Statistical Disclosure Control: Applications in Healthcare

James Miller

A Dissertation Submitted to the

University of Glasgow

for the degree of

Master of Science

Department of Statistics

December 2005

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Acknowledgements

A great deal of thanks goes to ISD Scotland from whom the funding was received which allowed this thesis to exist. I hope that my work will prove useful as the organisation continues to progress their thoughts and ideas in the statistical disclosure control field and if nothing else my work has prompted discussion which may prove helpful in the future. Special thanks go to Mary Sweetland and Peter Knight who proved valuable contact points in the organisation and shared their many ideas on the subject. Also thanks goes to Margaret MacLeod who supplied the data used in the thesis and was both friendly and helpful throughout.

The encouragement, support and most of all patience of my supervisor Professor E. Marian Scott were crucial elements in the production of this thesis. I cannot speak highly enough of Marian's ability to make time for all students, and especially myself, in what has proven to be a busy year. I would also like to thank Marian for her help on a personal level in the difficult times I had throughout the production of this thesis and let her know that all her support was much appreciated.

Finally I would like to thank my family for the example they have set for me in terms of love, respect and most of all strength of character. Through what has been a difficult time your actions have inspired me to put more into both life and work. I thank you all for not asking any more of me than I had to give and want you to know that the production of this thesis only came about thanks to your support and encouragement.

Abstract

Statistical disclosure control is a progressive subject which offers techniques with which tables of data intended for public release can be protected from the threat of disclosure. In this sense disclosure will usually mean information on an individual subject being revealed by the release of a table. The techniques used centre around detecting potential disclosure in a table and then removing this disclosure by somehow adjusting the original table.

This thesis has been produced in conjunction with Information and Services Division (Scotland) (ISD) and therefore will concentrate on the applications of statistical disclosure control in the field of healthcare with particular reference to the problems encountered by ISD. The thesis predominately aims to give an overview of current statistical disclosure control techniques. It will investigate how these techniques would work in the ISD scenario and will ultimately aim to provide ISD with advice on how they should proceed in any future update of their statistical disclosure control policy.

Chapter 1 introduces statistical disclosure and investigates some of the legal and social issues associated with the field. It also provides information on the techniques which are used by other organisations worldwide. Further there is an introduction to both the ISD scenario and a leading computing package in the area, Tau-Argus.

Chapter 2 gives an overview of the techniques currently used in statistical disclosure control. This overview includes technical justification for the techniques along with

the advantages and disadvantages associated with using each technique. Chapter 3 provides a decision rule approach to the selection of disclosure control techniques described in Chapter 2 and much of Chapter 3 revolves around a description of the implications derived from the choices made.

Chapter 4 presents the results from an application of statistical disclosure control techniques to a real ISD data set concerned with diabetes in children in Scotland. The results include a quantification of the information lost in the table when the disclosure control technique is applied. The investigation concentrated on two and three-dimensional tables and the analysis was carried out using the Tau-Argus computing package.

Chapter 5 concludes by providing a summary of the main findings of the thesis and providing recommendations based on these findings. There is also a discussion of potential further study which may be useful to ISD as they attempt to update their statistical disclosure control policy.

Contents

List of Tables

List of Figures

1. Issues in Statistical Disclosure Control

1.1 Need for Statistical Disclosure Control

In any Statistical survey the aim of the statistician is to provide information on a set of subjects to the general public. These subjects can range from individuals or households to businesses or organisations. What is common to all these subjects is that the information they supply should be provided in confidence. Any respectable data collector will keep all individual records confidential and this confidence is often essential to the quality of the data provided since such a promise should encourage participants to not only cooperate fully with the statistician but to provide accurate data. The concealment of data from potential intruders is the sole responsibility of the statistician and the statistical organisation they work for. This concealment has become much more difficult as both statistical methodology and computing technology have advanced. Statistical organisations need more detailed data to be released to carry out more recent and sophisticated techniques in certain areas such as modern advanced modelling techniques. For these data to be released the statistical organisations must have in place systems to protect each individual subject's data. The huge strides taken in computing technology, in particular the exponential increase in size of the World Wide Web, has also ensured that there are vast amounts of auxiliary data available to potential intruders.

It may be hard to believe that intruders to statistical studies exist and that they will go to great lengths to discover information on a subject but there are situations where it may be of great benefit to an organisation to gather sensitive data on individuals.

Take for example a study on annual turnover of businesses; it may be very helpful for a company preparing a potential takeover of a company to know the exact (or at least a good estimate of) turnover for that company. In the field of health care there are potential difficulties if for example life insurance companies could uncover information on individuals from statistical studies and alter their premiums accordingly. There is an on going argument between statisticians and in fact politicians as to what extent a statistical organisation should be expected to compensate for an intruder who is armed with masses of auxiliary data and whether information uncovered by such an intruder is indeed a breach of confidentiality. Nevertheless the issue of concealment of an individual's data is clearly important and there are in fact laws (to be discussed in section 1.2) governing the processing and dissemination of results from studies.

Another interesting point of discussion is what actually constitutes a potential disclosure risk. A glossary presented at the Joint ECE/EUROSTAT Work Session on Statistical Data Confidentiality [1] defines the following forms of disclosure:

'Disclosure: This relates to an inappropriate attribution of information to a data subject, whether an individual or an organisation.

Disclosure by matching: Disclosure can be accomplished with high resolution keys by matching the data set with a register which contains the keys and names and addresses.

Disclosure by response knowledge: This is the knowledge that a person was interviewed for a particular survey; if an investigator knows that a specific individual has participated in the survey, and that consequently his or her data are in the data set, identification and disclosure can be accomplished more easily.

Disclosure by spontaneous recognition: This means the recognition of rare persons; disclosure may occur by accident.'

Clearly there are many forms of potential disclosure. Each form of disclosure must be considered when thinking about how to protect an individual's data. There are certainly many issues to consider when defining what constitutes disclosure. For example there could be occasions where an outside intruder cannot discover information about an individual from a table but one subject in the table can discover the information about another subject they know is in the same table. It could be argued that it is difficult for the statistical organisation to protect against disclosure to another individual and that doing so would be too restrictive so protecting only against disclosure from an outside intruder is sufficient. On the other hand, many members of a study would not want it to be possible for another member in the study to infer further information about them so the statistical organisation should protect against this form of disclosure as well. It is generally accepted however that when there is any doubt, the statistical organisation should always lean towards over protection rather than under protection. In fact Dalenius (1978) explains statistical disclosure by saying 'statistical disclosure occurs when the release of a data product enables a third party to learn more about a respondent than originally known.' This would seem to imply that if any person (fellow subject or intruder) can infer extra

information on an individual due to the release of a table then this constitutes an unacceptable disclosure of this individual's personal data.

In response to these issues statisticians have been forced to discover various techniques to provide protection of the confidentiality rights of subjects. The blanket term for these techniques is statistical disclosure control. Statistical disclosure control is basically the issue of balancing the need to know for the greater good of society and the right to privacy of the subject. There are many issues in the process of statistical disclosure control. The Government Statistical Service Report of the Task Force on Disclosure [2] defines the process in this way;

'There are four processes in the reduction of disclosure risk, namely:

(i) the recognition of risk of disclosure from tables

(ii) the assessment of the degree of risk involved

(iii) the selection of an appropriate method to eliminate or, at least, reduce the risk of disclosure of tables

(iv) the quantification of any loss of information.'

1.2 History and Legal Aspects of Statistical Disclosure Control

Although statistical disclosure control has only become a serious concern in the last 20 years due to the increased power of computers and wide access to data due to the World Wide Web the issue dates back to the middle of the 20th century. A study of various forms of legislation by Scherff [3] in 1952 implies that a lack of confidentiality would result in distrust from respondents and therefore inaccurate data

would be provided if any data were provided at all. These arguments are very similar to those used today.

It was in the 1970's that the Western world first saw how the power of modern technology could affect the privacy of individuals. The web of political scandal in the USA from 1972 to 1974 known as the Watergate scandals showed the possibility for the authorities to intrude in the lives of individuals to discover a large amount of sensitive information. This coupled with increased knowledge of CIA and FBI surveillance and exponential growth of information technology resulted in the Privacy Act of 1974 [4]. The Privacy Act states: 'The purpose of this act is to provide certain safeguards for an individual against invasion of personal privacy by requiring Federal agencies... ....to collect, maintain, use or disseminate any record of identifiable personal information in a manner that assures that such action is for necessary and lawful purpose, that the information is current and accurate for its intended use, and that adequate safeguards are provided to prevent misuse of such information.' In 1980 Europe produced its own cross border regulations for the protection of data. This was found in the OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data [5] produced in 1980 which stated: 'The development of automatic data processing, which enables vast quantities of data to be transmitted within seconds across national frontiers, and indeed across continents, has made it necessary to consider privacy protection in relation to personal data. Privacy protection laws have been introduced, or will be introduced shortly, in approximately one half of OECD Member countries (Austria, Canada, Denmark, France, Germany, Luxembourg, Norway, Sweden and the United States have passed legislation. Belgium, Iceland, the Netherlands, Spain and Switzerland have prepared draft bills) to

prevent what are considered to be violations of fundamental human rights, such as the unlawful storage of personal data, the storage of inaccurate personal data, or the abuse or unauthorised disclosure of such data... ....OECD Member countries considered it necessary to develop Guidelines which would help to harmonise national privacy legislation and, while upholding such human rights, would at the same time prevent interruptions in international flows of data.' This was shortly followed by the Council of Europe's Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data in January 1981 [6]. This convention stated: 'The purpose of this convention is to secure in the territory of each Party for every individual, whatever his nationality or residence, respect for his rights and fundamental freedoms, and in particular his right to privacy, with regard to automatic processing of personal data relating to him ("data protection").' All of these acts were created to protect individual privacy in the face of an increase in the available technology to allow records to be transferred swiftly and easily. They were a reaction to the change in technology since previously records were often held by an organisation on paper or on their own computer network with the possibility of large amounts of information being shared considered negligible due to the logistical problems. These bills were all enacted in the mid 1970's to early 1980's and technology for data recording and transfer has moved forward since then. In fact in 1999 UC Berkley predicted that more data would be stored between 1999 and 2002 than in the whole history of data storage. It is also of interest to note that international organisations required acts to prevent the release of information in cross border scenarios.

The first UK bill to deal with the confidentiality of individual's data was the Data Protection Act of 1984. This has now been superseded by the Data Protection Act of 1998. The act requires personal data to be

- processed fairly and lawfully
- obtained only for a specific lawful purpose
- adequate, relevant and not excessive
- accurate and up to date
- kept for no longer than necessary
- processed in accordance with the rights of data subjects
- kept secure
- transferred outside the EEA only if there is adequate protection

A survey carried out by the UK Information Commission [7] found that the general public still regarded privacy an important issue in society and believe that it is a human right. The survey also found that the public viewed the general requirements of privacy as very similar to those laid out in the Privacy Act of 1998.

As recently as 2003 the EU produced a directive, The Privacy and Electronic Communications (EC Directive) Regulations 2003, to particularise some of the legislation to deal with the issues in improved telecommunication technology. This illustrates some of the problems facing legislators protecting against disclosure control. The advancement in data storage and transfer technology means that laws and practices for preventing disclosure of individual data must evolve rapidly to keep pace with this advancement. This places extra burden on statistical organisations to

ensure that they do not act in a way which either contravenes the new legislation or allows for potential disclosure as a result of new technology.

The most recent law to influence the process of disclosure control is the introduction in the UK of the Freedom of Information Act 2000 which came into force in 2005. The act allows an organisation or an individual to request information from organisations which hold data and this request must be granted unless the request falls under an exemption. In general this now means that if an organisation does not wish to divulge information it must provide a reason rather than simply reject the request. This change in the law has made it necessary for statistical organisations to have in place stringent disclosure control policy to avoid a potential disclosure through a forced release of data.

A major legal problem found by statistical organisations is the definitions of some of the terms in statistical disclosure control. For example what is disclosure? Dr Jan Holvast writes [8]'Cox [9] has given the most universal description. In his view, disclosure is breaching the pledge of confidentiality by revealing confidential respondent data.' This is very vague and does not answer many legal questions. There are many ways in which data can be revealed and a decision must be made on which of these are as a result of negligence on the part of the statistical organisation. This will be laid out in the many laws and treaties on the issue but as society and especially technology has evolved public perception of these issues may change. An important issue is the terminology pertaining to personal (or sensitive) data. This is important in terms of disclosure control since many of the laws and practices are in place to ensure that no sensitive data is revealed. What makes this issue so difficult is

that different people will find different topics sensitive. For example one subject in a survey may regard their employment status as a sensitive issue which should not be revealed to an outside party whereas another subject may have no problem with their employment status being revealed. This problem is a particular concern for the EU where countries may have different laws regarding what information on a citizen is public knowledge. For example in Sweden each citizen's annual salary is public knowledge whereas in the UK that information would be regarded as sensitive. In the UK The Data Protection Act 1998 [10] defines sensitive personal data:

'**2.** In this Act "sensitive personal data" means personal data consisting of information as to-

(a) the racial or ethnic origin of the data subject,

(b) his political opinions,

(c) his religious beliefs or other beliefs of a similar nature,

(d) whether he is a member of a trade union (within the meaning of the Trade Union and Labour Relations (Consolidation) Act 1992),

(e) his physical or mental health or condition,

(f) his sexual life,

(g) the commission or alleged commission by him of any offence, or

(h) any proceedings for any offence committed or alleged to have been committed by him, the disposal of such proceedings or the sentence of any court in such proceedings.'

Disclosure issues will without doubt change over the coming years as technology advances and public interest grows. Statistical organisations must ensure that their

policy follows evolving laws and necessary practices to ensure the privacy of respondents is maintained.

1.3 Social Issues associated with Statistical Disclosure Control

It is clear that the issue of statistical disclosure control has many social issues. The feeling among many statisticians is that certain types of data require much more stringent disclosure control procedures than others. For example in the field of health care clearly there are certain topics such as a subject's abortion history, especially if the data was at school level for example, or HIV category which would be deemed as sensitive whilst a topic such as a subject's history of in grown toe nails is not at all sensitive. In fact in a leading statistical disclosure computing package (Tau-Argus which is to be discussed later) the option is given to describe variables as 'very sensitive'. A comment given by Hundepool and Willenborg [11] on the assignation of these variables in the Tau-Argus program is that '*very sensitive variables* are sometimes taboo in microdata sets – such as the public use files at Statistics Netherlands -, particularly if there are several very sensitive categories (such as 'suicide' for 'cause of death').' Clearly there are differences in how sensitive different variables are but this is often a very personal opinion where there are many topics that one person would find sensitive that others would not. The difficulty for the statistician is that they do not know whether the data provided is deemed very sensitive by the subject or whether its potential disclosure does not cause the subject any real concern. As always the statistician should err on the side of over protecting information about an individual subject rather than risk disclosure.

Two important issues in the problem of statistical disclosure control are the protection of an individual's data and the attempt to release as much information to society as possible. Clearly there is a conflict between these two issues. The simplest way to protect the confidentiality of each individual's data is to not release any information gathered by the statistical agency to society although clearly this would also be unethical since the 'greater good of society' would suffer. This leaves the statistician with a trade off to make between the potential harm to the individual due to disclosure of their data and the potential harm to society caused by a lack of information. It is only recently that statisticians have begun to use statistical methods to address the problem of disclosure control. Fienberg [12] writes 'For far too long, confidentiality and disclosure limitation were relegated to the non-statistical part of large-scale data collection efforts and, as a consequence, the methods used to address them were ad hoc and conservative.' The use of statistical methods have allowed the release of improved levels of data free from disclosure risk although in some situations the data cannot be made safe and the release may disclose information about an individual. Cox [13] writes that in these situations 'incompatibilities are resolved in favour of preserving confidentiality'

The restriction of access to data collected by statistical organisations is also a social and ethical issue associated with statistical disclosure control. The main advantage of restricting access to the data is that the data would not have to be adjusted to avoid disclosure as access would only be granted to a few individuals who would have no interest in inferring information on certain individuals. Those who were granted access to the data would then be required to sign confidentiality agreements and have their use of the data monitored and subsequently controlled. This would allow

groups, such as policy makers in the government, to have access to the unadjusted data on which to base crucial policy decisions. Although in theory this would seem a reasonable suggestion there are many problems with it. Fienberg [12] writes 'Restricted access should only be justified in extreme situations where the confidentiality of data in the possession of an agency cannot be protected through some form of restriction on the information released.' A major argument for releasing as much data as possible is the hope that this will increase the trust of the public in the statistical organisations. If all the data is held by the organisation and only released to certain groups (e.g. policy makers or politicians) it could lead to a certain distrust of how the information is being used and result in questions about the relationship between the statistical organisation and the group and also call into question the integrity of the organisation. Any lack of trust in the organisations will result in poor (generally incorrect) data being submitted by subjects. Georges Als [14] writes 'Despite all the guarantees, the population distrusts the statistical services, and the responses it gives are no fuller and truer than those supplied to the Revenue or other authorities.' The more seemingly transparent the statistical organisations are the more likely subjects are to give time and thought towards the data they supply resulting in better responses.

Since the Freedom of Information Act [15] was introduced on January 1st 2005 it has become increasingly important for statistical offices to have in place stringent statistical disclosure control policies. These policies have become necessary since organisations will have more pressure placed on them than in the past by bodies requesting access to data, to which they are now entitled. In a talk given by Kevin Dunnion the Scottish Information Commissioner to the Edinburgh branch of the

Royal Statistical Society he stated that institutes must divulge information unless amongst other things:

1. the statistics were due to be published within 12 weeks
2. data were confidential – quality of confidence
3. data were personal data
4. data were part of a programme of research.

Clearly all these are important to statistical organisations but both 2 and 3 are examples of social issues connected with statistical disclosure control.

Statistical disclosure control is not a purely scientific topic. There are many social and ethical considerations to make. This thesis will concentrate mainly on the issues raised by the scientific workings of tests and procedures to deal with disclosure control but it is important that any statistical organisation constructing policy in the field keeps in mind the social and ethical ramifications of their work.

1.4 Issues Arising from Different Forms of Data Sets

The presentation of data in tables is a very common and practical way to disseminate data. Tables can come in all forms from a very simple 2x2 table to much more complex multi level tables. Most tables required by end users need the marginals to be present to allow both a more informed impression to be made of the data and for further analysis to be carried out. Unfortunately as with all public exposure of data tabular data carries a risk of disclosure. This risk generally occurs when the published

tables contain cells with low values and therefore few subjects although if many tables on the same data set are produced it may be possible for results to be matched and a subject recognised. Clearly the fewer subjects that are in a cell the greater the chance an infringement of their anonymity becomes. It would therefore be helpful to have efficient techniques to detect and remove this form of potential disclosure in tabular data. Fortunately these exist and this thesis will attempt to outline some of the techniques which are commonly used. It may be helpful to think of the process of reducing the disclosure risk in a table as having four stages:

1. detecting risk of disclosure from the table
2. discovering the degree of risk posed by the table
3. removing (or reducing) the risk of disclosure from the table
4. assessing the loss of information caused by removing the risk of disclosure

Stages 3 and 4 are the balancing act between publishing data that are both accurate and useful and the right of the subject to confidentiality. This thesis will attempt to cover methodology and issues raised by various techniques to deal with all four stages.

