Frequency Tables Disclosure Control for the Abu Dhabi Census 2011

Hanan AlDarmaki and Miriam Hodge
Statistics Centre – Abu Dhabi, Abu Dhabi, UAE
Email: hialdarmaki@scad.ae mchodge@scad.ae

Abstract

Statistics Centre-Abu Dhabi (SCAD) conducted a comprehensive population census for the emirate of Abu Dhabi in 2011 and released the results using online dissemination tools which allow the public to create frequency tables from census data down to small geographic areas having as few as 500 individuals. While increasing data availability, the flexibility of these tools introduced new risks of disclosing information about individual respondents. This paper presents a discussion of the risks for providing access to ad-hoc report generation and the method used by SCAD to mitigate these risks. The method presented is based on random rounding and controls for consistency across multiple output channels. While this method is applied on each output table individually, it ensures that the same frequency count in any table is always rounded to the same value.

Key Words: confidentiality, random rounding

1. Introduction

Collecting and publishing statistics about the population are the main functions of an official statistical organization. However, privacy concerns impose various limitations on the publication of collected data, and this tradeoff between data availability and confidentiality is one of the main challenges in statistical dissemination. SCAD's output tools for 2011 census data, which are designed to be flexible and user oriented, introduced new concerns of confidentiality, and various statistical disclosure control methods were reviewed to come up with a useful compromise. Some of the issues involved in the selection of the most suitable disclosure control method are: feasibility of implementation, transparency, consistency, and information availability.

This paper describes SCAD's output tools, some of the risks involved as a result of providing these tools to a wider public, a description of the most common approaches in statistical disclosure control, and the method used by SCAD as a variation of random rounding. Assessing the privacy obtained by implementing this method is beyond the scope of this paper.

2. Background

To increase the availability of Abu Dhabi Census 2011 data, SCAD has designed a number of online dissemination tools. These tools are named Table builder, Thematic maps, and Community tables, and they produce aggregate forms of census data; microdata are not yet provided. The Thematic maps tool provides a geographic representation of a predefined list of indicators. The tool allows users to find hotspots in the data and understand the geographic variation for selected indicators. The Community tables tool gives users access to predefined tables with customized geography. The customized geography allows users to view data aggregated for a number of minimum census geographies.

The largest disclosure risk comes from the Table builder tool. This tool is a real time online tool that produces frequency tables according to user needs. Users can select a geographic area for analysis and a list of census variables to create a custom cross tabulation. An example of a table builder cross tabulation is shown in Figure 1.

Usual Residents For Statistcal District by Age by 10 year age groups and Gender

Age by 10 year age groups	0-9		10-19		
Gender of Respondent	Male	Female	Male	Female	
Statistical District	---Usual Residents---				Row Total
District a	1,080	1,005	840	830	3,760
District b	395	355	380	220	1,350
District c	350	345	340	300	1,335
Column Total	1,820	1,710	1,560	1,345	6,445

Figure 1: Example of Table builder output

A basic level of disclosure control is applied on the raw data used to produce these tables. Sensitive identifying variables like names and home addresses are removed from the output dataset, and other detailed variables like age are categorized and output as age groups. The geographical aggregation of the output provides another layer of security since the smallest geographical unit is a sector, which has a minimum population of 500 individuals and 5 households.

3. Disclosure Risks in Census Data

In a sample survey, the data is protected by the sample design and population estimates, so the final weighted data cannot be mapped to individual respondents. In a census, however, each individual is represented once and the data is disseminated as collected, which puts the individuals at risk of identification even with the identifying variables removed.

By disseminating the data in frequency tables, the aggregation protects the individual respondents from being identified, especially in large geographical areas. As the dataset gets smaller, by geographical filtering for example, small cell values can be observed, which can lead to identification, especially if certain attributes or a combination of attributes are unique to one person in the specified geographical area. For example: a frequency table showing the number of female citizens holding a PhD in a statistical sector consisting of 500 persons, where there is only one female citizen holding a PhD; identifying the person from this table can lead to identification of other previously unknown information related to the same individual.

4. Review for Statistical Disclosure Control

In addition to protecting individual data, a number of other considerations need to be accounted for in a suitable disclosure control design. Data availability and accuracy are important characteristics of dissemination tools, so the method must avoid distorting the distribution of the data and minimize information loss. Another important characteristic is consistency in the same frequency table and among different tables from the same dataset. Also, respondents need to be ensured about the privacy of the data, and end users need to account for the distortions in their analysis, so the method must be transparent and easy to explain. In addition, simplicity and feasibility of implementation in a real-time system will ensure proper and efficient delivery of output.

