

A Method to Improve Data Swapping at the U.S. Census Bureau

Marlow Lemons, Ph.D.

Center for Disclosure Avoidance Research

April 1, 2015

Disclosure Avoidance: A Balancing Act

- Goal is to publish as much valuable statistical information as possible while protecting the confidentiality of our respondents
- Can accomplish both by applying disclosure avoidance (DA) techniques prior to releasing our data products

Data Swapping [Introduction]

- *Data swapping* is a disclosure avoidance method that involves interchanging a subset of records to protect respondent privacy.
- First introduced in 1978 by Dalenius and Reiss.
- The goal is to introduce uncertainty in sensitive records while preserving statistical properties in the data.

Data Swapping [Routines]

1. Selection

- Risk level is calculated for each record using the risk and geography variables.
- High risk records are flagged for swapping.



2. Matching

- Flagged records are matched with other flagged records based on matching variables.



3. Deselection

- Removes excessive swap matches based on the target swap rate.



4. Swapping

- Perform the swap by exchanging the variable values among matched records remaining after deselection.

Data Swapping [Issues]

Questions Concerning Data Swapping

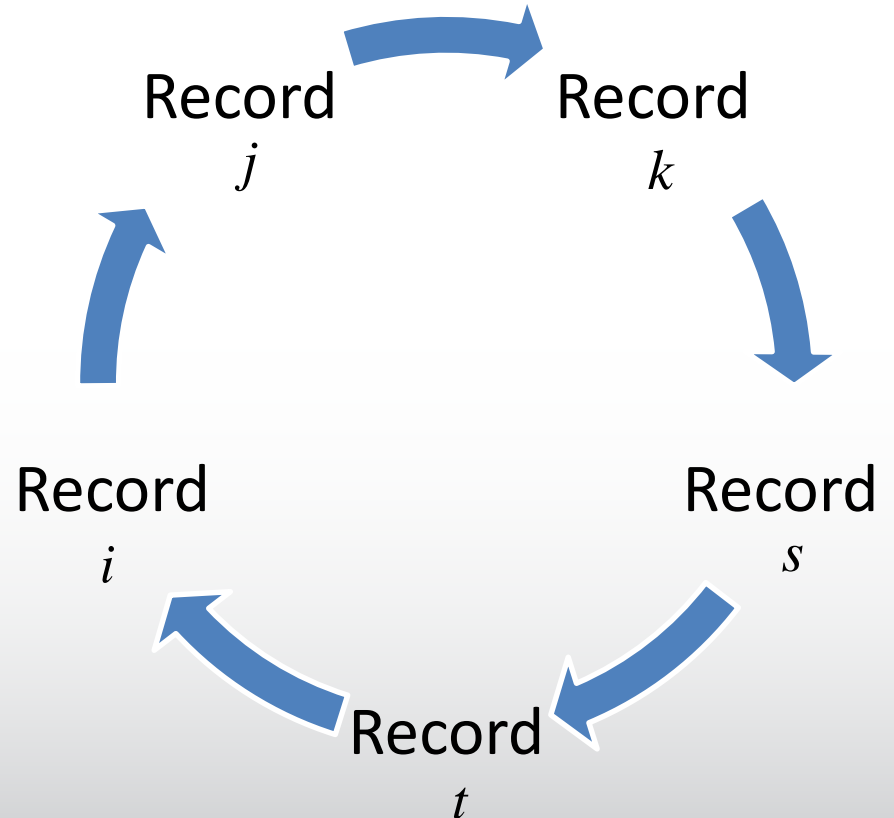
- What swapping rate should be used?
- How does one determine which records are considered “risky”?
- Are there restrictions in what variables are swapped?
- **How do you swap?**
- What are ways to measure the quality of the perturbed data?

Variations of Data Swapping

































































































































- There are several variations of data swapping:
 - Pair data swapping
 - Data shuffling
 - ***n*-cycle swapping**
- The final set of slides describe the ***n*-cycle swapping**, which has been the Center's latest research focus.

n-cycle Swapping

- The *n*-cycle swapping method interchanges records using permutations, that is an arrangement of a set of records using a one-to-one function.



n-cycle Swapping [Selection]

Variables		Identification	Geography	Risk	Match	Other	Risk Level	Risk Flag								
1																
2																
3																
4																
5																
6																
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$N-1$																
N																

n-cycle Swapping [Matching]

- The matching routine attempts to locate swapping partners for each flagged record while preserving the values of the match key variables.

