Generation of Synthetic Census Data based on Correlation

Amber Rawson, Dr. Anand Sarwate

Department of Electrical and Computer Engineering, Rutgers University, New Brunswick

Abstract

American Community Survey (ACS)

- Information about households from income to internet accessibility
- Used to analyze economic and social trends
- Data is “anonymized” to protect privacy
- New attacks on privacy may be able to “deanonymize” the data

Can we generate *synthetic data* that gives more privacy while still preserving the statistical trends?

Technical Challenges:

- How should missing entries in ACS data be handled?
- How can the statistical trends be preserved?

Method:

- Use Python packages (Numpy, Pandas)
- Use imputation methods to fill in missing entries
- Estimate correlation structures in the data that should be preserved

Background

Census Bureau

Conducts surveys and compiles sources including:

- US Department of Health and Human Services
- US Department of Housing and Urban Development
- US Department of Justice

The Issue

Sets of demographic and economic statistics are available for research but require special access for non public data

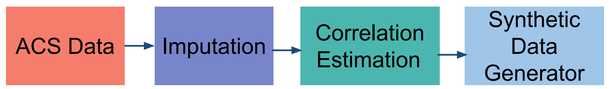
- Confidentiality laws protect citizens privacy
- Gaining special access can hinder research

One Current Privacy Protection Method

Introduce error into the data to anonymize information.
Must consider *Privacy Utility Tradeoff*

- Enough error to protect respondents’ data
- Still accurate enough to be used in research

Synthetic Data



Methodology and Analysis

Importing and Organzing the Data:

Data: 1-year 2015 data for NJ Housing
Dataset was focused on internet access and annual family income (23, 015 entries):

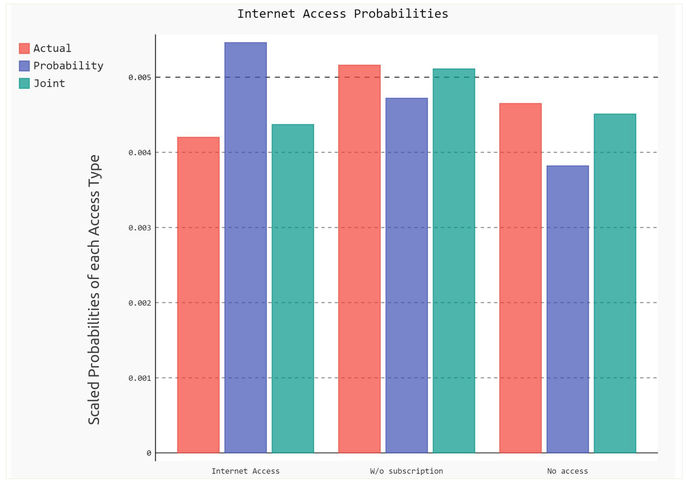


Figure 1. Probabilities of the actual data, imputation by probability and imputation based on correlation between Internet Access and Annual Income. Probabilities are scaled by 0.895, 0.021, 0.070

Analyzing Dependency of Internet Access on Family Income

By recognizing that statistics are correlated, better data can be imputed based on combined probabilities
Income was broken down into 5 brackets 0, $44000, $78000, $116000, $176500, $1787000

Mutual Information - Analyzing Correlation

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

I: (Mutual Information) Measures how much knowing one variable reduces the uncertainty of the other
X and Y: Two discrete random variables
p(x) and p(y): Marginal probability distribution functions
p(x,y): Joint probability distribution function

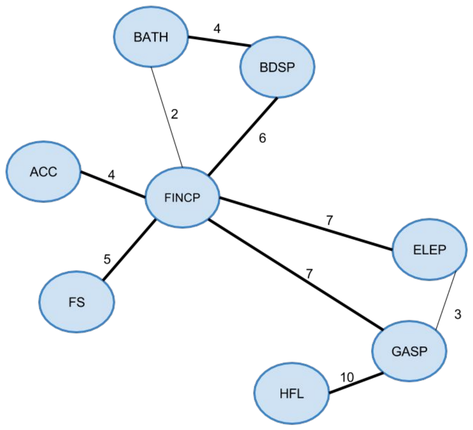


Figure 2. Maximum Weight Spanning Tree:
Represents a possible configuration for the connections between:

Family income(FINCP)	Internet access(ACC)
Shower access(BATH)	Bedrooms(BDSP)
Electricity monthly cost(ELEP)	Food stamp(FS)
Gas monthly cost(GASP)	House heating(HFL)

The edges (lines connecting the nodes) have a value representing how strongly the two variables are related.

A tree, indicated by the bolded lines, was chosen based on the highest numbers, and therefore the most correlated items. All nodes must be included in the path.

Future Direction

By finding the probability densities and weighing correlations between all the variables, a tree can be constructed that connects the data. This can then be used to generate accurate synthetic data given one data point.

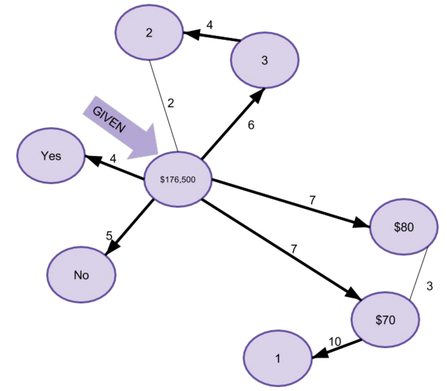


Figure 3. Constructed spanning tree

Big Picture:

This generated data can be released to the public for research purposes without fear that it can be reconstructed into its original respondent, because it does not represent one. Easier access to data will facilitate research.

Acknowledgements

I would like to thank Dr. Anand Sarwate for his guidance and support throughout this project. I would also like to extend my gratitude to the Douglass Project and Nicole Wodzinski for providing this research opportunity.

References

Dwork C. (2008) Differential Privacy: A Survey of Results. In: Agrawal M., Du D., Duan Z., Li A. (eds) Theory and Applications of Models of Computation. TAMC 2008. Lecture Notes in Computer Science, vol 4978. Springer, Berlin, Heidelberg

Mayer, Thomas S. "Privacy and confidentiality research and the us census bureau recommendations based on a review of the literature." Survey methodology (2002): 01.

McCaa R., Ruggles S., Davern M., Swenson T., Palipudi K.M. (2006) IPUMS-International High Precision Population Census Microdata Samples: Balancing the Privacy-Quality Tradeoff by Means of Restricted Access Extracts. In: Domingo-Ferrer J., Franconi L. (eds) Privacy in Statistical Databases. PSD 2006. Lecture Notes in Computer Science, vol 4302. Springer, Berlin, Heidelberg

RUTGERS