the number of suspected re-identifications or
 - b. The confirmed re-identification rate divided by the suspected re-identification rate.

5. The conditional re-identification rate is the critical statistic to determine whether to release a PUF.
 - a. External intruders may calculate low or high suspected re-identification rates, given the information they have available to them.
 - b. They may even purport that they successfully linked their external data to our Public Use File.
 - c. We know precisely how successful their re-identification attempt was, but only if we have access to the same external information.
6. The conditional re-identification rate, identical to the metric of precision in the record linkage and health science literature [Arbuckle; Herzog; Scaiano], represents the ratio of true positives to the sum of true positives and false positives. Data owners will not be alarmed if an external organization reports a relatively high suspected re-identification rate so long as the conditional re-identification rate is low.

1.3 The Four Re-Identification Metrics

There are four metrics calculated when conducting re-identification studies in CED-DA:

1. *Unicity*
2. *Taxicab*
3. *Euclidean*
4. *Adhoc* (currently used for re-identification studies sponsored by HUD)

1.4 Types of Linking Variables

There are two types of linking variables:

1. numeric (can be continuous)
2. categorical (can be ordinal)

2. CED-DA'S DEFINITION AND IMPLEMENTATION OF THE FOUR RE-IDENTIFICATION METRICS

This section provides the definitions of the scoring methods for each of the four re-identification metrics. It also provides instructions on how to implement the general course of a re-identification study for each of the four metrics.

Note that:

- For the *unicity* and *taxicab* methods, it is necessary to bin numeric variables into quintiles to make them all categorical.
- For the *euclidean* and *adhoc* methods, we use both numeric and categorical variables as they appear.

2.1 Unicity

1. Convert all numeric linking variables to categorical by quintile, with values 1-5.
2. Construct all potential interactions across linking variables.
 - We only consider observations with non-missing values for the selected interactions of linking variables.
3. Determine the set of sample uniques in the PUF.
 - Sample uniques are observations that are in a cell with a frequency of one.
4. Determine the set of sample uniques in the EIF.
5. Determine which sample uniques are in the same cell (identical values for all linking variables in that interaction) in both the PUF and the EIF. Those sample uniques are the suspected re-identification pairs.
6. Note: The number of sample uniques in both datasets, unduplicated by **pufid**, is the number of PUF observations suspected of re-identification.
 - They need to be unduplicated because a sample unique in a lower-order interaction of linking variables will also appear in any higher-order interaction of linking variables containing that lower-order interaction.
 - All orders of interaction are considered because missing values for any linking variable are excluded from consideration in that particular interaction of linking variables. Were there no missing values, only the highest order interaction would be considered.
 - The number of potential interactions for v variables is $2^v - 1$. Until the method is enhanced to run faster, we limit the number of combinations to 8,191: where $v=13$.

2.2 Taxicab, euclidean, and adhoc Commonalities

1. Construct the Cartesian product of the PUF and the EIF.
2. Score the pairs according the method (documented following for each metric, separately).
3. Retain only those pairs that meet the threshold (also documented following).
4. Rank the scores that meet the threshold and look at the distribution of ranks.
5. Retain only pairs where the cumulative frequency of distinct ranks is less than or equal to five.
 - For example, if the top three ranks have unique scores while there are twenty observations with the same fourth rank, then only the first three pairs are retained.
 - If the top rank is shared by twenty observations, then none of the pairs are retained for that **pufid**.
 - According to [Simon], "A cell size or count of 5 or 6 is often held out as a threshold for unacceptable risk of re-identification."

6. The number of pairs in both datasets, unduplicated by **pufid**, is the number of PUF observations suspected of re-identification.
 - They need to be unduplicated because the Cartesian product has every observation in the PUF paired to every observation in the EIF.

2.2.1 *Taxicab Details*

1. Convert all numeric linking variables to categorical by quintile with values 1-5.
2. For all linking variables (now categorical), define the **taxicab** score as follows:
 - 0 if the linking variable is in the same category
 - 1 if the linking variable is in a different category
 - α if the value of the linking variable in either the PUF or the EIF is missing
 - We currently set $\alpha = 0.5$.
3. Calculate the **taxicab** metric as the L1 norm (sum of absolute values) of the scores across all linking variables, divided by the number of linking variables: the mean absolute difference.
 - The minimum and maximum values of the **taxicab** metric are 0 and 1, respectively.
 - Set the threshold to be the value of **taxicab** where half the observations were matches and half the observations were missing.
 - With v linking variables, $\text{threshold} = [(0 * v/2) + (\alpha * v/2)] / v = \alpha / 2$
 - The L1 norm [Barille; Krause] is sometimes called the taxicab or Manhattan distance.
4. A pair is a suspected re-identification if the **taxicab** score is less than the threshold, $\alpha / 2$.

2.2.2 *Euclidean Details*

1. Retain numeric linking variables as numeric.
2. For categorical linking variables or when the value of the linking variable (whether categorical or numeric) for either the PUF or the EIF is missing, the **euclidean** score is identical to **taxicab**.
3. For all numeric linking variables with values that are not missing, define the **euclidean** score as follows:
 - Calculate the z-score for the value of the linking variable in the PUF and the EIF, where the means and standard deviations for each linking variable are those on the PUF to keep the z-scores comparable.
 - Calculate the absolute difference of the z-scores, capped at a maximum of six.
 - Take the logit of the absolute difference of the z-scores, where the logit is defined as $\text{logit}(x) = \frac{e^x}{1+e^x}$ and $x = \min(|\Delta z|, 6)$.