A number of approaches have been proposed and used to control disclosure risks for frequency tables. Some disclosure control methods modify the underlying microdata and then produce frequency tables on the modified data; these are called pre-tabular methods. Eliminating high risk records, data swapping, randomized responses, and synthetic data are some examples of pre-tabular approaches (Matthews, 2011). The advantage of such methods is that they are applied once on the data and no processing is required in real time. On the other hand, these approaches may distort the distribution of the data, and it can be difficult to locate all risks and ensure confidentiality with these methods alone.

Post-tabular disclosure control methods modify the resultant frequency tables. Cell suppression, cell perturbation and random rounding are a few examples of such methods. In cell suppression, sensitive values in the output table are removed and displayed as missing values (primary suppression). Secondary suppression on safe cells is required to ensure that the sensitive data cannot be guessed from marginal sums or other frequency tables. This can lead to increased information loss due to secondary suppression, and it can also be difficult to find an optimal solution when a large number of tables is produced, and confidentiality could be compromised as a result (ESSNet, 2010). In cell perturbation, some noise is added to all cells in the frequency table. This ensures that users cannot be certain about the real values of small cells. One disadvantage of this method is that it is less transparent and hard to explain for analysis (Salazar-Gonzalez, 2006). Other complex approaches are discussed in (Matthews, 2011). Random rounding is a common post-tabular method used to protect frequency tables, and it's explained in detail in the next section.

5. Review of random rounding.

This method is easy to implement and explain and provides the benefits of disclosure control without dramatically distorting the data. Conventional rounding to the nearest multiple of the rounding base provides minimum loss of information at the expense of effective protection since it reduces the range of possible values. More elaborate techniques can be used to overcome the limitations of simple rounding like random rounding or controlled random rounding (ESSNet, 2010).

In random rounding, a value is rounded up or down in a probabilistic random manner, where the probability of rounding to either multiple of the rounding base depends on the difference between the original value and each possible round number. For example, to round a number 23 with a rounding base of 10, the value will be rounded to 30 with probability 3/10 and to 20 with probability 7/10. This ensures minimum loss of information in most rounded values, while introducing an additional level of uncertainty to each rounded cell.

Since each table is handled separately, random rounding can result in the same cell value in different tables being rounded differently. Also, assuming unlimited access to an online table builder, the same table can be generated several times, and by observing the frequency of the rounded values, the original value can be narrowed down to a smaller range or even discovered. For this reason, consistency of rounded values is important for confidentiality protection.

Also, rounding table totals separately, which minimizes information loss, results in inconsistent tables where the cells do not add up to the totals. Controlled rounding can be used to ensure that table sums match cell totals, with the added complexity of implementation and the difficulty to keep it consistent across different tables (Salazar-Gonzalez, 2006).

Other disadvantages of random rounding methods include loss of information since all cells are adjusted even if they do not introduce a disclosure risk. However, in large datasets like a census, using a small base of 5 results in a small amount of information loss, and the distribution of data is not largely affected.

6. Consistent random rounding method

The approach used in SCAD is a variation of random rounding described in the previous section. This implementation ensures consistency between different tables that contain the same counts, while also ensuring randomness for the probability distribution.

Each cell in a frequency table can be a count of individuals, households, units, or buildings. Each of these record types has a unique identifier. The unique identifiers in the raw data are used to derive a random number that is eventually used for rounding. Random rounding is used to round the numbers up or down to a multiple of 5 according to the appropriate probabilities.

6.1 Probability distribution

The probability that a cell value x is rounded up to a multiple of 5 is (x mod 5) / 5. If x mod 5 is 0, the value remains unchanged. For example, if the residual after dividing the cell value by 5 is 3, the value will be rounded up with probability 3/5 or 60% of the time.

To achieve this probability distribution in a programming environment, a pseudorandom number generator (PRNG) is used to simulate a uniform distribution. If a PRNG generates random values from 0 to 1 in a uniform manner, then 50% of the numbers generated will be less than 0.5, 60% less 0.6, and so on.

6.2 MD5 as a random number generator

MD5 is a cryptographic hash function commonly used to check data integrity and store encrypted passwords (Rivest, 1992). It is used here as a pseudorandom number generator. The input of the function is a string of any length and the output is 128 bits hash value. The output of MD5 is usually expressed as a 32 digit hexadecimal number. The benefit of using this function as a PRNG is that the same input (seed) value will always produce the same pseudorandom number, which ensures that the same cell is always rounded the same way in any table. At the same time, the value is created randomly and cannot be predicted without access to the seed. The seed value is the last five digits of the record identifier. The seed of a group of records is the sum seeds of the records in the group.

The steps to confidentialize the data are as follows:

1. The 5 least significant digits of the identifier are extracted as the seed.
2. The seed values for all the records contributing to the cell value are summed.
3. The sum of the seed values is used as an input to the MD5 hash function.
4. The resultant hash value is used to decide whether to round a cell value up or down, according to the probability distribution for the cell value.