Many statistical organisations publish their findings in tabular form. Whether it be in published papers, on the web sites or in the national press, data in tabular form are simple, informative and user friendly. There is unfortunately the potential for individual's data to be disclosed by these tables. There are many separate issues to consider before protecting tables using the various techniques (these will be discussed later see chapter 2). One of these issues is whether the table contains magnitude data or frequency data. This distinction is essential to the techniques involved in

disclosure control and it is important that the difference is understood and the disclosure control process adjusted accordingly. A table that contains magnitude data (example *table 1.4.1*) is a table that quotes a quantitative value in each cell of a table, for example total business turnover or mean blood pressure. A table that contains frequency data (example *table 1.4.2*) is a table that simply quotes the number of subjects in a category in each cell, for example number of subjects who are male and earn over £30,000 a year or the number of subjects who are over 60 and have a clinically high blood pressure.

		Median Income(£)	
		Sex	
		Male	**Female**
Age Group	**30-40**	£21,000	£19,000
	40-50	£26,000	£23,000

Table 1.4.1: Fabricated example of table of magnitude data.

		Age Group	
		Under 60	**Over 60**
Blood Pressure Status	**High**	7	12
	Normal	45	32
	Low	2	14
	Total	54	58

Table 1.4.2: Fabricated example of table of frequency data.

The form of data in the tables makes a difference in both how potential disclosure of sensitive data is detected and also how the potential disclosure is removed. It is also important for the statistician to ensure that when multiple tables are being produced that the information from these tables cannot be combined to cause a disclosure.

A further form of data that may be found in a table is hierarchical data. A hierarchical variable is a categorical variable which contains various levels on the single data value. For example if there was a geography variable then the single variable may

contain three parts e.g. 'continent', 'country', 'region'. Then say the geographical area was Greater Glasgow that would be represented by a variable showing Europe, Scotland, Greater Glasgow e.g. EURSCOGG. Or say the geographical area was New York then this would be represented by North America, United States of America, New York e.g. NAUSANY. These hierarchical variables can be used to produce a hierarchical table which, as Willenborg and de Waal [16] write, 'is like an ordinary table with its marginals, except that it has additional subtotals (for each group at each level of the hierarchical variable).' A hierarchical table is simply a set of tables represented in one table using the hierarchy. The extra challenge presented by a hierarchical table in terms of disclosure control comes from the fact that the extra sub totals from the hierarchy can potentially be used to calculate the values of protected cells in the same way as an intruder could use the marginals to compute cell values.

An important issue with the design of the statistical disclosure control process is whether marginals are included in the table. The presence of marginals makes the table much easier to 'unpick' for any intruder wanting to make inferences. If no marginals are present any cells which risk disclosure could be removed from the table and the table published. Unfortunately having the marginal values included would allow any intruder, through relatively simple linear equations, to compute the values of these cells. This means the cells are no longer protected against the risk of disclosure. It therefore seems reasonable to exclude the marginals from the table however as mentioned earlier advances in statistical techniques have resulted in statistical organisations requiring more detailed data to carry out formal tests. For many formal tests and modelling procedures it is not sufficient for only tables with incomplete cell values but the table must contain all cell values and the marginals.

Many other users of the data (such as policy makers) also require complete tables with the marginals in order for the data to be useful.

It is important when carrying out the statistical disclosure control process that the needs of the final user of the data are discussed. Some of the various disclosure control methods may suit one user but useless to another user. For example if a newspaper was looking for a small table it would not want any of the cells in the table suppressed but may be more inclined to accept rounding of the cell values. Suppression is a statistical disclosure technique where cells in the table which are deemed sensitive are removed from the table. To allow the marginals to remain in the table further cells must be suppressed (a technique called secondary suppression) to avoid the value of the suppressed cell being attainable by simple linear equations (more detail on suppression can be found in section 2.2.2). Therefore if a group of policy makers were looking for a large table where only some of the information was crucial they might prefer suppression of some of the sensitive cells which were not crucial for their needs. This also highlights the need for the statistician to investigate the data before working with it. Certain techniques require a working knowledge of the data and in particular variables which can be grouped, are part of a hierarchy or have some unusual relationship.

1.5 Non-Statistical Methods for Avoiding Disclosure

Although this thesis will concentrate mainly on statistical methods for controlling disclosure risk it must be remembered that non-statistical methods also play an important role in the process. If these methods are used properly the need for

statistical intervention to the data can in some circumstances be greatly reduced with the result that more of the utility of the data is retained. It must be noted however that there are often disadvantages to these non-statistical methods as there are for statistical methods. These methods should though be central to any policy statistical organisations have on disclosure control. A good policy will usually contain a mixture of statistical and non-statistical methods for disclosure control.

One obvious method of disclosure control is to restrict access to the data. If the data is not released to the public there is clearly no (or at least a greatly reduced) risk of disclosure. For this to work everyone who has access to the data, whether they be the final policy maker or a data analyst, would be required to sign a legal disclaimer preventing them from revealing any of the data. This disclaimer could very easily be made part of a contract and in fact in many cases this will already be practice. The access to the data would then have to be monitored closely either by a specified employee or the management of the company to ensure no malpractice occurred. This method would allow the statisticians to work with the original data that were collected, this is an advantage since removing any potential error in the data will allow more accurate further analysis to be presented to policy makers. In theory this disclosure control method would seem to be advantageous to both the statisticians and society in general however in reality the public want to know how the data they provide are used and not have them analysed by statisticians behind close doors with no apparent end result. A mentioned in section 1.3 if the public loses trust in the statistical organisation the data they supply (if indeed they supply data at all) may have reduced accuracy. In a talk given by Gerald Gates [17] of the United Sates

Census Bureau it was suggested that the four major issues causing a negative public perception of privacy were:

- Fear
- Mistrust
- Misunderstanding
- Loss of Control

Any policy where access to the data is limited will only compound these negative public perceptions. In fact Gates also outlined an incident known as the Canadian record linkage incident where the data for American Citizens were being held and linked together without public knowledge and when this database was uncovered the public outcry was enormous. An interesting discussion centres on the public perception of privacy and whether the problem with this record linkage was the fact that a lot of data on each individual was being collected or whether the fact that it was done in secret was the major issue. In a talk given on a study on privacy violations and privacy perceptions in ubiquitous multimedia environments, Sasse [18] from the University College London found that people want feedback and control. Further more she commented on the fact that 'Users are usually prepared to accept some risks to privacy, if there are potential/expected benefits from the system that harbours the risk. Also it is important that users need to be aware of risk and have a chance to accept it.' This would seem to suggest that the public do not in general want organisations to adopt policy where access to the data is restricted but may be able to accept a degree of disclosure risk so long as there was a perceived benefit. This implies that statistical organisations should be as open as possible and take time to 'market' the work they are doing to convince the public it is in their interest.

A method which would allow both the raw data to be used, but in theory allow the public more access to the data, is one where a user can request certain statistical analysis of certain data from the data holder and the data holder must perform the test and provide the (edited slightly) results to the user. In the traditional sense Statistical Disclosure Control follows the form:

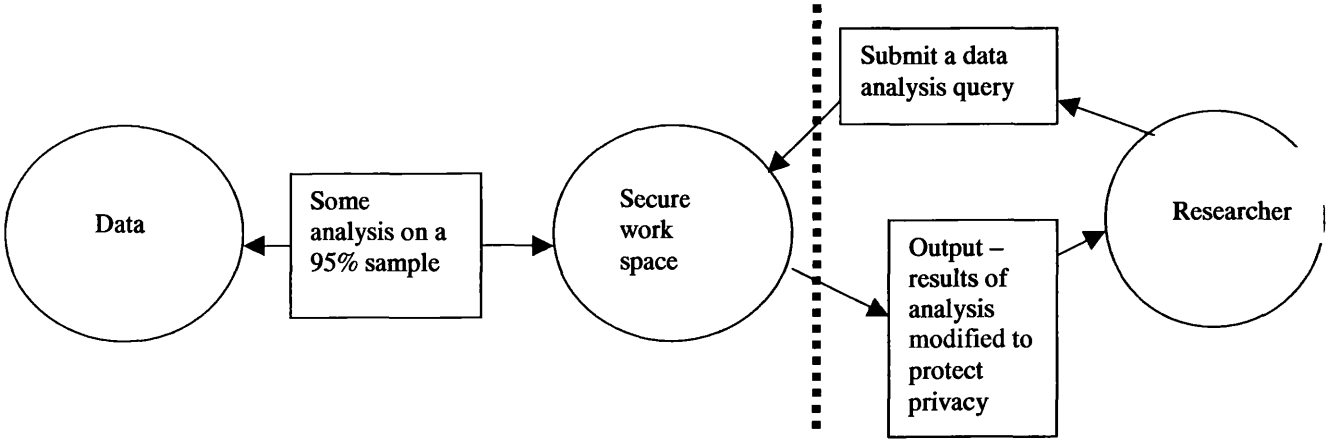
Data → Confidentialise → Analyse

However this technique (called privacy preserving analytics) follows the form:

Data → Analyse → Confidentialise

This would mean only the statistical organisation providing the data would have access to the raw data needed to run the analysis. The major advantage of this is that the analysis is run on the raw data so there is no inaccuracy in the results released to the public due to the fact that the data used has been adjusted to avoid disclosure. In general the user should be able to ask for any statistical analysis of the data and it should be provided in the full form although some analysis may yield results that cause potential disclosure and these results are in some way suppressed. One approach to do this is the remote server approach (*figure 1.5.1*).

Figure 1.5.1: The Remote Server Approach



Under this approach a researcher can request an analysis of certain data to be produced. This analysis will be carried out by the data provider in a secure workspace which will be subject to security procedures. Furthermore the analysis is run on a 95% sample of the data providing extra protection to the individual's data. Once the analysis has been carried out any results which could lead to direct identification of an individual subject are masked or suppressed so as not to reveal the individual. Although there is still the chance of some of the results being suppressed the results will be more accurate since the analysis has been run on the original data.

This technique has been investigated by CSIRO in Sydney and work on this is continuing. At present they have found that certain results cannot be released and must be replaced by some less disclosive results. For example scatterplots cannot be released so they have been replaced by parallel box plots, exact p-values cannot be revealed so a range of possible p-values is given and outliers cannot be revealed but the analysis can be carried out both with or without outliers. CSIRO have so far found that there is a place for both traditional disclosure control techniques and the remote server approach depending on the situation.

As mentioned above if the data disseminated is a sample taken from the raw data the protection given to each individual subject in the data increases. This is because any potential intruder who knows that a subject has been part of the study does not know for sure whether the data for that subject are included in the disseminated data. Clearly the smaller the sample the less likely each individual subject is to be included therefore the greater the protection. However the smaller the sample the less accurate the data released will be. This is the same trade off realised in all of the disclosure control techniques namely data utility versus disclosure risk. The major advantage of a sample is that regardless of how small the sample there is still doubt created about the presence of each individual subject. Using only a sample of the subjects can be easily used with other statistical disclosure techniques to increase the protection offered to each individual subject.

1.6 Techniques Used by Other Statistical Organisations

Statistical disclosure control is a growing topic around the world. The growth of technology worldwide has made storage of large amounts of data possible in any country in the world whilst the development of the World Wide Web and the grid has allowed for data to be transferred with ease between countries. Many different statistical organisations in many countries in the world have to deal with the problem of disclosure control. It has already been mentioned that the European Union has introduced legislation regarding the use of data within its organisation and that the CSIRO in Australia are working to develop a remote server approach. This section

will outline some of the laws and methodology used by different statistical organisations.

Statistics Netherlands is one of the forerunners in the statistical disclosure control field. The organisation has played a large part in the development of the Tau-Argus software and is in fact the consortium leader of the CASC programme. The chair of the CASC steering committee is Anco Hundepool from Statistics Netherlands. As one would expect Statistics Netherlands use Tau-Argus to protect the data it releases from potential disclosure. The most commonly used techniques are a dominance rule (see section 2.1) to detect potential disclosure and cell suppression with optimal secondary suppression (see section 2.2) to remove the potential disclosure from the table. There is also the opportunity for bona fide researchers to work on-site at Statistics Netherlands to allow the researcher access to data with more detail than would be released to the public. The researcher is allowed to use any standard software package or even their own software package. Nordholt [19] writes 'Like all employees of Statistics Netherlands, these people who work on-site have to swear an oath to the effect that they will not disclose the individual information of respondents.

The researchers who work on-site on economic data have to take the rules of Statistics Netherlands' Centre for Research of Economic Microdata (CEREM) into account. The most important rules are:

- Researchers must be associated with a recognised research institute (e.g. a university);
- There must be a research proposal that conforms to current scientific standards;

- The researcher and his superior have to sign a confidentiality warrant;
- The researcher obtains only access to the data needed for his project;
- The data do not contain information on names and addresses of the enterprises;
- Data related to the two most recent years will not be supplied;
- It is forbidden to let data, or not safeguarded intermediate results, leave the premises of Statistics Netherlands;
- All prospective publications will be screened with respect to risk of disclosure;
- All publications will be in the public domain;
- A public register contains the researcher's name(s), the research project, the publication(s) and the databases provided.'

These laws allow the researcher to work on much more detailed data however information can only be taken from the premises with the permission of the responsible statistical officer.

The Australian Bureau of Statistics produces a wide variety of statistics covering all aspects of life in Australia. As with all respectable statistical institutions they have a policy on disclosure control. The principles of the policy are consistent for all subject areas although it may be the case that particular methodology changes for the different subject areas. The Census and Statistics Act 1905 [20] requires that data taken in survey are protected against disclosure and that the act 'provides a fine of up to $5,000 and/or a penalty of 2 years imprisonment for an unauthorised disclosure of information collected under the Act.' Also the Statistics Determination 1983 – List of Regulations [21] states that.

'Information, being information included in a class of information to which this clause applies, may, with the approval in writing of the statistician, be disclosed except where:

(a) in the case of information relating to a person, being an individual – that person;
(b) in the case of information relating to an official body – the responsible Minister in relation to that official body; or
(c) in the case of information relating to an organization other than an official body – a responsible officer of that organisation,

has shown that such disclosure would be likely to enable the identification of that particular person or organization.'

It is clear that the Australian Bureau of Statistics have in place rules and regulations to ensure that they do not disclose data pertaining to individuals. As mentioned the separate departments will have individual methodologies for dealing with the problem. In the field of Health Data there are blanket rules which are used to protect tables from potential disclosure. These are:

- Any cell which contains 3 respondents or less is suppressed.
- Data are always weighted.
- Weighted data are always rounded to the nearest 100.

Clients are also discouraged from looking for tables containing too much detail, as these will often have large standard errors making the data unreliable.

The U.S. Census Bureau provides data to the public on the United States people and economy. An important aspect of the work they do is to release data whilst honouring privacy and protecting confidentiality. The Census Bureau use statistical methods to ensure the identity of individual subjects or businesses is not disclosed. The U.S. Census Bureau Website [22] states 'The Census Bureau has an internal Disclosure Review Board. This board sets the confidentiality rules for all data product releases. A checklist approach is used to ensure that potential risks to the confidentiality of the data are considered and addressed before any data are released.' The Census Bureau also has a set of 'Privacy Principles' [23] which help the organisation design surveys whilst considering the respondent's rights. These are:

- Necessity: Do we need to ask this question? Do we need to collect this information?
- Informed Consent: Do you know why we are collecting your information?
- Respectful Treatment of Respondents: Are our efforts reasonable and did we treat you with respect?
- Confidentiality: How do we protect your information?

Any violation of the confidentiality of the respondent is seen as a federal offence and can result in a fine of up to $250,000 and/or a five-year prison sentence. Furthermore the Census Bureau privacy principles [23] state that 'we promise that we will use every technology, statistical methodology, and physical security procedure at our

disposal to protect your information.' The Census Bureau frequently use cell suppression as a method to remove potential disclosure from a table and this method was used in 1980 Census of Population and Housing and is now used for economic surveys and censuses. Before the Census Bureau releases any data, a computer program is used to check published tables for primary and complementary disclosures. The Census in 2000 used a form of rounding to protect some tables from the threat of disclosure. Zayatz et al [24] write of the census 'All cell values are rounded according to the following scheme:

0 rounds to 0

1-7 rounds 4

8 or greater rounds to the nearest multiple of 5.

Totals are constructed before rounding, thus universes remain the same from table to table, but the tables are no longer additive.'

Other statistical organisations are currently working on specific disclosure control methodologies and policies. For example Statistics New Zealand is looking to implement a method of adding noise to the cells in the table. Camden et al [25] writes 'There appears to be good potential for extending the method (adding noise) to tables of counts, and we hope this will be explored further.' Also Statistics Norway is currently working on a policy/best practice document to deal with statistical disclosure control to be implemented autumn 2005.

Much of the statistical data released in the UK are released under the National Statistics Code of Practice: Protocol on Release Practices [26]. This is a document published in 2002 which outlines the principles which must be adhered to when data are released. Further the National Statistics Code of Practice: Statement of Principles [27] outlines the principles which must be adhered to when data is released. There are 7 main principles in the document including:

1. National Statistics will promote equality of access
2. Final Responsibility for the content, format and timing of release of National Statistics will rest with Heads of Profession (in devolved administrations, the Chief Statistician) acting in consultation with the National Statistician
7. As much detail as is reliable and practicable will be made available, subject to legal and confidentiality constraints

These are the three principles which have the largest bearing on the statistical disclosure control and public accountability area of data release whilst the other four principles are concerned with the timing of the data release. The protocol goes on to expand on the principles by stating that promoting equality of access requires the statistical data to be equally available to all at the same time and should be released in a format and at a time that will be suitable to the majority of users. This implies that data in the UK should not be restricted to certain groups of society (i.e. policy makers) and that the statistical organisation has a responsibility to ensure that the released data are of a form which is useful and informative to a wide range of potential users. The protocol also states that a fixed format should be adopted for data which are released regularly and that any significant change to this format should only

be done after consultation between the relevant statistical organisation and the end user. Furthermore the protocol states that although the initial release is targeted at a specific audience and as such is potentially focused and simplified there is often a large amount of data underlying the initial release and that these should be available to the greatest extent possible. However the protocol also states that 'in order to maintain the trust of respondents, information can only be accessible where it does not impinge on confidentiality constraints'. This statement is crucial to the field of statistical disclosure control since it makes it clear that it is the responsibility of every statistical organisation working under the National Statistics Code of Practice to release as much data as possible however this information can only be released if there is no danger to the confidentiality of individual subjects.