Tenure

Number of
Minors in
Household

Number of
Adults in
Household

- Matching takes place within county at all costs. The records must be in different Census tracts.

n-cycle Swapping [Matching]

Type	Description
1	County level n-Cycle where all n records are at risk for disclosure
2	PUMA level n-Cycle where all n records are at risk for disclosure
3	SPUMA level n-Cycle where all n records are at risk for disclosure
4	State level n-Cycle where all n records are at risk for disclosure
5	County level pair matching where one record is at risk and the other is not.
6	PUMA level pair matching where one record is at risk and the other is not.
7	SPUMA level pair matching where one record is at risk and the other is not.
8	State level pair matching where one record is at risk and the other is not.

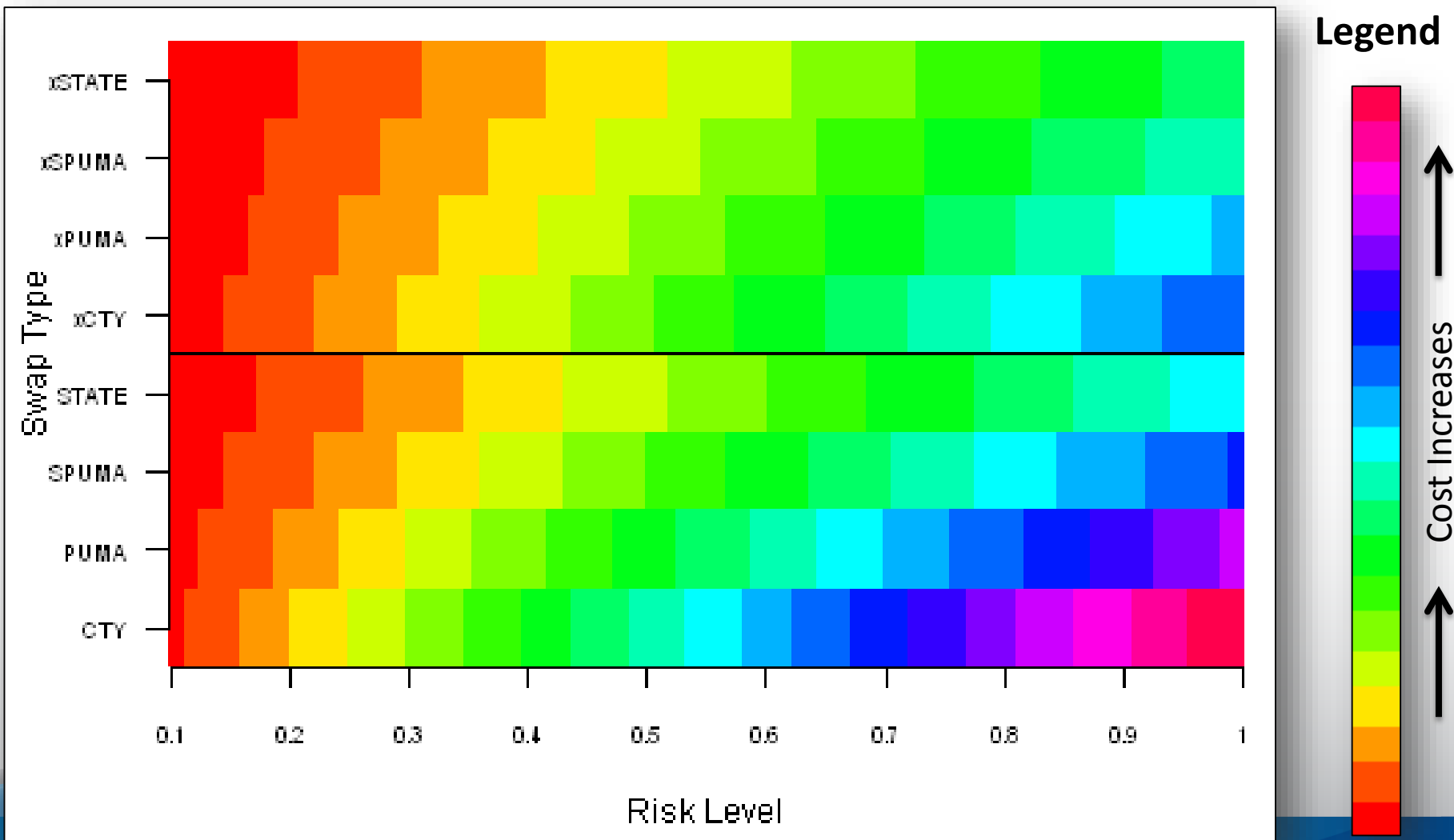
n-cycle Swapping [Deselection]

