Sri Lankan Journal of Applied Statistics, Vol.10, 2009, p.229-236

Balancing Disclosure Risk with Data Quality

A. RAMANAYAKE AND L. ZAYATZ[1]

U.S. Census Bureau, USA

ABSTRACT

As the leading source of data about the nation's people and the economy, the Census Bureau's mission is to provide high quality data while honoring the privacy and protecting the confidentiality of its survey respondents. Legal mandates aside, the Census Bureau needs to maintain the trust of the nation without which it will not be able to gather quality data and achieve high response rates. Thus the Census Bureau applies disclosure avoidance techniques to its data products prior to public release in order to protect the confidentiality of its survey respondents. This paper discusses three disclosure avoidance techniques that are currently used: cell suppression, data swapping and synthetic data.

Keywords: Confidentiality, Data Dissemination, Disclosure Avoidance

The U.S. Census Bureau is the leading federal agency that collects data about the nation's people and the economy. It conducts the decennial census, as well as the economic census and numerous sample surveys. Once data are gathered and processed, statistical summaries and public use microdata files (PUMS) are released to the public. These products are used for many purposes, including redrawing state legislative and congressional districts, determining distribution of funds for government programmes, and planning public facilities. A statistical disclosure happens when a data product inappropriately reveals information about a respondent which was not apparent prior to the release. The resulting disclosure risk depends on the likelihood and impact of the disclosure.

The Census Bureau is required by law to protect the respondents to its censuses and sample surveys from such disclosures. Thus the Census Bureau takes necessary measures before releasing its data products. First, the data are de-identified. That is, any direct identifiers such as name, address, etc., are removed from the released data products. However, even after this de-identification, there still could be other sensitive information about the respondents that could get revealed through the

---

[1] 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.

released data. Therefore, the Census Bureau takes additional steps to eliminate these more subtle disclosure risks before releasing its data products. At the same time, the Census Bureau needs to maintain the quality of the data products it provides. Methods with high disclosure protection lead to lower data quality in general. Thus the goal is to achieve an optimum data quality at a tolerable disclosure risk. Figure 1 shows this balance between data utility and disclosure risk. The dotted line shows the extent of disclosure risk that is tolerated.

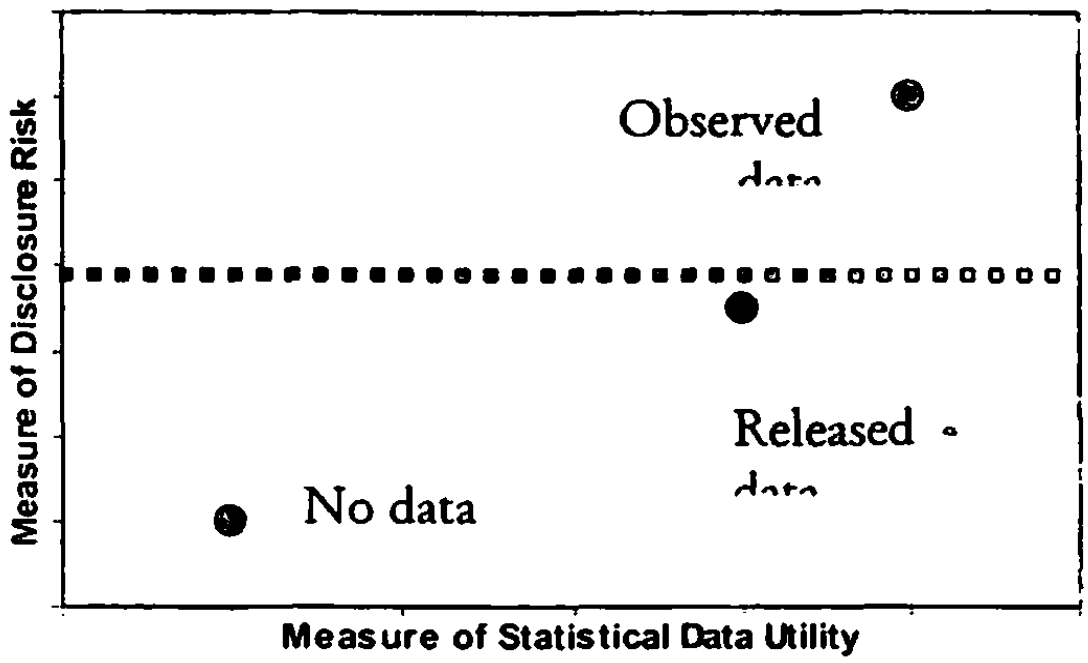

Figure 1: Risk-utility plot

The different disclosure masking techniques used are based on the type of data and the data product that is released to the public. In this paper, three disclosure masking techniques that are currently used by the Census Bureau are discussed.