Statistical disclosure control is considered an issue by many organisations worldwide. There are clearly many methodologies used by these organisations. It would appear that the choice of method is often dependent on the data that the organisation are dealing with. It is apparent that all statistical organisations feel the need to have a policy on statistical disclosure control where the methods they use are outlined in simple terms.

1.7 Tau-Argus

It is often the case that statistical agencies gather data in the form of surveys or research and want to disseminate the data in tabular form. The data collected will be entered into a computer system as microdata and this microdata will be used to construct the tables to be released. As mentioned earlier there are many techniques

that can be used to protect both tabular data and microdata against potential disclosure. To perform these techniques by hand would be extremely time consuming and to produce a program or macro for every disclosure scenario would be both complex and inefficient. This has led to the production of a set of programmes under the heading of Argus. There are two programmes in operation, Mu-Argus and Tau-Argus. Mu-Argus is aimed at making sets of microdata safe for dissemination and at the time of writing is very much in the development phase whilst Tau-Argus, which has been the focus of most of the development so far, is aimed at producing safe tables from a set of microdata. The Tau-Argus programme is evolving rapidly with small improvements being added regularly although it would appear that the framework is in place for a programme that allows the user a large amount of choice whilst removing much of the complicated calculations and programming. A major aim of the programme would appear to be to give the user control over which methods are used for the disclosure control. There are many options available to the user and these options cover a large variety of methods for both detecting and removing potential disclosure in the tables. The programme allows a lot of user interaction in the decisions made in the process of disclosure control. For example the user has the option to select cells which should not be used in secondary suppression and can give weights to certain cells influencing the likelihood of their use in secondary suppression.

The Tau-Argus programme should not be seen as an attempt to replace the role of the statistician in disclosure control but as a tool to allow the user to carry out different disclosure control techniques without the problems of complex and time consuming calculations and computing work. A perceived problem with the programme however

is that it is not very user friendly and can be complicated for users without a solid grounding in the techniques used in statistical disclosure control. It is hoped that with time (i.e. in the later versions of the programme) this problem will be minimised with the workings of the programme becoming less complex.

The work on the Argus programme has been carried out by CASC (Computational Aspects of Statistical Confidentiality). CASC is a group within the EU that works on developing new and improved techniques and programmes to deal with the issue of statistical disclosure control. The groups website [28] quotes its aims as

'The research, development and implementation of new techniques for Statistical Disclosure Control is the main objective of the CASC project.
The need for more statistical information has a consequence that we should pay more attention to the confidentiality aspects. The impact on the Statistical Offices is very big if we fail to solve this problem adequately. The project will aim at the development of practical tools as well as new research to support the development of these tools. Attention will be paid to both tabular data as well as microdata. The development of these standard tools serves the harmonisation of the statistical production in Europe and will enforce the leading role of Europe in this topic.'

Mu-Argus and Tau-Argus are the main practical tools that are mentioned in the aims of the group.

The CASC project began in January 2001 and received funding under the European Plan for Research in Official Statistics (EPROS) (see EPROS European Plan for

Research in Official Statistics 22 September 1999 [29]). Giessing and Hundepool [30] write 'The project is meant to be a follow up for the SDC-project which has been carried out 1996 to 1999 in the sense that it will build further on the achievements of that project, and take over the results and products emerging from the SDC-project.' The members of the CASC project team are from the Netherlands, Italy, Spain, Germany and the United Kingdom. The British institutes that are involved in the project are the University of Plymouth, the University of Southampton, the Victoria University of Manchester and the Office for National Statistics.

As the quoted aims of the CASC website suggest one of the hopes for the Tau-Argus programme is that it will become recognised as a Europe wide tool for producing disclosure protected tables. This implies that the programme must give a large amount of user choice to allow its use in many different scenarios for different types of data. Sarah Giessing (member of the CASC steering committee) and Anco Hundepool (chair of CASC steering committee) write [30] 'the software package must be able to deal with tables of any size and complexity of structure, facilities must be offered to deal with specific problems of particular situations in a flexible, user-friendly and comfortable way.' In the same paper they also list the projects aims to achieve that goal which include:

'

1. Refine and support the integration of desirable qualities and facilities of existing software systems for tabular data protection as identified in the best-practice study into Tau-Argus.

3. Significant improvement of the cell suppression algorithms based on the linear programming as already supplied along with the package, and supply of supplementary heuristic methodology.
5. Development and integration of table-perturbation methodology.'

It is clear that the major responsibility of the CASC project is the development and dissemination of the Argus packages. It is in fact true that the CASC project is designed around the Argus software and the workings of this software are integral to the work done by the project.

The nature of Tau-Argus is that it brings together widely used disclosure control techniques and integrates them into one package. As it progresses the idea is that it should include the most up to date techniques available. To this end another important part of the CASC project is research into new techniques. Anco Hundepool [31] states that 'the three main goals of innovation in the proposed project regarding tools for tabular data protection will be:

- Firstly to develop data-structures for Tau-Argus that are able to represent the cell suppression problem for hierarchical structured and linked tables.
- Secondly, GHQUAR (a further package for disclosure control produced in Germany which is ideally suited to large tables) will be integrated into the restructured version of Tau-Argus.
- The third main task will be to speed up the linear programming methodology in Argus as emerged from the 4th Framework project. This will make Tau-Argus capable of solving the larger problems that result from the

representation of real life tables with many sub-marginals in reasonable (computing) time.'

All of the methods that are research are tested on a given set of test data held by the CASC project to assess their effectiveness.

1.8 ISD Scenario

ISD is the abbreviation for the Information and Services Division of NHS National Services Scotland (formerly the Common Services Agency or CSA). It is the lead organisation for healthcare statistical service to NHS Scotland and the Scottish Executive Health Department. The official ISD website [32] states that they 'provide a wide range of information support including national data collection, analysis and publication of statistics, and the provision of expert advice on information matters.' The official ISD website [32] states the aim of the operation to 'be an essential support service to NHSScotland and the Scottish Executive Health Department; responsive to the needs of NHSScotland as the delivery of healthcare evolves; proactive in determining and advising how best to use Information and Information Technology to ensure efficient, effective delivery of patient care.'

Since January 2001 the statistics released by ISD have been covered by National Statistics. National Statistics was introduced in June 2000 which changed the way official statistics in the UK were governed and how accountability was assigned. ISD now adheres to the National Statistics Code of Practice Protocol on Release Practices as it is implemented in Scotland via the Scottish Executive's Statement of

Compliance. A key requirement of National Statistics is that statistical releases should be planned and the arrangements for the release should be made available. ISD complies with a policy of releasing statistics at the earliest possible date although in certain cases the releases are made in orderly manner. In certain cases this results in the release of data being delayed although the disadvantages of this delay are outweighed by the advantages of the increased understanding given by the ordered release of the data. The work of ISD also requires the organisation to comply with the Freedom of Information Act 2002. The act gives new powers to members of the public who wish to access information held by public bodies. The nature of the act means that public bodies (such as ISD) must have procedures ready to deal quickly with any such request.

The main aim therefore of ISD is to collect data on many aspects of the health service and to provide useful information from these data to inform policy making or let tax payers know what is being achieved in the NHS. Clearly these data are important for public accountability and decision making in Scotland since the trends and variations in the ISD data may effect where and on what resources are allocated. These data can come from many fields for example, from an ISD Scotland leaflet [33], ; 'information on hospital admissions, patients with cancer, waiting lists, childhood immunisations and earnings paid to NHS Scotland staff.' ISD works in partnership with most branches of the health service e.g. Health Boards, G.P.s, hospitals, local authorities and voluntary organisations to collect the necessary data. The nature of much of these data is that they can identify individual patient's personal health data which is sensitive and therefore ISD must ensure that steps are taken to protect these data.

As a standard, ISD have security guidelines on the holding of personal data and how they can be used. Authorisation is required for any employee to have access to personal health information and this authorisation and use of data is monitored closely while health boards and hospitals are only granted access to information on their own residents/patients. ISD also have specific posts to ensure the protection of patient's confidentiality. They have an individual, known as a 'Caldicott Guardian' with the sole job of ensuring that individual's personal data are being handled properly and also a Privacy Advisory Committee which comprises of mainly non-NHS members which advises ISD on the use of personal data for research. Furthermore it is a duty of all ISD staff (and honorary contributors such as students) to protect patient confidentiality. ISD also follow the laws given by the Data Protection Act 1998 which gives an individual:

- The right to know how ISD use your personal health information.
- The right to object to our use of your information. You can ask ISD to change or restrict the way we use your information.
- The right to access any personal information that we ISD may hold on you.

1.9 Aims of the Thesis

The major aim of this thesis is to investigate the effects of existing statistical disclosure control techniques on tabular data with specific mention being made to the issues surrounding statistical disclosure control in the ISD scenario. The practicalities and methodology of various techniques will be discussed with further discussion on the advantages and disadvantages of the techniques in certain scenarios. The

methodology will be described in detail so as to give the reader an understanding of the underlying theory behind the techniques however when using the techniques in practice knowledge of the underlying theory is not essential. This is due to the fact that many of the techniques are available in statistical computing packages designed to deal with potential disclosure.

Many of the social issues related to statistical disclosure control have already been discussed earlier in chapter 1 (in particular in section 1.3) and the effects of these social issues on the disclosure control process will be discussed throughout the thesis. Many of these issues will have an effect on which disclosure control technique will be favoured by a statistician and may also affect the parameters used in the techniques.

The effect of various disclosure control techniques will be investigated by applying the techniques to tables produced from real data. Where possible the techniques will be carried out using the Tau-Argus programme. The resulting tables will then be analysed using a measure of information loss between the actual table and the disclosure controlled table. The aim is to compare the information loss incurred when using the various disclosure control techniques and comparing these techniques to the tables which would be produced using current ISD policy. This should hopefully provide recommendations or at least general advise for ISD should they choose to update their statistical disclosure control policy.

2. Disclosure Control Techniques for Tabular Data

2.1 Theory and Practicalities of Detecting Potential Disclosure in Tabular Data

It is extremely important that a proper check for potential disclosure in the data is carried out before data are published. This section attempts to investigate advanced methods of disclosure detection. It was stressed in section 1.4 that it is important to distinguish between tables containing magnitude data and tables containing frequency data. This was partly due to the fact that when it comes to detecting potential disclosure in tables there are different techniques for the two separate forms of data. This section will show techniques for detecting potential disclosure for both forms. It is of interest to note that the technique used for detecting potential disclosure in the data should not be published with the data as this can give an intruder information which can be used to estimate the original cell value.

Statistical disclosure does not only occur when an intruder can explicitly uncover a data value for an individual but disclosure is also said to occur when an intruder can estimate the value to a certain range. An interval in which the true cell value lies can usually be derived, using the other values and a linear computing programme, for each cell in the table. This interval is called a feasibility interval. For a cell to be regarded as 'safe' the feasibility interval should be acceptably large. The statistician should specify this feasibility interval in advance ensuring that it is large enough to sufficiently protect the individual subjects.

2.1.1 Frequency Data

As mentioned in section 1.4 a table containing frequency data (example *table 1.4.2*) is a table that simply quotes the number of subjects in a category in each cell. Disclosure in this form of table can be quite difficult to detect. For example it seems reasonable to assume that there is a high risk of disclosure if there is a cell in the table that contains only one subject but this is not always the case. If we think of statistical disclosure as occurring when the release of the data enables a third party to learn more about the subject than already known the following simple example illustrates why. Imagine a small survey carried out looking into unemployment figures by sex and area released the following results (*Table 2.1.1 and 2.1.2*)

Area: A		**Sex**		
		Male	**Female**	**All**
Unemployed	**Yes**	4	1	5
	No	14	17	31
	All	18	18	36

Table 2.1.1: Fabricated Example of Employment Figures by Sex for Area A.

Area: B		**Sex**		
		Male	**Female**	**All**
Unemployed	**Yes**	6	2	8
	No	20	19	39
	All	26	21	47

Table 2.1.2: Fabricated Example of Employment Figures by Sex for Area B.

Ignoring the protection offered by the fact that the study is a sample, at first sight this table would appear to disclose information about the unemployed female in area A but this is not the case. Assuming that there are figures from other areas and this was not the only area studied there is no way of making an inference on one of the variables even by knowing the other two variables i.e. assuming the intruder knows the subject is from area A and unemployed there is no way of knowing whether they are male and female, similarly knowing that the subject is female and unemployed is not enough to say they are from area A.

This example shows that tables that sometimes can appear disclosive actually do not contain a large disclosure threat so formal techniques should be used to check for potential disclosure. This section will in the main focus on explaining the workings of three rules to detect potential statistical disclosure. These rules will be the so-called *across cell test*, a test measuring the percentage of subjects in sensitive cells (*p% sensitive cell rule* say) and the *minimum frequency rule.*

In this section a 2-way table of frequency data (*Table 2.1.*3) will be used to illustrate how the techniques described to tackle the problem in frequency data work in practice.

		Variable 2				
		E	**F**	**G**	**H**	**Total**
Variable 1	**A**	23	3	37	18	81
	B	1	15	12	119	147
	C	54	43	8	4	109
	D	19	16	22	10	67
	Total	97	77	79	151	404

Table 2.1.3: Fabricated Example: Frequency Data Table.

2.1.1.1 Across Cell Test

The across cell test is suggested as an efficient method for detecting disclosure in tabular data by the Government Statistical Service Methodology Series No.4: Report of the Task Force on Disclosure (December 1995) [2]. The report gives the definition of the rule as:

The 'across cells' test for non-disclosure in tables of counts or values

'A sufficient condition for non-disclosure in a table is that each non-empty one-dimensional subset of the table is distributed in such a way that every combination of (k-c) individuals, where k is the number in the line and c the maximum size of a co-operation group, satisfies the non-proximity conditions for that dimensional classification.'

The user must define the proximity conditions before carrying out the test. The proximity condition is the percentage (say p%) to which the true value can be estimated (+/- p%) which constitutes a disclosure risk. This condition must be specified for each case individually although it is believed that p=40 is a relatively safe whilst relatively unrestrictive proximity condition. Also in practice it is usually assumed that c=1 i.e. each subject in the cell does not work with any other subject to attempt to uncover information about the remaining subjects in the study. The rule is much simpler to use on data where the variable groups are defined by ranges, e.g. age groups of 20-30, 30-40, 40-50 etc, rather than on data where the variable groups are separate entities, e.g. different strains of a disease.

The test can be simplified by grouping the variables in such a way to reduce the proximity assumptions to a sufficient equivalent involving the number of empty cells. The definition of the simplified test (from [2]) is:

Simplified 'across cells' test for non-disclosure in tables of counts or values

'If the values of the dimension variables are grouped such that two non-zero cells in each line exceeds the minimum range defined by the proximity assumptions then a sufficient condition for non-disclosure is that:

every non-empty one-dimensional subset of the table must have

either three contiguous non-empty cells

or two contiguous non-empty cells, both greater than 1'

The simplified test cannot be used in most scenarios. The reason for having the simplified test is that it can be carried out manually for small two-dimensional tables, large tables or tables of higher dimensions however require automated checking.

This simplified version of the test is one of the major advantages of using the across cell test when compared to other rules. A further advantage of the across cell test is that it allows the user to take into account the fact that contributors in the cell may not only want to make inferences on other contributors in the cell but may work together with other cell members to do this. This is particularly prevalent in business statistics where coalitions are relatively likely to be formed between two similar businesses although it may also be applicable in other fields.

However the across cell test does have disadvantages. The major disadvantage of the rule is that it requires a large amount of computing power since all of the feasibility intervals must be calculated for each subject in each cell. For small tables this is not a problem however if the test is run on large, potentially multi dimensional, tables this becomes an important issue. Another disadvantage is that the user must predict the size of a potential coalition. This may cause a problem if the table is being protected by a statistician that is not up to date on the issues surrounding the data. Clearly the statistician could go to an expert in the area for advice but this is time consuming and inconvenient.

2.1.1.2 p% Sensitive Cell Rule

This method is outlined by Willenborg and de Waal [34].

The basic concept of this rule is that there is a higher disclosure risk when the number of subjects in one cell is close to the number of subjects in the margin. For example if the row contains 5 subjects and one cell contains 4 subjects then there is clearly a higher risk of disclosure. This risk of disclosure is especially prevalent in studies where subjects have some prior information on the other subjects, for example a business study where certain information may be public knowledge or may be shared among co-operative businesses. This scenario where one cell dominates the row may not in fact constitute disclosure but the statistician should be aware of the situations where it may cause a risk.

For this rule the variables are split into two groups, identifying variables and unknown variables. Identifying variables are variables which are common/public knowledge and are used to identify someone. Unknown variables are variables that are private to the individual and are not part of public knowledge. In *table 2.1.1* sex and area would be thought of as identifying variables since they are generally regarded as public knowledge from birth records (for sex if necessary) or the electoral roll (for area). Employment status would often be regarded as an unknown variable since this is usually sensitive information. It can be argued that employment status is public knowledge through benefit records etc., but a subject's employment status is far less likely to be public knowledge than their sex or area of residence and is probably sensitive enough to be thought of as an unknown variable. This distinction is an

important one since it can be assumed that an intruder has full knowledge of a subject's identifying variables making the risk of disclosure of the unknown variables of interest.

It is not always obvious which of the variables are identifying variables and which are unknown variables. There are also infinite combinations of identifying and unknown variables that can be used. Here the 'worst case' scenario is considered where all variables are identifying variables ($Z_1, Z_2, \ldots, Z_k$ say), which can be treated as one categorical variable (Z say), except one unknown variable (Y say). The disclosure risk is highest when the distribution of Y is highly concentrated for any value of Z. Clearly the highest concentration (and therefore the highest disclosure risk) occurs when the value of Y is the same for any value of Z. This theory can be applied to tables where if one cell in the row corresponding to variable Y contains all, or the majority, of the subjects this constitutes a disclosure risk. For example consider *table 2.1.3* and suppose that variable 1 is a variable which is unknown to the general public.

		Variable 2				
		E	**F**	**G**	**H**	**Total**
Variable 1	**A**	23	3	37	18	81
	B	**1**	**15**	**12**	**119**	**147**
	C	54	43	8	4	109
	D	19	16	22	10	67
	Total	97	77	79	151	404

Table 2.1.4: p% sensitive test applied to Table 2.1.3 – unsafe row is highlighted in bold.

Using the p% sensitive cell rule it may be that category B of variable 1 poses a potential disclosure threat due to the fact that 119 of the 147 subjects in the category are present in one cell. This is in fact 80.95% of the subjects in this category in the one cell. This appears a large amount however whether this is too large is a subjective decision here. In practice it is possible to adopt a simple rule such as; a row is sensitive if at least p% of the subjects fall into a sensitive category of Y. P must be specified in advance according to what the statistician perceives as a reasonable value.

