An implementation of a cell suppression algorithm for tabular data in R and its challenges

Institute of Statistical Mathematics　Kazuhiro Minami

1 Introduction

To perform statistical disclosure control (SDC) on tabular data is a difficult task because we need to make sure that each suppressed cell has s sufficient range of possible values under the presence of linear relations among cell variables in the tabular data. In addition, it is desirable to minimize the number of suppressed cells to reduce information loss of the original data. We therefore develop a SDC tool in *R* to supports both primary and secondary cell suppressions for tabular data.

2 Main functionality of the tool

We develop a prototype tool in R, which provides several public functions that allows users to perform primary and secondary cell suppressions on tabular data. We use the package *lpSolve* to solve linear programming problems. The major functions of the tool is listed as follows:

- Perform primary suppressions on a frequency table based on a threshold of the minimum unit number
- Perform primary suppressions on a magnitude table based on various occupancy rules, such as (n,k)-rule, p% rule and so on
- Perform secondary suppressions on both on frequency and magnitude tables
- Construct a magnitude table with auxiliary information such as the corresponding frequency table, the tables of the largest and the second largest unit values

We expect a researcher to use our tool to perform statistical disclosure control on tabular data and prepare all necessary information for output checking as shown in Figure 1.

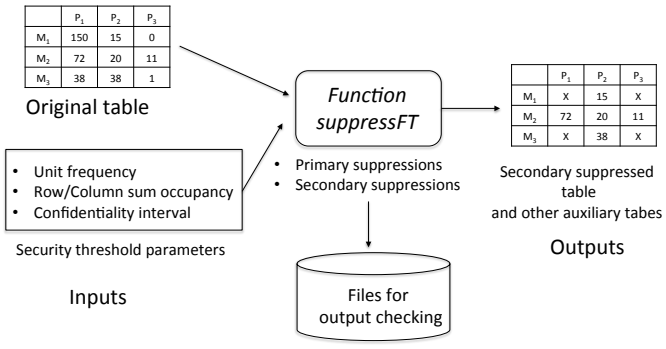

Figure 1: Cell suppressions with the function *suppressFT*

3 Future work

Our secondary suppression algorithm is a greedy-based approximation algorithm, which tends to suppress more cells than the optimal one. We thus plan to develop a new algorithm that produces the optimal solution.