It is important that the random number generator is uniform and dense over the range of possible values. To assess the uniformity of the MD5 function the distribution of n=10,024 random numbers was generated using the MD5 function

$$f(t) = MD5(t).$$

Each of the n random numbers is generated by the MD5 function using the input t, which is the sum of the last five digits of the record identifiers,

$$t = \sum_{i=1}^{T} k,$$

$$T = unif(0,\ 2 * 10^6)$$

where T is a random number of records, and k is the last five digits of the record identifier. The result of this analysis is shown in Figure2. Each of the $n = 10{,}024$ results are plotted, where the x-axis is the order that the numbers are generated $n = [0, 10024]$ and the y-axis is the value of the random number divided by the maximum value of the MD5 range $(2 * 10^6)$ so that the range is between 0 and 1. The plotted points confirm that the distribution of the numbers generated by the MD5 function is dense and uniform.

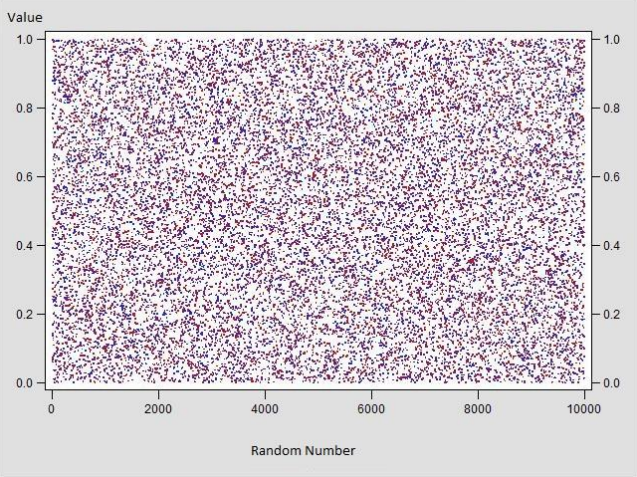

Figure 2: Distribution of 10,024 MD5 numbers using the sum of a random number of member ids

The goal of confidentialization is to preserve the characteristics of the data. The confidentialized data should have the same expected values as the original data, in other words, the expected value of the adjustment should be zero.

$$\mathrm{E[adj]} = \mathrm{p_0} \cdot \mathrm{E[n \bmod m = 0]} + \mathrm{p_1} \cdot \mathrm{E[n \bmod m = 1]} + \cdots$$

$$+\mathrm{p_{m-1}} \cdot \mathrm{E[n \bmod m = (m-1)]}$$

Where m is the base (in the implementation $m = 5$), $\mathrm{p_0}$ is the probability that the mod of the cell is equal to 0 and $\mathrm{E[n \bmod m = 0]}$ is the expected adjusted value of a cell whose mod is 0, $\mathrm{p_1}$ is the probability that the mod of the cell is equal to 1 and $\mathrm{E[n \bmod m = 1]}$ is the expected adjusted value of a cell whose mod is 1, and so on. The mod of any cell must be between 0 and $m - 1$, therefore the probability of these must sum to one,

$$\mathrm{p_0} + \mathrm{p_1} + \cdots + \mathrm{p_{m-1}} = 1.$$

The expected value for the adjustment of each of the mod values is zero,

$$\mathrm{E[n \bmod m =\ r]} = \frac{\mathrm{m - r}}{\mathrm{m}} \cdot (-\mathrm{r}) + \frac{\mathrm{r}}{\mathrm{m}} \cdot (+\mathrm{m - r}) =\ 0,$$

where r is the result of the mod, and m is the base. The overall expected value is also

zero,

$$E[adj] = 0.$$

To test that the program is producing the expected results, a frequency analysis was done on N=10,024 samples from the data. The results are shown in Table 1.

n Mod 5	trials	mean
0	1989	0
1	2111	-0.03126
2	1983	0.039838
3	1918	0.031803
4	2023	-0.06524
All	10024	-0.00579

Table 1: Experimental results on data adjustment of 10,024 trials

7. Conclusions

The method developed by SCAD for disclosure control of frequency tables met the desired goals of implementation simplicity, transparency, consistency, and information availability. Experimental results showed that the distribution of the data was preserved as a result of applying random rounding using MD5 hash function as a random number generator. The consistency achieved by implementing this method in Table builder allowed SCAD to provide small area results from census while reducing the risk of disclosing identifying data.

8. References

Handbook on Statistical Disclosure Control. (2010). ESSNet SDC.

Knuth, D. (1997). *The Art of Computer Programming.* 3rd. edition, chapter 3, pp. 1-193: Addison-Wesley.

Matthews, G. J. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, 1-29.

Rivest, R. (1992). *The MD5 Message-Digest Algorithm.* RFC 1321: MIT Laboratory for Computer Science.

Salazar-Gonzalez, J.-J. (2006). Controlled rounding and cell perturbation: statistical disclosure limitation methods for tabular data. *Mathematical Programming*, 583-603.