Table 2.1.4 shows that one of the major advantages of the p% sensitive cell rule is that it is a simple rule to implement. Very little computing power or indeed time will be needed to investigate the rule. The rule also takes into account the number of subjects in each row and uses the percentage of subjects in each cell as the test value rather than simply imposing an arbitrary minimum value required in each cell.

However there are certain disadvantages of the rule. It may be the case that although most of the subjects fall into one cell due to the make up of the table there is no disclosure risk due to a similar scenario as shown in *table 2.1.1* and *table 2.1.2*. Another disadvantage of the rule is that if a cell fails the test then the whole row/column has to be suppressed which may result in an unnecessarily large amount of information loss. Furthermore the statistician must make a decision on which variables are deemed sensitive. This is not always straightforward, as different people will have different views on whether certain information is sensitive.

2.1.1.3 Minimum Frequency Rule

The minimum frequency rule states that a cell is considered unsafe if it contains less than n subjects. The number of subjects required to make the cell safe must be specified by the statistician in advance. The idea is that the more subjects a cell contains, the lower the risk of an intruder identifying one subject in the cell and if enough subjects are in the cell this risk becomes acceptably low.

This rule is very simple but can in certain cases be very efficient. It is interesting to note that the minimum frequency rule is also often used for magnitude data. In the Tau-Argus programme the Minimum Frequency Rule can be used if the cell items are either a response variable or just simply the frequency. The rule simply states that a cell is unsafe if the number of contributors is less than a specified frequency (n say). The user must select a value for n. Popular choices for n include 3 (the default in Tau-Argus) and 5. When the Minimum Frequency Rule is chosen the user must specify a Range. This is to allow secondary suppressions to be carried out to give the required protection for each suppression. This range is given as a percentage and it means that a cell is only safe if it is suppressed and an intruder cannot predict the true value, using the other cell values and the marginals, to within the chosen percentage of the true value. The Tau-Argus user manual [35] states 'For example, if this value was set to equal 30%, it would mean an attacker would not be able to calculate an interval for this cell to within 30% of the actual value when looking at the safe output. Following this, the secondary suppressions may be carried out.'

If the minimum frequency rule was carried out on *table 2.1.3* and n was chosen to be 5 there would be three unsafe cells.

		Variable 2				
		E	**F**	**G**	**H**	**Total**
Variable 1	**A**	23	**3**	37	18	81
	B	**1**	15	12	119	147
	C	54	43	8	**4**	109
	D	19	16	22	10	67
	Total	97	77	79	151	404

Table 2.1.5: Minimum frequency test applied to Table 2.1.3 – unsafe cells are highlighted in bold.

This rule for detecting potential disclosure would have shown a risk in this case in three of the cells (in bold) and *table 2.1.5* would have to be adjusted in some way before it was released. The simplicity with which this is done is one of the major advantages of the minimum frequency rule. It can easily be seen if a cell entry is lower than the required minimum frequency. It also allows the user to make their own decision on what number of subjects are required in a cell for it to be considered safe. The minimum frequency rule can also be used on both frequency and magnitude data.

On the other hand a major disadvantage of the minimum frequency rule is that it is over simple. There are many occasions where a cell with (say) 1 subject may cause no disclosure risk and under this rule the cell would be considered unsafe. Therefore

this rule may result in a number of cells being unnecessarily suppressed resulting in an unnecessary loss of information.

2.1.2 Magnitude Data

In most cases the rules used to check for potential disclosure are different when magnitude data is used. As mentioned in section 1.4 a table containing magnitude data (example *table 1.4.1*) is a table that quotes a quantitative value in each cell of a table, for example median income or mean blood pressure. It should be clear that the minimum number of subjects in a non-zero cell must be three since if there was only one subject in the cell the value of the variable would be disclosed to the whole public and if there are two subjects in the cell the value of the variable for one subject would be disclosed to the other subject in the cell. Since the values in the cells are quantitative it is often of interest to know the range to which an intruder can estimate the true value. Disclosure can be said to occur if the intruder can estimate the true value to a small range not necessarily the exact value. Section 2.1.2 will in the main focus on explaining the workings of two rules for detecting the risk of statistical disclosure. These rules will be the so-called *across cell and within cell tests* and the *dominance rule.*

2.1.2.1 Across Cell and Within Cell (P,Q% Test) Tests

The across and within cell test is an extension of the rule for detecting potential disclosure for frequency data given in the Government Statistical Service Methodology Series No.4: Report of the Task Force on Disclosure [2]. The extension

comes from the within cell part of the test which is required when dealing with values. Firstly an across cell test must be performed on the data. This is done in exactly the same way as the across cell test in section 2.1.1. Once this test has been carried out a further test, the within cell test, must be carried out on the data.

The idea of the within cell test is to determine whether the range of values for which the value of the variable for one subject in the cell can be estimated by another subject in that cell is small enough to constitute a disclosure of information. Protecting against this sort of disclosure is often much more complex than protecting solely against an outside intruder although in many cases it is more likely that an intruder will be another subject in the study. An example of the desire of an intruder to disclose information on a subject from within the same cell would be in a business survey investigating turnover of businesses in a certain field. One of the businesses in the cell could be competing for a contract with another business in the same cell and having information on their turnover could be beneficial. Therefore in this circumstance the confidentiality of all the businesses should be safeguarded. Clearly in this type of scenario the businesses may have auxiliary information on each other making the task of disclosure control even more complicated.

To carry out the test a proximity condition must be specified in advance. As for the across cell test the proximity condition is the percentage (say p%) to which the true value can be estimated (+/- p%) which constitutes a disclosure risk. The assumption is that a subject in a cell knows his own value precisely and in spite of this must not be able to estimate the value of another contributor to within p%. It is also prudent to assume some prior knowledge of the other subjects in the cell (i.e. the subject can

estimate the value given by the other subjects to within q%). This would be especially true in an example such as the business example above. With these assumptions in mind the within cell test proceeds as follows (as printed in the Government Statistical Service Methodology Series No.4: Report of the Task Force on Disclosure [2]):

'Suppose $x_1, x_2, x_3, \ldots x_n$ are the values summed in a table cell and that contributor x_2 is attempting to disclose the value contributed by x_1. Assuming he can estimate $x_3, \ldots x_n$ to within q percent, then his estimate for x_1 is the range

$$cell \quad total - x_2 \pm \sum_{i=3}^{n} x_i \times \left(1 - \frac{q}{100}\right) \tag{1}$$

This simplifies to

$$x_1 \pm \left(\frac{q}{100} \times \sum_{3}^{n} x_i\right) \tag{2}$$

So the condition for this to be wider than the range $x_1 \times \left(1 \pm \frac{p}{100}\right)$ is

$$x_1 < \frac{q}{p} \times \sum_{i=3}^{n} x_i \tag{3}'$$

In theory this test should be carried out for each subject attempting to find out information on every other subject but it is clear from the formula that the most disclosive situation is when the subject that makes the second largest contribution is estimating the value for the subject that makes the largest contribution. Therefore if the data does not cause a disclosure risk in this scenario it is safe to assume that there is no disclosure risk from that cell. Unfortunately this method can be quite restrictive especially when there is a high prior knowledge (q%) assumed.

A further precaution must be taken if it is thought some of the subjects are cooperating to estimate information about a subject. If it is assumed that there will be c co-operators then the non-disclosure becomes:

$$x_1 < \frac{q}{p} \times \sum_{i=c+2}^{n} x_i \qquad (4)$$

The within cell test is often referred to as the p,q% rule. A special case of the within cell test is the *p% rule* where it appears that potential intruders are assumed to have no prior knowledge of the subjects in the table. In fact Loeve [36] writes about the rule 'it is a priori known that all contributions are non-negative. In fact, the *p% rule* is a special case of the *p-q rule* with an upper and a lower q parameter. The a priori information about every contribution is that it lies between zero and infinity.' Basically the rule states that any cell in which after publication the true value of any respondent can be estimated to within p% constitutes a potential disclosure risk. This rule has been used and was in fact used for the 1992 US Economic Census but Eric Schulte Nordholt [37] writes 'not many countries have already experience in using other rules than the dominance rule for the identification of sensitive cells in tables.

When the p-percent rule and the pq rule will become available in standard software packages for statistical disclosure control it can thus be expected that these rules will become more popular.' The p% rule is available in the Tau-Argus programme. The rule states that a cell is unsafe if the value of the largest contributor to the cell can be determined to an accuracy of p% by a coalition size N from within the group. It is often the case that N is taken to be 1 i.e. it is not often expected that coalitions are formed between subjects in the cell. The user selects the values of p and N required for the rule they wish to implement. A popular choice (the default choice) is p=10 and N=1 i.e. the second largest subject (worse case scenario with no coalition) cannot determine the value of the largest subject to an accuracy of better than 10%.

Clearly the rule has both the advantages and the disadvantages from the across cell test in spite of the fact the within cell test is also used. The within cell test however has further advantages and disadvantages. One of the advantages of the within cell test is that thought is given to the fact that members in the same cell may wish to disclose information on each other and furthermore takes into account that these cell members may work together to achieve this. The test also takes into account the fact that the value of the second largest contributor to a cell influences the degree of protection for the largest contributor to a cell. This is often important but is overlooked by many rules. Furthermore the rule takes into account the fact that the intruder(s) may have prior knowledge on other members in the cell other than the one they wish to disclose information on.

As with the across cell test a major disadvantage of the within cell test is that a large amount of computing power may be needed since proximity ranges for each cell must

be calculated. A disadvantage also found with the across cell test is that the user must predict the size of a potential coalition and furthermore must predict the potential prior knowledge the intruder has of subjects in the cell. This may cause a problem if the table is being protected by a statistician with no expertise in the area of the data. Clearly the statistician could consult an expert for advice however this will cause the process to potentially lengthen in time scale and cost.

2.1.2.2 Dominance Rule

The dominance rule, or (n,k)-test as it is sometimes known, is a common and relatively simple rule used to detect potential disclosure in tables of magnitude data. The idea behind the rule is to find cells in which few subjects account for a large proportion of the cell total. Westlake [38] writes 'This rule is widely used, but it measures Dominance, not Sensitivity and is thus disliked by theoreticians.' There can clearly be disclosure issues if one or two subjects contribute most of the data in a cell. For example if there was a survey of the number of brain operations carried out in British hospitals in a year and in one hospital there were found to be 400 and one of five surgeons knew they had carried out 300 operations this may cause internal problems in the hospital. There are many situations where this problem could arise so it is prudent to check for this disclosure before releasing any tables.

To carry out the test the parameters n and k must be specified. The parameter n is the size of a (if any) potential coalition. This is usually taken as one, since it is assumed that no coalition will be formed, although occasionally it will be taken as two. The parameter k is the threshold value. In general the rule says that a cell is potentially

disclosive if the largest n values account for more than k% of the cell total. Formally (with notation from Willenborg/de Waal [34]) this can be written as:

'Suppose $x_1, x_2, x_3, \ldots x_n$ are the values summed in a table cell and that these values are ordered with x_1 the largest i.e. $x_1 \geq x_2 \geq \ldots \geq x_n$. Also denote the sensitivity measure (i.e. the disclosure risk of the cell) as $S(X)$. Then for the nth cell

$$S_n(X) = \sum_{i=1}^{n} x_i \Big/ \sum_{i=1}^{N(X)} x_i \qquad (5)$$

The cell does not have a potential disclosure risk if $S_n(X) \leq k$.'

The dominance rule is the most widely used rule for detecting disclosure although it does have some theoretical problems. Cuppen [39] writes 'Although the (n,k)-dominance is quite standard, the protection level it offers at the respondent level is not directly clear. This is because the (n,k)-dominance rule does not entirely account for the internal structure of the cell. It compares the relative size of the sum of the n largest contributors to the size of the cell total, but it does not account for the relative size of largest contributor versus the other n − 1 largest contributors or versus the remaining contributors.' In the Tau-Argus programme the Dominance Rule is used if the cell items are a response variable. The rule states that a cell is unsafe if n subjects in the cell contribute more than k% of the total value of the cell. The user must select the values of n and k they wish to use. A popular choice (in fact the default choice) is n=3 and k=75 although any sensible option is acceptable since other factors may affect the choice of these parameters.

An advantage of the Dominance Rule, which may have been considered by the programmers of Tau-Argus, is that it is relatively simple. It does not (especially for a small table) take a lot of computing power to determine whether a cell fails the test. It also offers good protection to the large contributors to the cell since if the large contributors contribute too much of the total of the cell it is suppressed.

However as mentioned earlier the dominance rule is disliked by theoreticians since it measures dominance and not sensitivity. This is clearly a disadvantage since the user wants to protect any cell which is regarded as sensitive in terms of risk of potential disclosure and a test of dominance (which is simply a measure of the size of a value compared to the other values in the cell) is not as helpful as a test of sensitivity. Another disadvantage of the dominance rule is that it does not take into account the fact that the value of the second largest contributor to a cell influences the degree of protection for the largest contributor to a cell.

2.1.3 Summary of Techniques Used to Detect Potential Disclosure in Tabular Data

The techniques which have been described to detect potential disclosure are summarised in *table 2.1.6.*

Technique	Data Used On	Parameters
Dominance Rule	Magnitude	n – number of contributors k – percentage contributed
P% Rule (across and within cell test)	Magnitude	p – accuracy of estimation n – size of coalition
P,Q% Rule	Magnitude	p – accuracy of estimation q – accuracy of intruder prior knowledge of other subjects n – size of coalition
Minimum Frequency	Magnitude and Frequency	n – minimum number of contributors
P% Sensitive Cell Rule	Frequency	p – percentage of subjects in sensitive category
Across Cell Test	Frequency	p – accuracy of estimation c – size of coalition

Table 2.1.6: Summary of Potential Disclosure Detection Techniques. Techniques available in Tau-Argus are shown in italics.

2.2 Theory and Practicalities of Removing Potential Disclosure in Tabular Data

Should the tests for potential disclosure indicate a risk of disclosure the table can clearly not be printed in its current form. This situation gives a statistician a problem. The idea of the study carried out will have been to produce results that can be viewed by some public body and usually to simply not publish any results due to potential disclosure is unacceptable. This necessitates the need to have techniques to in some way modify the table so that it can be viewed by the public but not result in potential disclosure of information on an individual subject. Some of the steps outlined in section 1.5 to protect against disclosure, such as running the analysis on a sample of

he data rather than the whole data, may have already been implemented or could now e implemented but there are statistical techniques which can be used to protect the roduced table. The techniques used to remove potential disclosure in tables fall into wo categories: perturbative and non-perturbative. The idea behind perturbative nethods is to adjust the data (i.e. give it error) in such a way as to protect the identity of individuals. Non-perturbative methods are techniques to alter the table (mainly by suppressing some results) whilst not altering the data. Clearly both methods result in a loss of information in the table but are obviously preferable to no table being published due to potential disclosure. This section will attempt to cover methodology and issues raised by both categories of techniques used to remove potential disclosure in the data.

2.2.1 Perturbative Techniques

As mentioned earlier perturbative techniques attempt to change the data in such a way that the disclosure risk is decreased whilst trying to retain as much of the information as possible. The clear disadvantage of this is that the data in the output are different from the raw data collected. This can result in a potential reduction in the quality of the data. It could also lead to those without statistical knowledge to dismiss the results as a fabrication and give them no credibility. It is important therefore that the techniques are statistically sound so as to allow the statistician to have confidence that the results produced still have statistical merit. This is especially important to statisticians who have to produce results for a non-statistician e.g. a statistician presenting results to a government minister. A major advantage of this form of disclosure removal is that no cells in the table are suppressed so no values are

'hidden' from the public. This section aims to investigate some of these perturbative techniques and show the theory behind them. Two of the most common perburbative techniques are *adding noise* and *rounding*.

2.2.1.1 Adding Noise

Adding noise is a relatively simple technique that can be used to purposefully alter the data so as to mask the true values given by individual subjects. It involves adding random variation to all the cells in the table. It can be a very effective way to protect against potential disclosure since an intruder should not be able to make any inferences on the data in the table since these data have (potentially) been subject to the addition of noise. Unfortunately adding noise to the cells in the table means that additivity of the table is not guaranteed so if additivity is important this method clearly cannot be used. When adding noise to the cells it is important to ensure that no systematic bias is added to the data. Therefore the expected value of each cell should be the same as the original value after adding some noise. The technique of adding noise is widely accepted as effective and is outlined by Willenborg and de Waal [34]. The notation used in this section is the same as in the book.

There are a variety of ways in which noise can be added to the table. In all these ways it is important that a random value is added to each cell in the table, although it is sometimes possible to exclude some cells such as empty cells. It is also important that the type and amount of noise added must be consistent with the type of table e.g. if the original table contains frequency data then the table after adding noise must also contain frequency data.

Suppose there is a cell with value a then adding noise to the cell will change the value to $a' = a + \varepsilon_a$ where $E(\varepsilon_a) = 0$ and $Var(\varepsilon_a) = \sigma_a^2$. The expectation of the error is zero to ensure no systematic bias is introduced and this must hold in all cases. The variation of the error can be adjusted for different tables and circumstances. A very simple system for adding noise is to add either 1, 0 or –1 to each cell with the value to be added to each cell selected randomly. There are more complex systems to add noise. Willenborg and de Waal [35] suggest adding proportional noise to the data claiming that this is effective when the cell values are spread over a wide range. This is again a relatively simple technique where σ is proportional to a, e.g. $\sigma = a\sigma_0$ for some $\sigma_0 > 0$, and error given by this proportional variance is then added to a to give $a' = a + \sigma$.

An alternative to additive noise, that is in some cases preferable, is multiplicative noise. To introduce multiplicative noise there is the original cell value a and a random variable γ. The original cell value is multiplied with this random variable to give the the new value $a' = a\gamma_a$ where $E(\gamma_a) = 1$ and $Var(\gamma_a) = \sigma_a^2$. In this case the expectation of the variable is one (since it is now multiplicative noise and not additive noise) to ensure there is no systematic bias introduced and the variance of the variable can again be adjusted.

Using this technique of multiplicative noise it is possible to add noise to the table and protect additivity. To achieve this the marginals must be adjusted using multiplicative noise and then iterated proportional fitting (IPF) is used to spread the adjusted

marginals around the individual cells in the table. This results in a table, which is additive, that has been protected against the risk of disclosure due to the introduction of noise.

Table 2.2.1 gives an example of how *table 2.1.3* would look if additive noise was used and the simple system for adding noise of adding either 1, 0 or –1 to each cell, with the value to be added to each cell selected randomly, was also used.

		Variable 2				
		E	**F**	**G**	**H**	**Total**
Variable 1	**A**	23	2	38	17	81
	B	1	14	11	119	147
	C	54	42	8	5	110
	D	18	17	23	10	67
	Total	98	77	79	150	403

Table 2.2.1: Adding noise technique applied to Table 2.1.3.