CELL SUPPRESSION

The most common method employed by the Census Bureau to release data to the public is statistical tables. The Census Bureau publishes frequency count data mainly from the Decennial Census and the American Community Survey (ACS). Tables of frequency count data present the number of units in each cell. For example, a table may have rows representing the marital status of respondents and columns representing their citizenship status. The cell values reflect the number of people in a given geographic area having the various combinations of marital status and citizenship status. Statistical disclosure from frequency count tables occurs when a data intruder can link a respondent to a small cell count. This would be a violation of a threshold rule. For example, Table 1 below contains a sensitive cell

which reveals that exactly two of the widowed are non citizens. One of the methods used to prevent disclosure in such tables is cell suppression. When a sensitive cell is suppressed, other cells must be suppressed or else the value of the cell could be calculated from the marginal totals. These additional cells are called complementary suppressions.

Table 1: Example of a Frequency Table

	Born in the US	**Naturalized**	**Not a citizen**	**Total**
Married	77	12	13	102
Widowed	46	0	2	48
Separated/Divorced	12	18	12	42
Never married	619	23	22	664
Total	754	53	49	856

The Census Bureau also publishes magnitude data from its economic censuses and sample surveys. Tables of magnitude data often contain the frequency counts of establishments in each cell, but they also contain the aggregate of some quantity of interest over all units of analysis (establishments) in each cell. For example, Table 2 below presents the total value of shipments within the manufacturing sector by North American Industry Classification System (NAICS) code by county within a state. The frequency counts in these tables are not considered sensitive because so much information about establishments, particularly classifications that would be used in frequency count tables, is publicly available. The magnitude values, however, are considered sensitive and must be protected. The most common technique that is used for these tables is also cell suppression.

Table 2: Example of a magnitude data table

NAICS	Description	Establishments	Value of shipments
21	Mining	186	2,180,922
211	Oil and gas extraction	17	10,073
2111	Oil and gas extraction	17	10,073
21111	Oil and gas extraction	17	10,073
211111	Crude petroleum and natural gas extraction	16	D
211112	Natural gas liquid extraction	1	D

In this case, any table cell value that could allow users to estimate a responding company's value "too closely" is not shown. The value is suppressed and replaced with a "D" for disclosure. These values are called primary suppressions or sensitive cells. They are identified using the p% rule (Cox, 1980). Consider a cell with value T. Suppose

$$T = \sum_{i=1}^{N} x_i ,$$

where N is the total number of establishments contributing to T and $x_i \geq 0$, are the individual contributions of the establishments, ordered from the largest to the smallest. The p% rule makes sure that the second largest contributor is not able to estimate the largest contributor's value to within p% of its value. Thus the cell will result in disclosure if

$$T - x_2 < \left[1 + \frac{p}{100}\right] x_1.$$

More conveniently this could be written as,

$$\frac{p}{100} * x_1 - \sum_{i=3}^{N} x_i > 0.$$

Any cell that satisfies the above condition thus will be suppressed under the p% rule. Because marginal totals are shown in tables, complementary suppression must be performed, so that primary suppression values cannot be derived or estimated too closely via addition and subtraction of published values. The last two values in Table 2 were suppressed due to disclosure risk. While the one in the 6th row is a primary suppression, the one on the 5th row is a complementary suppression. Given a set of sensitive cells, the goal in cell suppression is to find a set of complementary suppressions that maximize the data utility while providing sufficient protection. The Census Bureau uses linear programming techniques applied to multi-dimensional tables to achieve this goal.

