Abstract of the research on PRAM and the disclosure protection of microdata in the employee registrations

A new measure of safety - M.H.Rienstra

In this report PRAM was discussed. The P(ost) RA(ndomization) M(ethod) is a new method in order to protect microdata. When applying PRAM, every value of a variable has a certain probability to change in another value according to a prescribed probability mechanism. Because of this probability mechanism the traditional safety rules cannot be used. Especially the rarity rule provides difficulties. This rule prescribes a minimal frequency (a so-called threshold) of a combination of variables in the safe file, thus after application of protection methods. When PRAM is used, the frequencies of combinations are stochastic. For this reason a new measure was introduced. This new measure is based on the probability of a combination to change in another, combined with the frequency of the combination in the original file.

This new measure is called 'Expectation Section' (ES). The ES respresents the probability of a certain combination in the released file (after applying PRAM) to have remained unchanged compared to the original file (before applying PRAM). The ES has been used to define new rules of safety. Two rules are introduced.

The first safety rule corresponds to a threshold value, which for instance can be chosen to be $ES \leq 0.5$. A threshold of 0.5 will intuitively provide a large uncertainty level, which reduces the chance of the file being treated as the original file.

Another safety rule uses the relation between the frequency of a rare combination and the threshold (the minimal frequency). Combined with the ES this provides the second safety rule. This measure will be illustrated with an example. In this example the threshold will be taken to be equal to 10. This means that a combination with a frequency equal to 10 in the original file is safe. A combination which frequency lies below this threshold has to be protected. We now can make a difference between combinations that occur only once in the original file and combinations which frequency lies just below this threshold (9, for instance). The latter are more safe since they can be linked between themselves. Therefore the frequency and the threshold do influence the level of safety. Furthermore we need the ES to provide information about probabilities when this measure is used on data protected by PRAM.

In this report we recommend the second rule. The first rule ($ES \leq 0.5$) is useful as well, though, because it also gives us much information about the data. The following two documents use both rules.

The off-diagonal probabilities in a PRAMmatrix - M.H.Rienstra

This report introduces a measure of information loss. We use the 1-dimensional variance (which is the variance per variable), with which we can compute the so-called variation coefficient.

All probabilities of change from one category into another are elements of a so-called PRAMmatrix. This matrix consists of probabilities of a category to remain unchanged (leading diagonal elements) and probabilities of change into another category (off-diagonal elements). PRAMmatrices can be combined into matrices which provide probabilities of combinations of categories.

This report discusses several methods to arrange the off-diagonal probabilities in a PRAMmatrix. Two methods have been compared empirically, using the employee registrations '98. The first method assigns equal probabilities to each off-diagonal category. The second method uses the frequencies of the categories (not a combination of categories) in the original file. This method is based on the idea that categories with high frequencies in the original file should be used to provide protection to the categories with low frequencies. This second method turns out to yield the best results. Hence we recommend the use of this second method when protecting the employee registrations. In the next report this practice will be discussed.

Protection of the employee registrations by PRAM - M.H.Rienstra

This report describes the use of PRAM in the protection of the employee registrations of 1998. The results of several PRAMmatrices have been tested against the prescribed measures of safety and information loss. The research has shown that in order to satisfy the rules, PRAM has to be applied on all variables. Moreover probabilities on the leading diagonal are needed to be in order of 0.7. Not all rare combinations can achieve the ES threshold value. Yet this number of combinations with high ES value is very small. This suggests a slightly more flexible safety rule, for instance to allow a somewhat higher ES at a certain percentage of the combinations. Doing this, one has to take into account the type of these rare combinations. An ES of 0.9 will sometimes satisfy the original safety rule when one of the variables scores is 'unknown'.

The measure of information loss is not often exceeded. Yet when it is, one could for instance gather categories of a variable. Here, the users wish has to be taken into account.

From this research we can conclude that PRAM can be used in protection of

microdata. However the mode of application is dependent on the file. For this reason we have to deal with each file seperately.

Further research questions are the following. What influence does the adaption of the diagonal probabilities per category have on the results of the ES? What should be the exact way of delimiting the ES? Another important issue is PRAM from a users point of view. In chapter 'Conclusions and recommendations' some suggestions for the continuation of this research have been made.