Table 2.2.1 has now been protected against potential disclosure due to the uncertainty added to each of the cells however it is no longer additive. A major disadvantage with adding noise, as illustrated in *table 2.2.1*, is that making the table additive is not practical in many cases. This is because doing so requires fairly complex and time-consuming proportional fitting. Any table which is not additive is in general disliked, especially by those with no (or limited) statistical knowledge. This is understandable because to those unfamiliar with these techniques the table will look incorrect. Similarly it is difficult to explain to an end user the process of adding noise and how

results which have simply had numbers added to them can still give statistically sound results.

However adding noise to the table to prevent potential disclosure has certain advantages. One advantage of adding noise is that it is very simple to ensure that no systematic bias is added to the data. This is done by ensuring that the random variable which is added to the cell value has an expectation of zero. In general adding noise gives the user a large amount of choice when it comes to deciding how the data are protected. For example the user can define the noise to be simply randomly adding 1, 0 or –1 to the original value or they can make a complex statistical model to model the noise. The user also has the choice between additive or multiplicative noise and can choose any random variable from which to generate the noise. These decisions can be made by considering the size and spread of the data present in the table. Another advantage is that adding noise means that no cells have to be suppressed. The advantage of this is that the table appears complete and it doesn't appear as if some data has been 'hidden' from the public or end user.

2.2.1.2 Rounding

Another technique used to adjust the data in the table to avoid potential disclosure is to carry out rounding on the cell values in the table. This offers protection against disclosure since an intruder cannot make clear inferences on the value of the cell since they do not know the true value of the cell. Unfortunately it also results in a loss of precision in the data and it could be argued that presenting rounded data to a non-statistician is problematic since the data have been visibly adjusted.

The idea of rounding is that the original data are replaced by a multiple of a given rounding base. The choice of rounding base will usually depend on many factors such as spread of the data, the size of the data or the information contained in the table. There are many different forms of rounding and the choice of which form of rounding to use is very important. Unsophisticated and ill-informed rounding can result in the table losing its additivity. Additivity in a table is preferable wherever possible since most users find it simpler to understand and accept and also makes it much more difficult for an intruder to 'unpick' cell contents. Therefore it is important that a sensible rounding technique is used. It is also important that the rounding technique used is not revealed since a well-informed intruder using powerful software could use this information to make inferences about the data in the table. This section will look at the workings of some of the available rounding techniques and look at the protection they provide to tabular data.

As with adding noise one of the major advantages of rounding is that no cells are suppressed. In general the public or end user will prefer to have values in all the cells of the table as this appears much more complete. In general rounding is a relatively simple operation where the only choice required by the statistician is the size of the rounding base.

The major disadvantage of rounding is that a table which has been rounded gives the impression of being crudely and obviously adjusted. Even a casual observer will notice that every value has been rounded to a certain base and end users tend to feel uncomfortable with this. Also a disadvantage of both conventional and random

rounding is that additivity is not retained in the table which causes problems for the end users as outlined above in the adding noise section. A further disadvantage with rounding is that for the table to be protected every cell must be rounded. This means that even many of the cells which contain no disclosure risk will be adjusted resulting in a loss of data utility.

2.2.1.3 Conventional Rounding

Conventional rounding is the simplest rounding technique. It is rarely used in practice and is included here simply to illustrate the general idea of rounding. To carry out conventional rounding the value in each cell in the table is rounded to the nearest multiple of the rounding base. An example of conventional rounding is given below:

Suppose the rounding base is 5 and the original table is *table 2.1.3*. Using the conventional rounding technique the rounded table would become *table 2.2.2*.

		Variable 2				
		E	**F**	**G**	**H**	**Total**
Variable 1	**A**	25	5	35	20	80
	B	0	15	10	120	145
	C	55	45	10	5	110
	D	20	15	20	10	65
	Total	95	75	80	150	405

Table 2.2.2: Conventional rounding applied to Table 2.1.3.

As can be seen the table is not additive. As mentioned conventional rounding is rarely the preferred rounding technique of most statisticians due to both the loss of additivity and the fact that little protection is offered. Conventional rounding does however result in little information loss since all values are rounded to the nearest multiple.

2.2.1.4 Random Rounding

Random rounding is another form of rounding that can be used to protect the table against potential disclosure. The method behind random rounding is that each cell value that is not a multiple of the rounding base is rounded up, to the nearest multiple of the base, with probability p and rounded down, to the nearest multiple of the base, with probability $1-p$. All cells which are multiples of the rounding base stay the same and p remains constant for all the cells in the table.

One variant of random rounding is *unbiased random rounding*. The procedure is unbiased since the expectation of each cell is equal to the original value of the cell. This is achieved by making the probability of rounding up or down dependent on the value of the cell. An example of how this works with a rounding base 5 is given in *table 2.2.3*.

Original Value	Rounded Value	Probability
5	5	1
6	5	4/5
	10	1/5
7	5	3/5
	10	2/5
9	5	1/5
	10	4/5

Table 2.2.3: Random Rounding Probabilities for base 5.

It is clear that the values in the table lead to an unbiased procedure. For example the expected rounded value for a cell with original value 6 is $\left(5\times\frac{4}{5}\right)+\left(10\times\frac{1}{5}\right)=4+2=6$

i.e. the expected rounded value is the same as the original value.

Random rounding offers more protection against disclosure than conventional rounding due to the ambiguity created by the fact each cell can be rounded either up or down in a random fashion. Random rounding does however cause information loss that is both greater than conventional rounding and harder to control. An example of how random rounding may work on *table 2.1.3* is given in *table 2.2.4.*

		Variable 2				
		E	**F**	**G**	**H**	**Total**
Variable 1	**A**	25	5	40	15	80
	B	5	15	15	120	150
	C	55	40	5	5	110
	D	20	15	25	10	65
	Total	100	80	80	150	400

Table 2.2.4: Random rounding applied to Table 2.1.3.

It is clear that whilst random rounding has resulted in *table 2.2.4* being protected from potential disclosure the table is not additive which is undesirable to both the statistician and the casual observer. However the protection offered here is far greater than in the conventional rounding technique since there is doubt as to whether each cell has been 'rounded up' or 'rounded down'.

2.2.1.5 Controlled Rounding

Controlled rounding is a very powerful, widely used technique for protecting tabular data from potential disclosure. It is slightly more complex than both conventional rounding and random rounding. It is said that a table has been controlled rounded if:

1. for every value in the cell, the rounded value is the largest multiple of the base that is smaller than the original value or the rounded value is the smallest multiple of the base that is larger than the original value (i.e. rounded to an adjacent multiple of the base)

2. the rounded table is additive

The procedure used to carry out controlled rounding will be illustrated here using the notation used by Willenborg and de Waal [34].

Suppose there is a cell value x and the value of the rounding base is b. Then x can be written as

$$x = kb + r, \quad 0 \leq r < b \tag{6}$$

where r is the remainder after dividing x by b and k is the whole number given when dividing x by b. For example if $x = 7$ and $b = 5$ then k would be equal to 1 and r would be equal to 2.

A rounding variable, $\phi(r)$ say, must then be created. $\phi(r) = b$ (with probability p say) if x is rounded to the smallest multiple of the base larger than itself. $\phi(r) = 0$ (with probability $1 - p$ say) if x is rounded to the largest multiple of the base smaller than itself. Now let $[x]$ be the randomly rounded cell value. $[x]$ can be written as

$$[x] = kb + \phi(r) \tag{7}$$

The expected value of $[x]$ is then

$$E([x]) = kb + pb \tag{8}$$

The value of p can be chosen by the statistician but it is important to note that for the process to be unbiased p must be equal to $\frac{r}{b}$ giving $E([x]) = x$.

This procedure is carried out for all the cells in the table to give the completed randomly rounded table. Controlled rounding is the most effective of the three rounding procedures outlined in this section since it both preserves additivity and the randomness of the rounding offers strong protection against disclosure. Dr. J. J. Salazar-Gonzalez developed a Controlled Rounding Program for the ONS which, as Lowthian and Merola [40] write, was 'based on sophisticated optimization techniques and computes solutions by the following criteria:

a) each rounded value is a multiple of the base adjacent to the original value;

b) the rounded values, y_i, must satisfy given constraints defined as $lb \le y_i \le y_i^-$ and $ub \le y_i \le y_i^+$;

c) the rounded table, y, satisfies My=0;

d) if more than one solution satisfying a) and b) exists, the solution chosen is the one that minimizes the distance function: $\delta(a, y) = \sum_{a_i \in a} w_i |a_i - y_i|$, where the w_i's are given weights.

Where a is such vector, then the additive structure of the table can be represented by the equation:

$$Ma=0,$$

where M is a matrix of coefficients (0, 1 or –1) that describe the additive relationships among the cells of the table.'

An example of how controlled rounding would work on *table 2.1.3* with a rounding base of 5 is given in *table2.2.5.*

		Variable 2				
		E	**F**	**G**	**H**	**Total**
Variable 1	**A**	25	5	35	15	80
	B	0	15	10	120	145
	C	55	40	10	5	110
	D	20	15	25	10	70
	Total	100	75	80	150	405

Table 2.2.5: Controlled rounding applied to Table 2.1.3.

Table 2.2.5 has now been protected against potential disclosure due to the uncertainty created by the rounding of the cells. This illustrates an advantage of controlled rounding which is that the table retains additivity. Furthermore when either controlled rounding or random rounding are used the process can be designed to ensure the rounded value of the cell is unbiased. The advantages of controlled rounding in particular are that a complete, additive and unbiased table is produced.

2.2.2 Non-Perturbative Techniques

As with perturbative statistical disclosure control techniques, the aim behind non-perturbative statistical disclosure control techniques is to reduce the potential risk of an intruder being able to obtain data on an individual subject whilst retaining the

utility of the data. The major difference between perturbative and non perturbative techniques is that whilst perturbative techniques adjust the data to reduce the risk of disclosure, non perturbative techniques suppress some of the data to reduce the risk of disclosure. An advantage of this is that the tables produced are based solely on the raw data collected. This means that the statistician does not have to explain why the data can be adjusted whilst still producing sound statistical results. On the other hand a suppressed cell in a table will both arouse suspicion about that table and potentially give the table a sense of incompleteness. These and further advantages and disadvantages will be discussed throughout this section. Two of the most common non-perturbative techniques are *table redesign* and *cell suppression.*

2.2.2.1 Table Redesign

When the table that is to be protected has many sensitive cells it is often useful to redesign the table. That is variables should be grouped together so as to increase the number of subjects in each cell. There are no specific statistical rules for redesigning a table and this is done using intuitively sensible variable groupings. Table redesign requires certain categories of variables which can be combined sensibly, if this is not the case table redesign is a very difficult technique to carry out and may ultimately result in useless tables being produced. It may also be the case that table redesign does not remove all the potential disclosure in the table and further disclosure control techniques may be required to protect the redesigned table from potential dislcosure. A simple fabricated example of table redesign is found in *table 2.2.6* where the data in cells for 17-24 and 25-35 year olds is considered (after formal testing) sensitive.

Disease / Age Group	No	Yes	Total
0-16	72	12	84
17-24	75	3	78
25-35	89	1	90
Total	236	16	252

Table 2.2.6: Fabricated example of Disease Status by Age Group.

The potential disclosure in this table could possibly be avoided by combining age groups 17-24 and 25-35. This combination of groups makes intuitive sense since the new groups could now represent children (0-16) and young adults (17-35). This table redesign would give a new table (*table 2.2.7*).

Disease / Age Group	No	Yes	Total
0-16	72	12	84
17-35	164	4	168
Total	236	16	252

Table 2.2.7: Table redesign applied to Table 2.2.6.

Formal tests of potential disclosure risk would be carried out on the redesigned table and if required further disclosure control techniques could be carried out on the redesigned table. Not all tables will be as simple to redesign although in many cases groups can be combined in effective and sensible ways to reduce the necessity of the use of other disclosure control techniques (such as suppression and rounding).

The major advantage of table redesign is that none of the data are adjusted or suppressed so therefore none of the data utility is lost. Redesigning the table is generally regarded as the most efficient form of disclosure control provided the rows and/or columns fall into sensible groups. Another advantage of table redesign is that when the table is redesigned it is often the case that the table is made simpler and more user friendly. Clearly any row or column which is required by the end user cannot be collapsed into a redesigned row/column.

There is however a major disadvantage in table redesign which is that in many cases there are no natural groupings for the rows/columns to be collapsed into. This results in the method being unavailable or in a worse scenario a statistician attempting to combine rows/columns which should not intuitively be combined. Table redesign is a very effective method but should only be used in situations where there are intuitive groupings.

Table Redesign is available in Tau-Argus under the term Recoding. Recoding is often useful in protecting a table against disclosure since collapsed cells usually contain more contributors therefore reducing the risk of disclosure. There are two types of recoding; Hierarchical Recoding and Non-Hierarchical Recoding. If the variable to be recoded is a hierarchical variable Tau-Argus shows a hierarchical tree in the recode section. Consider the example, from the Tau-Argus manual, where there is data for 12 regions (1-12) in four geographical areas (Nr, Os, Ws, Zd) where regions 1,2,3 are in area Nr, regions 4,5,6,7 are in area Os, regions 8,9,10 are in area Ws and regions

11,12 are in area Zd. If this geographical hierarchical variable were to be recoded the hierarchical tree in Tau-Argus would be in the form of *figure 2.2.1.*

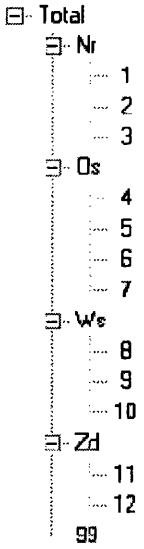


Figure 2.2.1: Hierarchical tree as would appear in Tau-Argus

To collapse the cells the user selects the '-' box next to the area. For example if the user wanted to collapse the cells for all the regions into the geographical areas the '-' boxes next to each of the four areas would be selected leaving the hierarchical tree in the form given in *figure 2.2.2.*

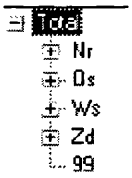


Figure 2.2.2: Hierarchical tree with collapsed cells as would appear in Tau-Argus

Alternatively if the user wanted to open the area to include cells on regions the '+' box is selected. Once the user has selected the form of recoding in which they wish the data to appear the Apply option and then the Close option is selected to incorporate these changes in the table.

If the variable to be recoded is non-hierarchical, the user is required to enter the recoding into Tau-Argus. To illustrate how the recode is entered considered an example where there are data for children aged 10, 11, 12, 13, 14, 15, 16, 17, 18. It may be sensible to recode these in to three groups; Group 1 containing 10, 11, 12 year olds, Group 2 containing 13, 14, 15 year olds and Group 3 containing 16, 17, 18 year olds. In the 'Edit box for global recode' this would be entered in the form;

1:10-12

2:13-15

3:16-18

2.2.2.2 Cell Suppression

Cell suppression is a more complex but more powerful non-perturbative technique than table redesign. The idea behind cell suppression is that any cell in the table that is sensitive is not published (in practice the cell value in the table is usually replaced by a symbol, + say). These suppressions are called primary suppressions. Usually the primary suppressions alone do not provide sufficient protection against potential disclosure. Consider that table *2.1.3* has been tested for potential disclosure using a minimum frequency test with n = 3 and cell suppression will be used to protect the table. The primary suppressions would result in the table being of the form of *table 2.2.8*.

		Variable 2				
		E	**F**	**G**	**H**	**Total**
Variable 1	**A**	23	+	37	18	81
	B	+	15	12	119	147
	C	54	43	8	4	109
	D	19	16	22	10	67
	Total	97	77	79	151	404

Table 2.2.8: Primary suppressions removed from Table 2.1.3.

It is obvious that *table 2.2.8* offers no protection to the suppressed cell. The number of subjects in group A for variable 1 and group F for variable 2 is clearly 3. This can be deduced using the other cell values and the marginals. Removing the marginals would significantly increase the protection offered by the table but often (and this will be assumed in this section) the user requires the marginal values to be retained in the published tables.

This shows that primary suppressions are not sufficient to protect a table against potential disclosure therefore further non-sensitive cells must be suppressed to protect the table. These further suppressions are called secondary suppressions. The choice of secondary suppressions should be such that the value of the cell that is subject to the primary suppression cannot be computed exactly. Unfortunately, as Evans et.al. [41] write, 'While the concepts behind determining whether a particular cell is a disclosure risk are relatively simple, the process of choosing complementary suppressions to protect these sensitive cells is very complicated. The methodology by which complementary suppressions are chosen, as well as the accompanying

computer software, is very difficult to understand for anyone without a background in linear programming.' For a table to have sufficient protection from potential disclosure there must be either no suppressions or at least two suppressions in each row and column. It will be possible however for a range of potential values for the suppressed cell to be calculated. An example of how secondary suppressions could be used on *table 2.2.8* is given in *table 2.2.9*.

		Variable 2				
		E	**F**	**G**	**H**	**Total**
Variable 1	**A**	23	+	+	18	81
	B	+	15	+	119	147
	C	54	43	8	4	109
	D	+	+	22	10	67
	Total	97	77	79	151	404

Table 2.2.9: Secondary suppressions removed from Table 2.2.8.

In *table 2.2.9* it is not possible to compute the exact value for any of the cells. Often tables are much larger than the example given and when there is a large table with a number of primary suppressions the choice of secondary suppressions becomes relatively complex. In this scenario the aim of the secondary suppressions is to sufficiently protect the primary suppression whilst limiting the information loss caused by carrying them out. Willenborg and de Waal [16] suggest that there are three important aspects to secondary cell suppression;

a. The sensitive cells should be adequately protected by the choice of the secondary suppressions; the ranges in which the values of the suppressed cells lie, should not be too narrow. It should be borne in mind that calculating ranges is possible when the values of the cells are restricted in some way, e.g. by a requirement that they be nonnegative.
b. The loss of information due to the secondary suppressions should be minimized.
c. No zero-valued cells or empty cells should be suppressed.'

The choice of secondary suppression affects the size of the range of values in which it is known the suppressed cell lies. This range of values is known as a feasibility interval and can be constructed using linear equations of the marginals and the remaining cell values. This can be done for small tables by hand or by using computing software for larger more complex tables. The statistician must decide before the suppression how large the feasibility interval should be to offer sufficient protection and then ensure that the suppression carried out has provided this protection.

The justification behind not suppressing empty cells is that in many cases empty cells will be known to be empty before publication of the data e.g. in the brain surgery example if one of the procedures was some form of brain surgery and a hospital did not have facilities to carry out that operation it would be known that this cell will be empty. Suppressing an empty cell would not aid disclosure in this case since the prior knowledge combined with the values in the marginals and non-suppressed cells may result in disclosure of the value of the primary suppressed cell.