- Deselection involves removing matches to meet the target swap rate.

$$Cost_i = \begin{cases} e^{-\frac{type}{a_1}} \left(risk_i + risk_{i+1} - .1 - \frac{b_1}{9-type} \right) & , \quad type \in \{1,2,3,4\} \\ e^{-\frac{type}{a_2}} \left(risk_i - \frac{b_2}{9-type} \right) & , \quad type \in \{5,6,7,8\} \end{cases}$$

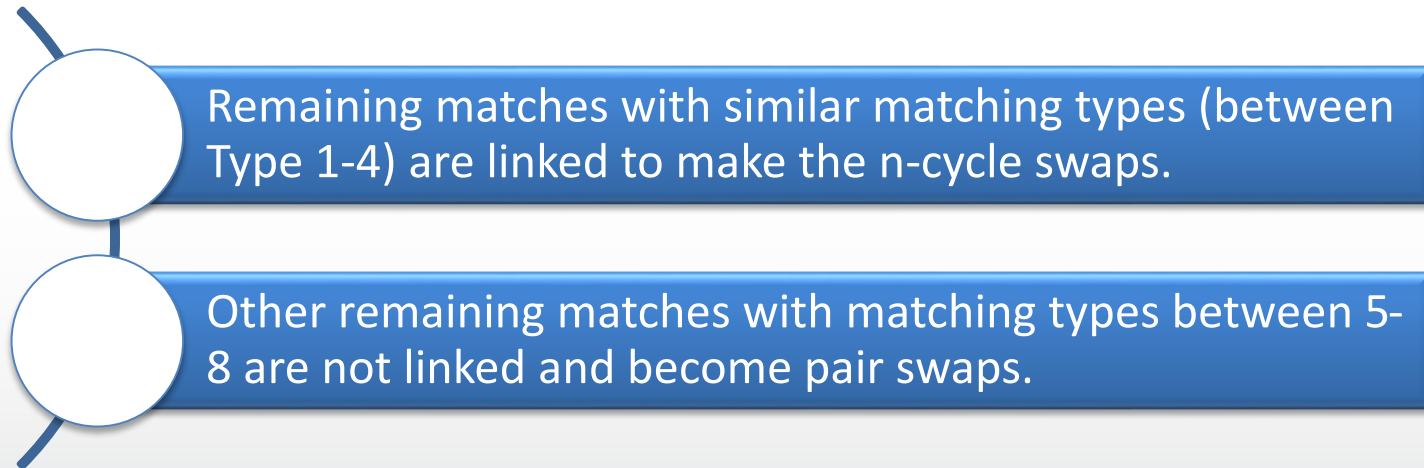
- A cost is calculated for each match. The cost function, along with the risk score, is used to deselect matches.

n-cycle Swapping [Deselection]



n-cycle Swapping [Swapping]

- All remaining matches, after deselection, are swapped.



- Swapping involves exchanging the geography information.

Next Steps

Comparing the data utility of the pair swapping and n -cycle swapping procedures.

Specifications for incorporating n -swapping into Census 2020.

Comparing quality of the perturbed data between the pair swapping and n -cycle procedures.

Reidentification risk measures on perturbed data due to n -cycle swapping.

Conversion of the n -swapping algorithm into SAS.

Thank You!