- For $x > 0$, To calibrate the logit to be between 0 and 1, with 0 for minimum distance and 1 for maximum distance, use the following formula:
 - *euclidean* linking variable score = $(2 * \text{logit}(x) - 1)$.
4. The *euclidean* metric is calculated as the L2 norm (root of the sum of squares) of the differences in scores across the linking variables, divided by the number of linking variables. This is equivalent to the standard deviation divided by the square root of the number of linking variables.
 - The minimum and maximum of the *euclidean* metric are 0 and 1, respectively.
 - Set the threshold to be the value of *euclidean* where half the observations were matches and half the observations were missing.
 - With v linking variables, $\text{threshold} = \frac{\sqrt{(0*v/2)^2 + (\alpha*v/2)^2}}{v} = \alpha/2$
 - The L2 norm [Torra] is sometimes called the Euclidean distance. It is related to the Mahalanobis distance.
 5. A pair is a suspected re-identification if the *euclidean* score is less than the threshold.

2.2.3 Adhoc Details

1. Retain numeric linking variables as numeric.
2. Score by linking variable as defined by the information product originator. The scores range from zero to five. A high score indicates a match while a low score indicates a non-match.
3. If the value of the linking variable in either the PUF or the EIF is missing, consider the pair to be a non-match and give it an *adhoc* score with a value of zero.
4. The *adhoc* metric is the sum of the scores across the linking variables.
 - The minimum and maximum of the *adhoc* metric are 0 and $5*v$, respectively.
 - Set the threshold to be the value of *adhoc* where half the observations were matches and half the observations were missing.
 - With v linking variables: $\text{threshold} = (5 * v/2) + (0 * v/2) = 5 * v/2$
5. A pair is a suspected re-identification if the *adhoc* score is greater than the threshold.

2.3 Important Note

The set of suspected re-identifications will not be identical for *taxicab* and *euclidean*, even if all linking variables are categorical, because of the use of L1 and L2 norms.

3. CHANGES TO OUR RE-IDENTIFICATION METHODOLOGY: WHY AND HOW

The four re-identification methods described in this working paper are impacted by both the number of linking variables and the sizes of the PUF and the EIF. As the number of variables, the size of the PUF, and the size of the EIF increase, together or separately, the re-identification programs sometimes do not run to completion. When they run to completion, the time to completion can be extremely long, especially on shared servers. Sometimes, programs will use too many computing resources. It became obvious that we would need a way to control the number of pairs to compare.

To control the number of pairs to compare, we ranked the paired observations by their *taxicab*, *Euclidean*, or *adhoc* scores and restricted our attention to the top (or bottom) ranked ranks, with a cumulative number of suspected pairs of at most five. Using thresholds to determine which pairs were suspected increased computer efficiency. Using ranks in this manner mitigated the ambiguity for suspected pairs that were later classified as confirmed re-identifications. For example, if the top rank contained twenty observations with one of them correct, then there is ambiguity about which of these twenty putative matches is actually correct.

Another way to control the number of pairs to compare is to parse the databases through defining strata with finer granularity than that which Census will release in the final information product. If the strata defined with finer granularity result in acceptable disclosure risk, then Census will release the final information product with acceptable disclosure risk.

For example, if there are n observations in both the EIF and the PUF, then there would be n^2 pairs of observations. If we were able to divide each dataset into two equally allocated strata, then there would be two strata in each dataset, each with $n/2$ observations. Each stratum would have $n^2/4$ pairs, resulting in $n^2/2$ pairs across the two strata, cutting the number of pairs for comparison precisely in half.

If the database passes the re-identification study with finer strata than what Census intends to release, that may indicate that the PUF may be released in the future at finer levels of strata, such as lower levels of geography.

The re-identification studies developed and conducted by CED-DA are used to determine whether a database has obvious vulnerabilities to re-identification attacks. Because the study links two datasets at a single point in time, they do not confirm that any dataset is safe for release, just that there are not endemic problems in its release.

These studies may identify problems that can direct improvements to our disclosure avoidance methods. Their purpose has never been to replace the legacy or modern provable privacy methods at the Census Bureau, but to act as a quality control to verify that the methods, old and new, protect as they are designed.

4. REFERENCES

- Arbuckle L, Emam KE (2013) *Anonymizing Health Data: Case Studies and Methods to Get you Started*. Sebastopol (CA): O'Reilly Media Inc.
- Barile M. Taxicab Metric. From MathWorld--A Wolfram Web Resource, created by Eric W. Weisstein. <http://mathworld.wolfram.com/TaxicabMetric.html>
- Herzog TN, Scheuren FJ, Winkler WE (2007) *Data Quality and Record Linkage Techniques*. New York/London: Springer.
- Krause EF (1986) *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*. New York: Dover Publications, Inc.
- Scaiano M, Middleton G, Arbuckle L, Kolhatkara V, Peyton L, Dowling M, Gipson DS, Emam KE (2016) A unified framework for evaluating the risk of re-identification of text de-identification tools, *Journal of Biomedical Informatics* Volume 63, October 2016: 174-183.
- Simon GE, Shortreed SM, Coley RY, Penfold RB, Rossom RC, Waitzfelder BE, Sanchez K, Lynch FL (2006) Assessing and Minimizing Re-identification Risk in Research Data Derived from Health Care Records. *EGEMS* (Wash DC). 2019 Mar 29;7(1):6.
- Torra V, Abowd JM, and Domingo-Ferrer J (2006) Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment. In: Domingo-Ferrer J., Franconi L. (eds) *Privacy in Statistical Databases*. PSD 2006. Lecture Notes in Computer Science, vol 4302. Springer, Berlin, Heidelberg.