The aim of the secondary suppressions is to sufficiently protect the primary suppression whilst limiting the information loss. To quantify the information loss each cell must be given a weight. There are many ways to assign weights (in fact any sensible system is acceptable) to the cells in the table and the choice of weights can often be dependent on the aim of the study or the nature of the data. The idea of these weights is to target the damage so as to maintain the primary purpose of the output as much as possible, i.e. the intention is to concentrate the damage as far as possible on the 'least useful' data. This often means the cells containing the lower frequencies. Three potential choices of weights are:

1. Weights to minimize the total number of cells suppressed. Here all cells are given the same weight i.e. $w_{ij} = c$ where $c > 0$.
2. Weights to minimize the value of the suppressed cells. Here the weight of the cells is equal to the value of the cell.
3. Weights to minimize the number of respondents suppressed. Here the weight of the cell is the same as the number of subjects that contributed to the cell.

Clearly the second choice of weights would only be applicable if the table contained magnitude data and not possible if the table contained frequency data showing how the choice of weights should depend on the nature of the data.

Cell Suppression is an extremely popular technique for protecting tables against potential disclosure. The major advantage of cell suppression as a dislcosure control technique is that since some of the cells are suppressed there is no need to perturb the

values in the table. This means that all the information provided in the table is accurate which is helpful for any end user that wishes to use the information to run further statistical tests. A further advantage is that the marginals are retained in their original form in the table (assuming they are protected from secondary suppressions) which is often helpful to the end user. This form of suppression is not difficult to explain to a non-statistical minded end user since only sensitive cells are suppressed and none of the values are adjusted.

In spite of the popularity of cell suppression as a disclosure control technique there are certain disadvantages encountered when implementing it. A major disadvantage is that end users can often feel uncomfortable about cell values being suppressed in tables. This is because either: suppressed values may make the end user (particularly if the end user is the general public) feel that the statistical organisation is attempting to 'hide' information or if the end user is attempting to carry out further analysis these suppressed values can cause serious problems. The issue of the end user feeling that the statistical organisation is attempting to 'hide' information is especially important due to the importance of public trust in the organisation. Another disadvantage of cell suppression is that since secondary suppressions are required to protect the marginals the information from the secondary suppressed cells is lost even though those cells are safe.

On a small table secondary cell suppression is not difficult and can often be simply calculated by hand however when the table becomes large and/or there are a large number of primary suppressions, the problem can become complex. The complexity arises from the fact that the potential number of combinations of secondary

suppressions can become very large. There are many methods currently used to compute secondary suppressions. For large tables these methods often require a significant amount of computing power. Four of these methods are hypercube, modular, optimal and network.

2.2.2.3 Hypercube Approach

To attempt to simplify the secondary suppression problem, the hypercube method considers the primary suppressed cells consecutively. For each cell, a hypercube is chosen where the primary suppressed cell is one of the corner points. Once the hypercube is chosen all its corner points are suppressed and the suppressed cells are given a large negative weight. In the two-dimensional case that is mentioned in this section this hypercube is a rectangle. The rectangle is chosen so as to minimise the loss of information that is incurred by the suppression. The loss of information is quantified by the weights (as discussed above) that are assigned to the cells and the chosen hypercube (rectangle) is called the suppression hypercube. This procedure is carried out for all the primary suppressed cells. It is often the case that cells that have already been suppressed will be included in future hypercubes due to their large negative weight therefore hypercubes for different primary suppressed cells may contain some of the same cells.

Once all the hypercubes have been selected the width of the feasibility intervals of the primary suppressed cells must be calculated to ensure that they are sufficiently wide. There are two recognised techniques for dealing with this problem. One of these techniques involves carrying out the above procedure and when all the suppressions

due to the hypercubes are completed, checking by means of a linear programming problem, whether all the feasibility intervals are sufficiently wide. A major disadvantage of this method is that if even one feasibility interval is too narrow the whole suppression pattern is rejected and it is then difficult to adjust this pattern resulting in a waste of time and resources.

The other recognised technique is to take into account the necessary width of the feasibility interval for each suppressed cell and include this in the decision of the suppression hypercube. This implies that the hypercube chosen as the suppression hypercube is the one that minimises the information loss whilst ensuring the feasibility interval is sufficiently wide. An example of this technique is a heuristic proposed by Repsilber [42]. Suppose there is a sensitive cell that has to be protected against disclosure and there is a proposed hypercube. Each corner point of the hypercube is either an even number or an odd number of steps from the sensitive cell. If the corner point is an even number of steps from the sensitive cell it is called an even corner point and if it is an odd number of steps from the sensitive cell it is called an odd corner point. Note:- the sensitive cell is an even corner point.

Suppose now that a value, ε_-, is subtracted from the sensitive cell. This implies that, in order to preserve additivity of the table, ε_- must be added to all odd corner points and subtracted from all even corner points. Also, in order to preserve non-negativity, the minimum value of the even corner points is equal to the maximum value of ε_-. Conversely when a value, ε_+, is added to the sensitive cell the minimum value of the odd corner points is equal to the maximum value of ε_+. Therefore an intruder can estimate the value of an even corner point to lie in the interval

$$(x_e - \max \varepsilon_-, x_e + \max \varepsilon_+)$$

where x_e is the true value of the even corner point. The value of an odd corner point can be seen to lie in the interval

$$(x_o - \max \varepsilon_+, x_o + \max \varepsilon_-)$$

where x_o is the true value of the odd corner point. If the values $\max \varepsilon_-$ and $\max \varepsilon_+$ are large enough, i.e. the feasibility interval is sufficiently wide, then the proposed hypercube becomes a candidate suppression hypercube. The suppression hypercube should then be the candidate suppression hypercube that results in the least information loss. This technique should be carried out on all primary suppressed cells to produce a protected table. A major advantage of this is that complex linear programming problems are avoided, however it takes no account of cells that have already been suppressed in previous secondary suppressions. One of the limitations of the hypercube method is that, Massel [43] writes, it 'often oversuppresses (estimate of 30%) and it may not find the best suppression pattern even for a single sensitive cell. This occurs because it is considering only the simplest types of suppression patterns.'

2.2.2.4 Modular Approach

If a table contains a hierarchical variable any secondary suppressions might lead to more secondary suppressions being required in related tables. It has been suggested

by Fischetti and Salazar Gonzalez [44] that adding additional constraints to a linear programming problem will allow tables from hierarchical variables and also linked tables to be protected. The problem with this is that when dealing with hierarchical variables the number of constraints grows rapidly due to the inter dependency of the many possible sub tables and sub totals. This results in the computing power and time required to carry out these calculations also increasing.

The modular (HiTaS) approach put forward by de Wolf [45] for secondary cell suppression is a heuristic approach used for suppressing hierarchical tables. This approach deals with a large set of sub-tables individually in a structured order to reduce the computation time. A potential problem with this method is that it produces a solution that minimises the information loss in the individual sub-tables but does not necessarily minimise the total information loss.

The modular approach uses a top-down approach to deal with the secondary suppressions. The first stage of the process is to identify the unsafe cells (i.e. the primary suppressions) in the base table which is produced by crossing the hierarchical variables. The secondary suppressions must now be calculated. The fact that hierarchical variables are being used means that the secondary suppressions must be chosen in such a way that different sub-tables cannot be combined to unpick the values of sub-tables further up the hierarchy. The basic idea of the approach is to firstly calculate the secondary suppressions for the highest-level table (i.e. the table with the highest level for all hierarchical variables). The interior cell values (whether suppressed or not) then become the marginal values in lower level tables. The process then moves down to the next level of tables below the highest level. The

marginals (from the interior cell values of the highest level table) in these lower level tables stay fixed (or protected) and secondary suppressions are calculated for these tables. The secondary suppressions calculated here can be either to protect any of the inserted marginals which have been suppressed or any primary suppressions in the interior cells of the lower level table. This process is repeated until the lowest level table has been protected.

Since the secondary suppressions are restricted to interior cells, if there is a lower level table with several empty cells it may be the case that it is impossible to find a solution. In this scenario the process carries out the suppression on the table from the level above the problem table again with the suppression pattern from the lower level table considered i.e. the tables cannot be considered independently.
It is clear that the order in which the tables are suppressed is important. It is crucial that this order is planned out in advance. The process deals with this by defining tables into certain classes. To illustrate how this is done suppose there are two hierarchical explanatory variables each with three levels (0,1,2). The groups are defined by a crossing of the levels of the explanatory variables. The classes associated with these variables and the groups that would be members of each class are given in *table 2.2.10*.

Class	Groups
0	00
1	10, 01
2	20, 11, 02
3	21, 12
4	22

Table 2.2.10: How groups are defined into classes for the modular approach to secondary suppression.

De Wolf [46] writes 'Defined in this way, marginals of the tables in class i have been dealt with as the interior of tables in a class j with $j < i$. As a result, each table in class i can be protected independently of the other tables in that particular class, whenever the tables in classes j with $j < i$ have been dealt with.'

Clearly the number of tables in each group differs depending on the nature of the hierarchical variable. The number of tables in a group is in fact dependent on the number of 'parent categories' the variables have one level up in the hierarchy. The Tau-Argus user manual [35] notes 'A parent category is defined as a category that has one or more sub-categories.' For example suppose there is a hierarchical variable with three levels of the form given in *figure 2.2.3.*

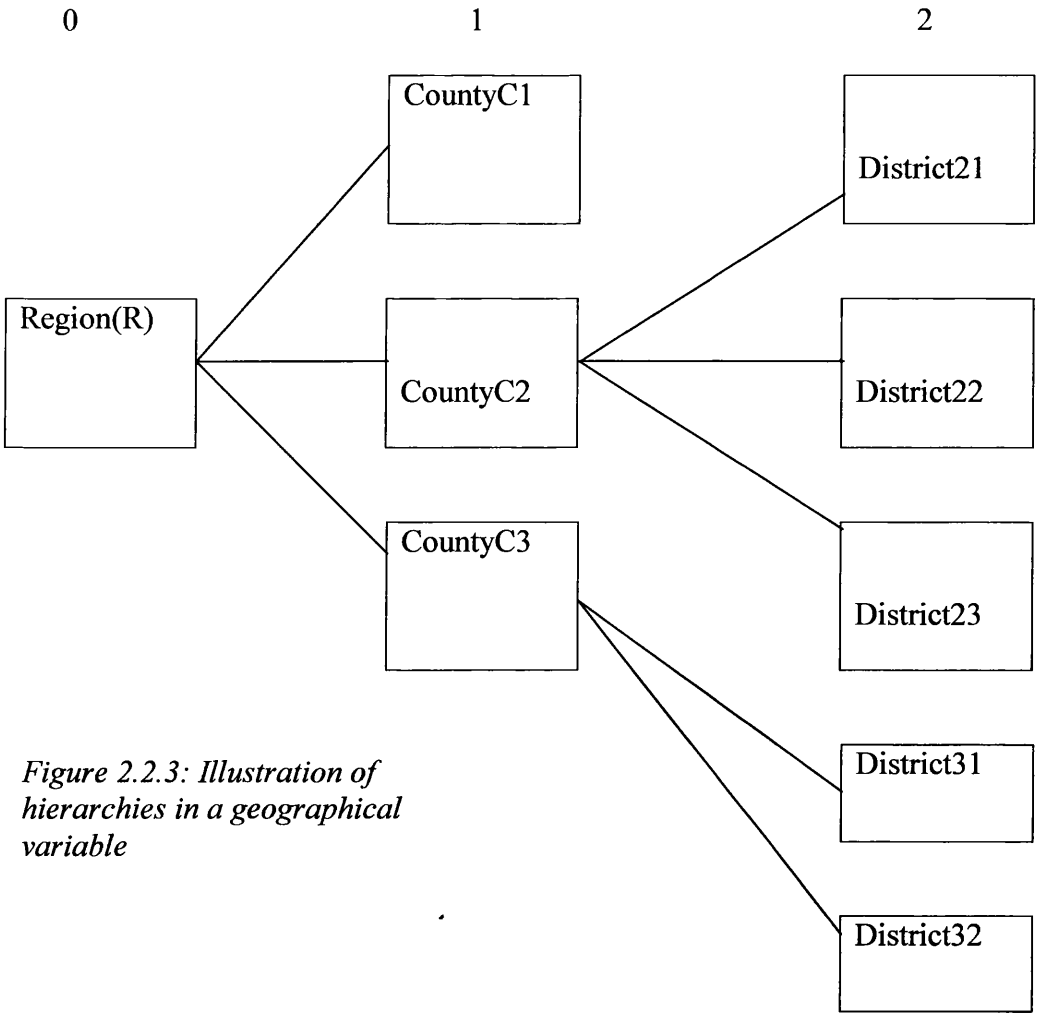


Figure 2.2.3: Illustration of hierarchies in a geographical variable

In *figure 2.2.3* level 1 has 1 parent category and level 2 has 2 parent categories. The number of tables in each group is equal to the product of the number of parent categories of each level.

2.2.2.5 Optimal Approach

The optimal approach to secondary cell suppression attempts to find optimal or at least near optimal solutions to the problem. The optimization models used have been developed for Tau-Argus by a team of researchers headed by Juan-Jose Salazar-

Gonzalez of the University La Laguna, Tenerife, Spain. The approach according to the Tau-Argus user manual [35] 'is based on a Mathematical Programming technique which consists of solving Integer Linear Programming programs modelling the combinatorial problems.' The major characteristic in the model is that it is based on a 0-1 variable for each cell where the value of the variable is 1 if the cell is to be suppressed and 0 if it is not to be suppressed. Therefore the number of variables in the model is the number of cells to be protected. Also included in the approach is a second model to ensure the safety ranges are maintained by the suppression. This model takes into account the fact that an intruder trying to compute the value of a sensitive cell will attempt to use two linear programming programmes (attacker problems) to calculate the suppressed values. The approach is called a branch-and-cut algorithm since the presence of the 0-1 variable makes a branching stage necessary.

According to the Tau-Argus user manual [35] 'Branch-and-cut algorithms are modern techniques in Operations Research that provide excellent results when solving larger and complicated combinatorial problems arising in many applied fields. Shortly, the idea is to solve a compact 0-1 model containing a large number of linear inequalities through an iterative procedure that does not consider all the inequalities at the same time, but generates the important ones when needed.' This allows large models to be split into a short sequence of small models.

Also important to the whole process is the pre-processing approach. This eliminates redundant equations defining the table, removes variables associated to non-relevant cells and detects dominated protection levels. This pre-processing is crucial to the process since it makes the problem as small as possible before the optimization stage.

The heuristic routine also plays a key role in the approach. It allows the process to start with an upper bound for the optimal loss of information and different protected patterns are produced as the process runs so that if the process is stopped before optimality is proven a near optimal solution is given. The best solution is called an optimal solution if its loss of information is equal to the lower bound for optimal loss of information where 'the lower bound is computed by solving a relaxed model, which consists of removing the integrability condition on the integer model.'

2.2.2.6 Network Approach

The Network approach to secondary cell suppression is used to suppress two-dimensional tables with one hierarchical variable. The heuristic for this approach computes a sequence of shortest-path sub problems to give a feasible solution to the problem which is hopefully close to the optimal solution.

The approach requires tables to be modelled as networks. To do this the procedure firstly lists all the possible tables in a hierarchical tree. The highest-level table is then extracted and a network created for it.

The original hierarchical network will be successively updated. This is done by an iterative procedure for each of the other lower level tables which extracts one table, creates its network and then updates the hierarchical network. When all the tables have been used to update the hierarchical network this network models the hierarchical table.

Once the hierarchical network has been produced one iterative procedure is carried out for each primary suppression. If not already protected by the protection of previous primary suppressions the primary suppression is protected by setting the cost of arcs in the network. The arc is the path from one cell to another. Any arcs relating to cells which cannot be suppressed are given a very high cost. A shortest path is then from one end of the network to the other and is defined as the path which results in the least information loss. All cells in the shortest path will then become secondary suppressions. The procedure protects both the upper and lower protection bounds for each of the primary suppressions. These bounds are updated throughout the process, Domingo-Ferrer and Torra [47] writes that this is 'to avoid the solution of unnecessary shortest-path subproblems for next primaries, we update not only the protection levels of p (primary suppression of interest in iteration), but also of all the primary cells in the shortest path'. Once all the iterations in the procedure have been carried out the table will be an adequate solution to the secondary suppression option and it is hoped the solution will be, whilst not exactly optimal, close to optimal.

Whilst this procedure has the advantage of being extremely fast computationally it has certain limitations. These include, according to Massel [43], 'oversuppression is common for all cases but usually tolerable, limited types of cost functions can be expressed' and undersuppression can occur for hierarchies with more than 2 levels.

Cell Suppression is one of the major tools used by Tau-Argus. All the secondary suppression methods mentioned above are available although clearly since some methods relate to certain types of data they can only be used for certain data sets. Tau-Argus also allows the user to manually decide which cells are suppressed. In the

program the user can view the status of the cell. The status of the cell can be any of the options given in *table 2.2.11.*

Status	Definition of Status
Safe	Does not violate the safety rule
Safe (from manual)	Manually made safe during the session
Unsafe	According to the safety rule
Unsafe (request)	Unsafe according to the Request rule
Unsafe (frequency)	Unsafe according to the minimum frequency rule
Unsafe (singleton)	Unsafe due to singleton suppression
Unsafe (singleton)(manual)	Unsafe due to singleton suppression but primary suppression carried out manually
Unsafe (from manual)	Manually made unsafe during this session
Protected	Cannot be selected as a candidate for secondary cell suppression
Secondary	Cell selected for secondary suppression
Secondary (from manual)	Unsafe due to secondary suppressions after primary suppressions carried out manually
Zero	Value is zero and cannot be suppressed
Empty	No records contributed to this cell and the cell cannot be suppressed

Table 2.2.11: Definitions of the various statuses that cells can take in Tau-Argus.

There is an option in Tau-Argus for the user to change the status of the cell. The options are:

- Set to Safe: A cell which has failed the safety tests is still to be considered safe.
- Set to Unsafe: A cell which has passed the safety tests is to be considered unsafe.
- Set to Protected: A safe cell that is not to be considered for secondary suppression.

These options can be extremely useful in circumstances where the user needs to publish data on certain sections of the table. For example there may be a number of tables on aborted pregnancies for different age groups across different areas where the primary subject of interest is abortions in middle-aged women. It may be the case that there are many unsafe cells for abortions in teenage pregnancies in certain areas and that for these to be suppressed the secondary suppressions may remove some of the data from abortions in middle aged women. By setting the cells of primary interest as Protected this problem would be removed.

2.2.3 Summary of Techniques Used to Remove Potential Disclosure in Tabular Data

The techniques which have been described to protect against potential disclosure in tabular data are summarised in *table 2.2.12*.

Technique	Variants	Parameters
Rounding	*Conventional Rounding* *Random Rounding* *Controlled Rounding*	*b - rounding base*
Adding Noise	Additive Noise Multiplicative Noise	ε - non-biased random variable
Cell Suppression	*None*	*secondary suppression technique*
Table Redesign	*None*	*variables to be grouped*

Table 2.1.12: Summary of Disclosure Removal Techniques. Techniques available in Tau Argus are shown in italics.