DATA SWAPPING

The main procedure used for protecting decennial census and ACS tabulations is data swapping (Dalenius and Reiss, 1982). In each case, a small percent of household records are swapped. Pairs of households that are in different geographic regions are swapped across geographies. The Census Bureau uses a combination of targeted and random swaps, but the number of targeted swaps is generally higher. Targeted swaps focus on households that are viewed as at-risk of disclosure. For example, these could include households in very small geographic areas and those that are racially isolated. These household records are 'flagged' as potential swap candidates. A smaller percentage of randomly picked households are also 'flagged' as potential swap candidates. The flagged households are paired with households in the same state, but in different tracts. Only households that match on a cross tabulation of certain (key) variables are allowed to be paired. Next the pairs are ranked so that pairs with the highest disclosure risk and geographic proximity are given the highest rank. The decision variables in the swapping programme are the swap rate, the flagging variables, the matching key variables and the ranking of pairs. These variables can be altered to change the results of the programme. The

selection of these quantities plays an important role in preserving data quality while reducing disclosure risk.

- A small swap rate will ensure statistical accuracy of the data, but might not provide disclosure protection.
- If more criteria are used for flagging households then it will reduce potential disclosure risk, but more pairs will get deselected at a fixed swap rate. This could increase the possibility of not swapping a household that is at a high disclosure risk.
- If more variables are added to the matching key, it will improve the data quality but it will hinder the pairing process and increase the number of match key uniques. On the other hand, if fewer variables are used when matching, it will weaken the data quality while helping the pairing efficiency.

SYNTHETIC DATA

The Census Bureau uses partially synthetic data to protect ACS group quarters (GQ) data (Hawala, 2008). First the values of identifying variables (variables that are commonly available) are deleted from the at-risk respondent records and are treated as missing. Next, multiple imputation techniques are used to impute these missing values. To achieve this, the blanked values are replaced with a random sample of remaining values of the variable, and then predicted values are obtained for all respondents for this variable based on a model that uses the other non-identifying variables as predictors. Predictive mean matching is used to find synthetic values for the blanked values. For each blanked value, the absolute distances between its predicted and the other predicted values of the non-blanked data values are computed. Then the synthetic value for the deleted value is taken to be the observed value of the respondent that has the closest predicted value. Once values are imputed for a variable it is also used as a predictor in the next stage.

If the blanked value is for a continuous variable, a generalized additive model is used to compute predicted values. If the blanked value is an ordinal variable, a proportional logistic model is used draw prediction values, and if the blanked variable is categorical, a logit model for multinomial responses is used to draw prediction categories. When synthesizing variables it is also necessary to impose constraints on the model in order to prevent illogical response combinations in the

data. For example, it is not possible to allow a synthetically generated age to be 10 years for a mother of three children.

CONCLUSION

The disclosure avoidance procedures used by the Census Bureau depend on the type of data, the data products released to the public, and the tolerable risk threshold.

Cell suppression has no effect on unsuppressed cells, thus the values that appear in the published cells are in fact the unaltered values. The marginal totals will be preserved and there will be consistency between tables. But cell suppression causes additional loss of statistical information because of the complimentary suppressions that must be done. Cell suppression is applied at the tabular level. Since the Census Bureau releases a large amount of tables, cell suppression is difficult to implement even though the process has been automated.

Data swapping is applied at the micro data level, so once implemented, the Census Bureau is able to make all tables it requires. Swapping will preserve the marginal totals for the key variables and associations between the key variables at any level of geography. It will preserve the data completely at the state level. There will be consistency between all the released tables. But swapping will not preserve the marginal totals of non-key variables for small geographies. Also, the associations between non-key variables and key variables will not be maintained at these lower geographies. While a lower swap rate will increase the data quality, it will also increase the disclosure risk.

The synthetic data allowed the Census Bureau to mask Group Quarters since data swapping could not be used due to the difficulty in finding swapping partners. When an analysis is done by a data user, using the model that the synthesizer used, the loss in data utility will be a minimum. But if the user conducts an analysis on a more complex model, then it can lead to less accurate results. Since the data is only partially synthetic, there still could be some disclosure risk in the data products, but by targeting high risk records and variables, the risk will be minimized.

REFERENCES

Cox L.H. (1980). Suppression methodology and statistical disclosure control, *Journal of American Statistical Association* 75: 377-385.

Dalenius, T. and S.P. Reiss (1982), Data swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6: 73-85.

Hawala, S. (2008). Producing Partially Synthetic Data to Avoid Disclosure. Proceedings of the Section on Government Statistics, American Statistical Association.

Zayatz, L.V. (2005). Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update," Research Report Series, 2005-06, U.S. Census Bureau, Washington D.C.