2.3 Quantifying Information Loss

The aim of statistical disclosure control is to produce a table that is free from (or at least has a very small probability of) potential disclosure risk. However whilst the table must be protected it is important that as much information as possible is retained by the table. Therefore to examine the performance of a particular disclosure control technique on a table of data the information loss must be quantified. To do this each cell must firstly be given a weight. As in section 2.2.2 three potential weights for cells are:

1. Unity – Each cell has a weight of 1.
2. Frequency – The weight of each cell is the number of contributors to that cell.
3. Variable – The weight of each cell is the total value of the variable in the cell. This can only be used if the table contains magnitude data.

Cells can be weighted using other techniques if the user requires a certain style of weighting although this is unusual.

Each method of disclosure control requires slightly different techniques to quantify the information loss in the table:

1. Cell Suppression – The information loss is the total of the weight values of all the suppressed cells.
2. Rounding – The information loss for each individual cell is $w|x-y|$ where w is the weight of the cell, x is the original value and y is the rounded value.

Then the overall information loss is the sum of the information loss over all the cells.

3. Adding Noise – The information loss is calculated in exactly the same way as for Rounding with y the value with noise added.
4. Table Redesign – The information loss is the sum of all the weights of the collapsed columns and/or rows.

The disclosure method chosen will usually be the one that minimises information loss although this may not always be the case. There may be situations where the user wishes to use a certain method of disclosure control. For example suppression may result in the smallest information loss but the user may not want suppressed cells in the table. Also it is often the case that tables have more categories than is necessary, e.g. having individual ages rather than age groups, and although table redesign may result in a fair amount of information loss it is information that is not crucial to the user that is lost therefore table redesign may be preferable. In general though the most effective disclosure control method is the one which results in the least information loss.

3. Selection of Statistical Disclosure Control Method

3.1 Practical Issues in Deciding Between Disclosure Methods

As has been shown statistical disclosure control has both scientific and social influences. If the problem were purely statistical it would be relatively simple to quantify the information loss produced by each disclosure control method and select the method which results in the least information loss. However there are often other practical issues which must be taken into account when the table is protected. These practical issues often require the statistician to subjectively investigate the data before they proceed with the disclosure control and to liaise with both the provider of the data and the end user of the data. This communication between the statistician and the data user is essential as the statistician will not only receive an idea of what the user requires from the data but may gain extra background information on the data.

The most subjective stage of statistical disclosure control is the decision on whether a cell contains a potential disclosure risk. It is at this stage that the statistician must take the most care over how much control is given to the user regarding the test used to detect potential disclosure. The reason for this is two fold. Firstly if there is any breach of confidentiality the responsibility ultimately lies with the statistician and their organisation and such a breach may result in future lack of trust in the statistical organisation or possibly even legal action. Secondly the end user may appear to be credible and well intentioned however it may be that they are trying to expose information on an individual and if they have too much knowledge of the safety test it may make it possible for the disclosure control method to be 'unpicked'. However,

useful information can still be gathered from the end user to assist in the process of discovering potential disclosure risks. For example if the statistician has decided on a minimum frequency rule it may be possible to learn from the end user (if they are an expert in the field) how many subjects would likely be required to reduce the disclosure risk to a reasonable level. Also if the statistician wished to use the P% test it might be useful to learn the perceived possibility of coalitions being formed by subjects in the table. It is important in these circumstances for the statistician to consult with the end user whilst retaining control over the decisions regarding detecting potential disclosure risk.

The statistical disclosure control technique which often provides the most useful results to the end user is table redesign. It is often the case that a redesigned table can still contain all the information required by the end user while removing the potentially disclosive aspects of the data. In many cases data are gathered in great detail and while this is often necessary for studies into exact conditions it may be the case that another study on the same data does not require such detail and can be simplified. For example there may be a large amount of data on a condition for subjects of every age. One study may be concentrating on children and therefore requires the data split up by year whereas a larger study on the population as a whole may not require individual years of age therefore age groups may provide less disclosive but equally useful data. The application of table redesign has many practical issues. The statistician needs to be aware of how detailed or otherwise the end users requires each variable to be or whether the user requires a certain level of a variable to remain and not be merged with other levels. Another important aspect of table redesign is that the table will only retain its usefulness if levels of the variable

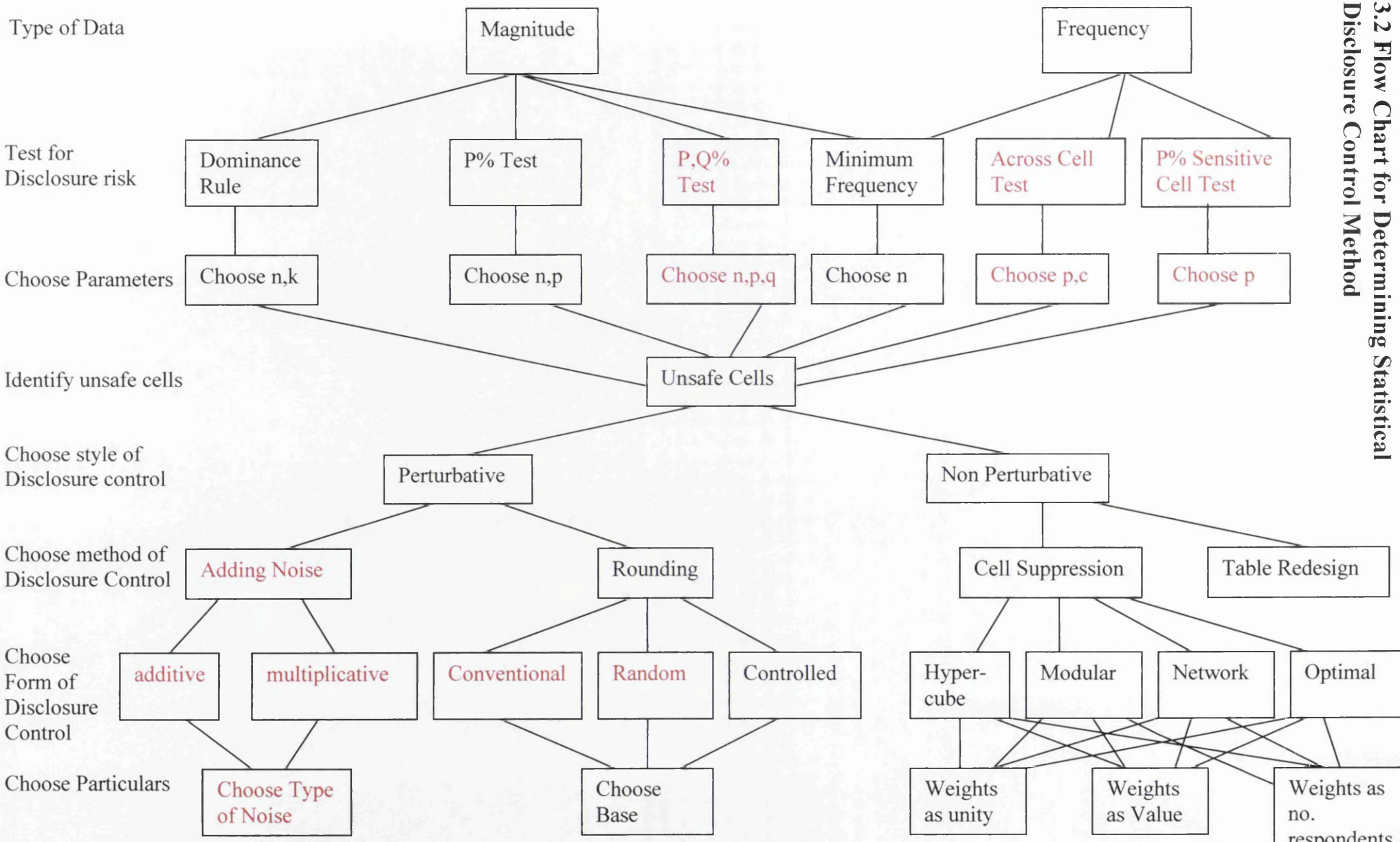
can be sensibly merged. For example if there is a condition which is prevalent in adults but not so in children or the elderly it does not make sense to protect the table against potential disclosure by merging the data for children and the elderly. In this example it is obvious to even the casual observer that this redesign is flawed however for more complex and specialised data it may not be so clear. In these situations the statistician is required to communicate with an expert (preferably the end user) to gather information on potential groupings.

One of the major differences between the disclosure control methods is the appearance of the final table. In many situations the end user will have a preference as to how the table is set out. One of the major decisions to be made is whether the end user prefers a table which has the values in each cell perturbed but is complete or whether they would rather a table which has some cells suppressed but has the original data in the other cells. There are many factors which may affect this choice. For example if the cell values were on the whole large except for a few entries it may be preferable to use rounding since the rounding may not have such a relative effect on the final results. On the other hand the table may contain cells which are not of primary interest to the final user but are found in the data. In some circumstances these cells may be suppressed (as either primary or secondary suppressions) to reduce the disclosure risk to a safe level without affecting the cells of interest. Again the importance of communication between statistician and the end user is essential to ensure a solution is reached which is mutually satisfactory. This choice can often be made by the end user, unfortunately there are occasions when certain disclosure control techniques (in particular cell suppression) are unavailable and the statistician

may have to insist a certain technique is used to avoid releasing potentially disclosive data.

It is clear that it is helpful to the statistician to have input from the end user in the statistical disclosure control process. The end user can often supply useful information to the statistician to assist in the disclosure control process. It is the job of the statistician to provide the information that is of most use to the end user and consultations between the two parties is crucial for this. It is important to remember however that the statistician should ensure that there is an acceptably low risk of disclosure. This may in some cases result in the user being unaware of certain techniques which have been used in the process or techniques being used against the end user's wishes. This is unfortunate but the statistician should always err on the side of ensuring low disclosure risk ahead of high data utility. Further the statistician should always endeavour to communicate with an expert in the field, even when the end user is unavailable/unhelpful, to assist with the practical aspects of statistical disclosure control.

3.2 Flow Chart for Determining Statistical Disclosure Control Method



Note - Procedures not available in Tau-Argus indicated in red.

3.3 Formal Discussion of Disclosure Control Flow Chart

The process of statistical disclosure control is a rather complex process with many user choices. In both the attempt to detect any disclosure risk in the data and the process of protecting against any such disclosure the statistician must make choices between a variety of techniques and the parameters associated to these techniques. These choices can significantly affect the construction of tables of protected data and it is important that the methods are planned in advance. A major advantage of the reduced computation time and large number of user options offered by the Tau-Argus program is that a number of disclosure control methods can be attempted on unsafe data to investigate the information loss due to each method. Section 3.2 contains a flow chart for determining the statistical disclosure control process which will be implemented on the data. This flow chart was produced at the request of ISD as an aide to allow staff to be trained in the techniques of statistical disclosure control since this is not an area covered in an undergraduate degree in statistics. This section will briefly discuss the choices to be made at each stage of the flow chart and discuss their implications.

3.3.1 Type of Data

Before even starting the process of disclosure control it is important that the statistician is aware of the composition of the data being used and what the particular use of the final table is. The data must be subjectively investigated to provide the statistician with an overview of such things as potential groupings and the size of magnitude data. It is also useful for the statistician to spend some time getting a

general 'feel' for the data. At this stage the user must know the important information that should be retained, if at all possible, in the safe table. This knowledge allows the statistician to make decisions later in the process that allow this information to be retained.

As mentioned in section 1.4 an important distinction that must be made is one between frequency data and magnitude data. This distinction must be made since the different types of data correspond to different tests for disclosure risk.

3.3.2 Test for Disclosure Risk

Of all the decisions in the statistical disclosure control the choice of safety test is probably the most important. Without properly testing for potential disclosure there is a risk that tables are produced that carry potential disclosure about an individual subject and therefore the potential for harm to the individual subject. In a simplistic world, as long as tests for disclosure risk were carried out, it would be possible to avoid any disclosure by not publishing any table that carries any risk. This is clearly not practical due to the need society has for the information, hence the disclosure control methods outlined in this project, but it shows the importance this stage has in protecting individual subjects.

The flow chart shows that when the data being protected are magnitude data there are 4 potential safety tests which can be selected. These are; the dominance rule, the p% test, the p,q% test and the minimum frequency test. An overview of the theory and practicalities of these tests is given in section 2.1.2. Either one or a combination of all

of these tests can be used to discover the cells in the table which contain a potential disclosure risk. It is often beneficial to use a combination of the tests to improve the protection given in the table. For example the minimum frequency test can be used in conjunction with any of the other tests to ensure there are no cells with few subjects contained and also that the magnitude values given by the subjects do not pose a potential disclosure risk.

The flow chart shows that when a table containing frequency data is being protected there are 2 potential safety tests which can be selected. These are; the p% sensitive cell test, the across test and the minimum frequency test. An overview of the theory and practicalities of these tests is given in section 2.1.1. These tests should be used individually and often the minimum frequency test is the preferred option due to the simplicity of the test.

Once the safety test has been selected the parameters for the required test must be selected. The choice of these parameters is crucial to the protection given by the safety rule. The flow chart shows the parameters which must be chosen for each of the tests. The choice of parameters is discussed in section 3.4.

3.3.3 Identify Unsafe Cells

Once the test for disclosure risk and the associated parameters have been chosen the process can identify the unsafe cells. These cells now have to be adjusted and/or suppressed in some way to ensure the final table is free of disclosure risk and safe for dissemination. In some circumstances the statistician will be able to set an unsafe cell

to safe since either they believe the issue to not be sensitive enough to require disclosure control or the cell will be retained at the expense of another (or other) cell(s). There are many techniques that can be used to protect the table and the choice of these is the next stage in the disclosure control.

3.3.4 Style of Disclosure Control

Once the unsafe cells have been discovered the statistician must decide on a style of disclosure control. As the flow chart shows all disclosure control techniques fall into two separate groups regarding the style of the disclosure control. These are:

- Perturbative techniques – the data are adjusted in some way to provide uncertainty to the true value of each cell.
- Non-perturbative techniques - the data are not adjusted but often some cells or groups are not published and are hidden from public view.

The choice of the style of disclosure control is an extremely important stage in the process. An important consideration to make is who the end user of the data is. Some end users may prefer to have a table that is complete but has all the cells adjusted whereas other users may prefer less detail in the table as long as the original values are used. The decision should be made by the statistician with input from the end user to ensure that they are satisfied with the final table provided. Methods from both styles of disclosure control are available in Tau-Argus.

3.3.5 Method of Disclosure Control

Once the style of disclosure control has been chosen the statistician must choose a method of disclosure from the available methods from that style. Both styles have a variety of methods from which the final one can be chosen. The choice of method of disclosure control determines how the final table is produced. The differences can be relatively significant and affect potential further study making the choice of method crucial.

The flow chart shows that if perturbative techniques are to be used there are two main methods for disclosure control: adding noise and rounding. Rounding is the most commonly used perturbative disclosure control technique. Rounding is the only perturbative technique available in Tau-Argus. Another perturbative technique that is not as popular as rounding, and therefore not available in Tau-Argus, is adding noise to the data. An overview of the theory and practicalities of both methods is given in section 2.2.1.

The flow chart shows that if non-perturbative techniques are to be used there are two main methods for disclosure control: cell suppression and table redesign. Table redesign is usually preferable and should be used whenever possible however it is not always a sensible option so cannot always be used. Cell suppression on the other hand is almost always possible and is the most commonly used non-perturbative technique. Both methods are available in Tau-Argus. An overview of the theory and practicalities of both methods is given in section 2.2.2.

3.3.6 Form of Disclosure Control

Once the method of disclosure control has been selected the statistician must make a decision on the form the disclosure control method will take. Each method of disclosure control can take a different form depending on the choice made by the statistician. This choice can in fact significantly affect the make up of the protected table with the different forms of disclosure working in different ways. Once again the form of disclosure control chosen is dependent on the data but is also dependent on which method the statistician finds more effective and/or user friendly. Consultation with the end user can again be useful at this stage.

The flow chart shows that if adding noise has been chosen as the method of disclosure control there are two potential forms of the method which can be used: additive noise or multiplicative noise. Clearly the difference between these two forms is quite large. A major difference (and in fact a disadvantage of additive noise) is that when additive noise is used it is difficult to retain the additivity of the table whereas when multiplicative noise is used an iterated proportional fitting (IPF) method can be used to retain additivity. However additive noise is simpler to apply and in many cases results in less information loss so is often the preferred form. Once the form of noise has been decided a random variable must also be chosen. Any sensible unbiased random variable, with a variance chosen according to the data, can be selected according to the wishes of the statistician.

The flow chart shows that if rounding has been chosen as the method of disclosure control there are three potential forms of the method which can be used: conventional

rounding, random rounding or controlled rounding. All three forms of rounding offer various advantages and disadvantages and should be used accordingly. Conventional rounding is rarely used since it offers little protection to the data. Controlled rounding is often chosen over random rounding since it offers good protection to the data whilst retaining additivity in the table (an advantage that no other rounding achieves). Once the form of rounding has been decided a rounding base must be chosen. This rounding base can be any sensible base according to the size of the data.

If cell suppression has been chosen as the method of disclosure controlled the statistician must choose between the four possible secondary suppression techniques. The flow chart shows that these are: hypercube, modular, network or optimal. The choice of secondary suppression technique is not simply determined by the statistician's preference since individual methods are necessary for certain data types. The hypercube and optimal secondary suppression techniques can be used on most forms of data and the hypercube method is the most commonly used however, as the name suggests, the optimal technique can in some cases give a slightly improved suppression but in large tables this may be time consuming and require large amounts of computing power. The modular secondary suppression technique is used on data which contains at least one hierarchical variable. The network secondary suppression technique is used on two-dimensional tables with one hierarchical dimension and is faster and requires less computing power than the modular technique. All four types of secondary cell suppression are available in Tau-Argus.

To allow secondary suppression to be carried out effectively all the cells need to have a value for the information loss of suppressing them. This information loss can be

measured as weights for each cell. The way in which cells are weighted is the decision of the statistician with any sensible method acceptable. The flow chart shows three common methods of weighting cells are:

1. Weights to minimize the total number of cells suppressed. Here all cells are given the same weight i.e. $w_{ij} = c$ where $c > 0$.
2. Weights to minimize the value of the suppressed cells. Here the weight of the cells is equal to the value of the cell.
3. Weights to minimize the number of respondents suppressed. Here the weight of the cell is the same as the number of subjects that contributed to the cell.

3.4 Parameter Selection Protocol

The statistician must make a number of choices in the statistical disclosure control process. These choices are not restricted to deciding on which particular disclosure control method should be used. Once the statistician has decided on both a test for potential disclosure and a technique for removing any potential disclosure, a choice must be made regarding the parameters for each test. Of the many methods available almost all require some decisions to be made regarding the values used. These values can make a significant difference to the final table, affecting which cells are considered safe and also the final make up of the table. It is important for the statistician to study the data before selecting parameters since different sets of data will require different parameters to ensure the data are made safe for release.

The selection of parameters in tests for detecting potential disclosure in a table is essential to ensure confidentiality is protected. If the parameters are too lenient there may be a risk of potentially disclosive information being released however if the parameters are too restrictive the released table will have lost a great deal of its data utility. Once again the process is a trade off between disclosure risk and data utility even down to the level of parameter selection. Each safety test requires different parameters to be selected. *Table 3.4.1* shows the different parameter selections which are required for five of the common safety tests. As mentioned the selection of these parameter values is very subjective and the data themselves have a large bearing on the chosen values. Also consultation between the statistician and an expert should allow more accurate selection of parameters especially for parameters such as size of coalition and intruders prior knowledge of other subjects. If the statistician is in any doubt over the selection of a parameter they should always select a parameter which they are sure will provide adequate disclosure protection.

Method	Data Used On	Parameters	Frequently Chosen Values
Dominance Rule	Magnitude	n – number of contributors k – percentage contributed	n – 2,**3** k – 70,**75**
P% Test (across and within cell test)	Magnitude	p – accuracy of estimation n – size of coalition	p – **10**,15 n – **1** (no coalition), 2
P,Q% Test	Magnitude	p – accuracy of estimation q – accuracy of intruder prior knowledge of other subjects n – size of coalition	p – 10,15 q – 10,15 n – 1,2
Minimum Frequency	Magnitude and Frequency	n – minimum number of contributors	n – **3**,5
P% Sensitive Cell Test	Frequency	p – percentage of subjects in sensitive category	p – 70, 80
Across Cell Test	Frequency	p – accuracy of estimation c – size of coalition	p – 30, 40 c – 1,2

Table 3.4.1: Parameter selections required for various safety tests. Tests in italic are available in Tau-Argus and the default selections are given in bold.

Once the test to detect unsafe cells has been selected the statistician must select a technique to remove the disclosure risk from the table. As with the tests for detecting potential disclosure risk the techniques to remove disclosure risk require further selections of the style of each technique to be used. *Table 3.4.2* shows the parameters to be chosen for each of the four main techniques used to remove potential disclosure.

Technique	Parameters	Frequent Choices
Rounding	Style of rounding b – rounding base	Controlled Rounding b – 5,10
Adding Noise	Type of noise ε - non biased random variable	Additive Noise ε - Norm(0,σ)
Cell Suppression	Secondary suppression technique	Hypercube, Modular, Network, Optimal
Table Redesign	Variables to be grouped	Any sensible Grouping

Table 3.4.2. Parameter selections required for various techniques to remove potential disclosure. Techniques in italic are available in Tau-Argus.

The parameters chosen for each test are often dependent on the data. For example if the table contained cells with generally large values the rounding base or the variance of the noise may be large as the protection would be increased without a significant effect on the utility of the data. As mentioned in section 2.2.2 the choice of secondary suppression for the cell suppression technique is explicitly dependent on the data and in particular the make up of the variables, therefore this choice is not a selection to be made by the statistician but is decided by nature of the data. In the case of table redesign it is important (as mentioned in section 3.1) that the statistician consults with the end user to identify potential groupings that are both sensible and retain the utility of the data.

3.5 Practical Applications of Statistical Disclosure Control Methods

It has become clear throughout chapter 3 that there are many choices facing the statistician in the statistical disclosure control process. The theoretical effect of these choices has already been discussed and when a disclosure control procedure is being

discussed it is prudent for the statistician to consider the theoretical effect each procedure will have on the table. However it is of interest to investigate further the statistical effect these choices would have in a real scenario. It has already been noted that there may be many influences on the choice of disclosure control procedure however it is useful for the statistician to consider how effective each disclosure control procedure would be if these outside influences were disregarded. This could be achieved by carrying out a variety of statistical disclosure control procedures on a real set of data and comparing the information loss between the actual table and the disclosure controlled table. This will lead to a multitude of comparisons between the information loss produced by different techniques for both detecting potential disclosure in the data and removing potential disclosure from the data. The information taken from these comparisons may be useful to the statistician in selecting potential disclosure control process that they will follow in a variety of scenarios. Chapter 4 will concentrate on evaluating the performance of a variety of selected disclosure control methods on actual data.

4. Statistical Disclosure Control in Practice: An Application to ISD Data

4.1 Data Problem

In order to compare different methods used to control disclosure, data is required from which potentially disclosive tables may be produced. This comparison of methods will be carried out in this thesis, however steps will be taken to ensure that no potential disclosure results from the publication of the work. The data used to evaluate the performance of various disclosure control methods in this thesis are real data provided by ISD and report the results of a study carried out in diabetes in children in Scotland. Information was collected on children who had been diagnosed with diabetes and also on a control set. There were data from 365 children with diabetes and 499428 control children. A selection of variables was chosen that might have had a causal affect on the chance of a child developing diabetes. All the variables are categorical (meaning the resulting tables will be frequency tables) and a list of the variables and the number of categories it contains is given in *table 4.1.1*.

Variable	Description
Sex	2 categories
Age	10 categories
Geographical Indicator	Hierarchical variable, 26 categories (including 1 category for missing values)
Deprivation Decile	12 categories (including 1 category for missing values)
Year of Treatment	3 categories
Diabetes/No Diabetes	2 categories

Table 4.1.1: Description of variables in ISD data set.

The geographical indicator is a mixture of health board and the first 2 letters of the postcode. For example if the subject was from Greater Glasgow Health Board and their postcode was G12 8HX the geographical indicator would simply be GG1.

There are many characteristics of the data that give a high chance of potential disclosure. One of those characteristics is that there are so few children with diabetes compared to those control children which may result in cells with few subjects. Also the fact that some of the variables (in particular variables 2, 3 and 4) have so many categories may result in cells which have relatively few subjects. There is also a chance that when the missing values are included in the table that if the data are good (i.e. there are few missing values) the cells of the table pertaining to missing values will contain few subjects.

There is also an issue with regards potential disclosure of these data in the sense that diabetes in children may be regarded as sensitive data since illness in children and

causes of such are always issues which the media have a great deal of interest in. In fact the idea of an outsider being able to uncover the identity of a child with diabetes makes ISD uncomfortable in general. Unfortunately there are issues such as prejudices due to misinformation that follow people who are diabetic and it may be the case that the family of a certain child do not want their condition to be widely known. It is the responsibility of ISD to ensure that a disclosure situation of this type does not occur. This places added importance on the disclosure control method implemented by the organisation since there may be outside bodies wishing to uncover disclosure to benefit their own needs. On the other hand it is extremely important that researchers have good data from which to investigate potential factors which may potentially cause diabetes in children. Good research in this area may allow changes in policy to help reduce instances of the disease in children which is clearly beneficial to society. At the request of ISD, due to the fact that the data set contains sensitive data and potential disclosure may occur from tables of the data, no tables of the data will appear in this thesis.

4.2 Current ISD Policy

ISD currently have draft guidelines in place regarding potential disclosure from small numbers. The guidelines are regarded as advice which should be used along with the National Statistics Protocol on Disclosure. The advice pertains only to the potential disclosure caused by having cell values which are close to unity and only considers frequency data. Clearly this over simplifies the disclosure risk to the table since no consideration is made of the fact that there are many more complex settings in which

potential disclosure occurs. The guidelines provide a method of deciding on a treatment for each cell using four criteria:

- The size of geographical areas to be presented.
- The population size i.e. the marginal value.
- The numerator size i.e. the cell value.
- Whether or not the data in question is 'sensitive'.

The guidelines provide a description as to what should be deemed sensitive by ISD analysts. In fact a list is given which includes the major areas which should be deemed sensitive, although other topics can be included if required.

Sensitive Health-Related Data include:

- Sexually Transmitted diseases
- Abortions
- Pregnancies in girls aged under 16
- Suicides
- Self-harm
- Mental health diagnoses
- Mental health conditions
- Alcohol and Drugs misuse
- Prescriptions for contraceptives

The list is considered a framework but often requests come for lower level data and the analyst must consider whether the condition falls under one of these categories. This is not always as simple as it may seem e.g. for Kaposi's Sarcoma data requests the analyst checking would find that this condition is related to AIDS, thus linked to both a sexually transmitted disease and possibly falling within the umbrella of drugs misuse. Such an analysis should therefore be treated as sensitive. Also if an analyst is asked to prepare data on individuals with Korsakoff's Syndrome, a condition that is related to alcoholism, the analyst would be expected to treat it as sensitive data. It is important that the analyst investigates the data they are asked to reveal and ensures that they know whether the data fall into the sensitive category. This may require them to either refer to a medical dictionary or confer with a Consultant or some expert with specialist knowledge. Furthermore if there is a situation where the content of the data may allow the identity of a subject to be uncovered the data can be deemed as sensitive and not released. This decision is taken in consultation with the Caldicott guardian and/or the Head of Statistics.

ISD guidelines suggest using a technique called the barnardising method to protect unsafe cells. This is a perturbative method which adjusts the value of cells which contain small numbers to mask the true cell values. The method is only used on cells which contain less than 5 subjects. The adjustments of the original value are given in *table 4.2.1.*

True Value	Possible Presented Values
0	0
1	1, 2
2	1, 2, 3
3	2, 3, 4
4	3, 4

Table 4.2.1: Possible adjustments of the true cell value due to barnardisation

Possible values do not occur with equal probability. The choice of possible value is weighted with the true value having a larger possibility of selection than one of the false values. The ratio for weight is 1:4, e.g.

True Value = 1 Possible presented value = 1 (with probability 0.8) or
2 (with probability 0.2)

True Value = 2 Possible presented value = 1 (with probability 0.1) or
2 (with probability 0.8) or
3 (with probability 0.1)

The marginals in the table are also adjusted so as to protect the additivity of the table.

4.3 Application to ISD Scenario

As mentioned in section 4.1 the data used in this thesis come from working data used by ISD. This data set was thought to have potential disclosure issues due to the fact that there were few diabetes cases in rural areas such as the Highlands. However the study was an extensive piece of work carried out over a number of health boards which contained some interesting results therefore it was important that ISD used statistical disclosure control to allow data to be published. This led ISD to consider the number of options available to them to protect the table they had produced from this potential disclosure whilst retaining the utility of the data for further analysis. At the moment ISD have a policy on disclosure control which makes use of a technique called barnardisation (see section 4.2 for details). This technique could be applied to these data however it is of interest to discover whether other techniques may provide a more effective solution to the problem. This allows for a comparison between the effectiveness of the barnardisation technique and the various techniques discussed in this thesis.

The policy ISD implement takes account of whether or not the information gathered is regarded as 'sensitive'. There is a list of the areas regarded as 'sensitive' in section 4.2 and it is of interest to note that diabetes is not regarded as a sensitive area. It may be argued that any data relating to the health of children are sensitive especially due to potential interest from the media that would be generated by any startling results in the field of child health care. However no account is taken of the age of the patients in the ISD policy and it is assumed that the presence or otherwise of diabetes does not cause sufficient distress if discovered to the subject to be regarded as 'sensitive'.

There is though a sufficient risk to the confidentiality of the individual children in the study for disclosure control techniques to be implemented.

The aim of the study, as with most studies undertaken by ISD, is to provide results which may be presented to the public and can be used for both further analysis and for general information whilst not causing any risk to the confidentiality of the individuals. This provides ISD with its own problem in terms of how the data should be presented. This is due to the fact that it may be the case that one statistical disclosure technique is more informative for producing tables which are useful for further analysis whereas one technique may be preferential when it comes to disseminating general information. ISD cannot produce the same table which has been subject to two different disclosure control rules since this will significantly increase the possibility of disclosure and should be avoided. This scenario is made simpler when the table produced is part of a data request and the disclosure technique can be discussed with the end user before handover however in many cases the above scenario occurs and ISD must think about how best to balance the needs of academics with those of the general public.

4.4 Comparing Different Disclosure Methods on Frequency Data Using Tau-Argus

ISD do not currently use Tau-Argus to provide disclosure control for the data they produce. There is an interest from the management in potentially implementing Tau-Argus to protect their data. There is the suggestion that Tau-Argus can be used to provide a large number of disclosure control methods which provide adequate

protection to the data and the program is fast and simple to use. This chapter will investigate various methods provided by Tau-Argus on ISD data and evaluate their performance on ISD data with particular comparison to current ISD practice.

The flow chart in section 3.2 and the parameter selection protocol outlined in section 3.4 show the choices that must be made when selecting the disclosure method. This chapter will investigate the effects of the different choices using Tau-Argus when applied to the ISD data set with the aim of discovering how different choices affect the utility of the final table. As mentioned all the variables in the data are categorical which means that all the tables produced will be frequency tables and therefore the only method for detecting potential disclosure available through Tau-Argus is the minimum frequency rule. Once the unsafe cells have been identified Tau-Argus offers one perturbative and two non perturbative techniques for removing the potential disclosure from the data. The perturbative technique offered is rounding and the specific form of rounding offered by Tau-Argus is controlled rounding. The two non perturbative techniques offered are cell suppression and table redesign. When cell suppression is chosen the user can select from optimal, modular or hypercube suppression techniques and must also choose between weighting the cells due to the frequency of each cell or weighting each cell equally (unity). When table redesign is chosen the user must select the levels of each variable that can be grouped together. Also once table redesign has been carried out either rounding or cell suppression can be implemented on the new table.

There were many tables produced from the data which resulted in no potential disclosure risk according to the safety rule. These tables will not be considered and

onlytables where there was a potential risk present will be discussed. One of the advantages of Tau-Argus is that it can protect tables of many levels with relative ease however in this thesis only tables of 2 and 3 dimensions will be considered. This resulted in three 2-dimension tables and six 3-dimension tables being considered. The list of tables protected using the variety of rules was:

Geog Indicator x Deprivation Decile

Age x Deprivation Decile

Age x Geog Indicator

Sex x Geog Indicator x Deprivation Decile

Sex x Age x Deprivation Decile

Sex x Age x Geog Indicator

Diabetes/No Diabetes x Geog Indicator x Deprivation Decile

Diabetes/No Diabetes x Age x Deprivation Decile

Diabetes/No Diabetes x Geog Indicator

Furthermore it was decided that it was possible to collapse Geographic Indicator, Age and Deprivation Decile into subgroups to allow for table redesign to be carried out. Therefore every table has been subject to some form of table redesign to attempt to remove the disclosure or at least to aid the further disclosure methods. Table redesign has successfully made the following tables safe from potential disclosure:

Age x Geog Indicator (Redesigned)

Age (Redesigned) x Geog Indicator (Redesigned)

Sex x Age x Geog Indicator (Redesigned)

Sex x Age (Redesigned) x Geog Indicator (Redesigned)

Outcome x Age (Redesigned) x Geog Indicator (Redesigned) (Minimum Frequency 5 and 10 only)

A technique using information loss in the table to evaluate the performance of the disclosure control method selected has been described in Quantifying Information Loss (section 2.3). This technique is simple and gives a quick comparison of the disclosure control method without the need for further analysis. It is of primary use when comparing different forms of the same method of disclosure control. For example the information loss technique is a reasonable way to compare the effect of using the optimal cell suppression technique as opposed to using the hypercube cell suppression technique or to compare the effect of using 5 as the rounding base as opposed to using 3 as the rounding base. However it may be argued that the performance of the disclosure control technique can only properly be investigated by quantifying the effect that the disclosure control has on further analysis carried out on the table i.e. how does the further analysis differ when using the disclosure controlled table as opposed to the raw table. Therefore it is sensible to carry out some further analysis on the tables to compare the performance of the different techniques. This also allows comparisons to be made between the different disclosure control techniques. This would have been problematic if the information loss technique had been used since the method for quantifying information loss was different for the different techniques meaning that it may have been possible to attribute any difference in information loss to the differing quantification techniques as opposed to a difference in the effectiveness of the techniques.

The further analysis used to compare the different disclosure control methods can take many forms, in fact any reasonable form of further analysis is acceptable. The further analysis that will be used in this thesis involves constructing a log linear model from each table. The model was constructed from the actual table and then for all the tables produced by the various disclosure control methods. It was decided that a reasonable barometer of how the disclosure control method has performed is how close the residual deviance from the log linear model produced using the disclosure controlled table is to the residual deviance from the log linear model produced using the actual table. To produce a value for this the absolute value of the difference between the residual deviances of the models from the two tables was calculated and this was divided by the residual deviance from the model from the actual table and multiplied by one hundred thus giving a percentage difference between the residual deviance from a model from the actual table and the residual deviance from a model from the disclosure controlled table. It is assumed that all disclosure control methods have made the respective tables safe from potential disclosure and that the concern is how different the further analysis is using the disclosure controlled table instead of the actual table rather than if there is any risk of disclosure.

The results, and the analysis of these results, produced by different choices in the statistical disclosure control process will be investigated and reported under six subsections:

4.4.1 Choice of Minimum Safe Frequency

4.4.2 Choice of Rounding Base

4.4.3 Choice of Secondary Suppression Technique in All Tables

4.4.4 Choice of Secondary Suppression Technique in Tables Containing a Hienrchical Variable

4.4.5 Choice between Rounding and Cell Suppression

4.4.6 Choice between Rounding and Barnardisation

In al tables the information loss quoted in the cells is calculated as the average, over eachtable of the specified design, of the percentage difference between the residual deviance from the log linear model from the disclosure controlled table and the residual deviance from the log linear model from the actual table.

4.4.1 Choice of Minimum Safe Frequency

As mentioned above it is clear that the choice of minimum safe frequency will have an effect on the final disclosure controlled table. It would seem reasonable to assume that as the minimum safe frequency increases the percentage difference in residual deviances from the log linear models will increase. The effect of the choice of mininum safe frequency is only really prominent when cell suppression is chosen as the method of disclosure control. This is the case since when rounding is chosen the mininum frequency rule merely flags a table as unsafe if one of the cells is deemed unsafe and then plays no further role in the disclosure control process whereas when cell suppression is chosen the minimum frequency rule determines how many cells will be primary suppressed clearly effecting the make up of the disclosure controlled table. Therefore the effect of the minimum safe frequency can only be properly investigated by comparing the cell suppression methods for different minimum safe

frequencies. The minimum safe frequencies that were applied to the tables from these data were 3 (the default in Tau-Argus), 5 (a very common choice) and 10 (a very cautious choice). *Table 4.3.1* and *Table 4.3.2* show the average, over each table of the specified design, of the percentage difference between the residual deviance from the log linear model from the disclosure controlled table and the residual deviance from the log linear model from the actual table for two different cell suppression methods for the three different minimum frequencies.

Design / Min Freq	2-dim	2-dim(1 var redesign)	2-dim(2 var redesign)	3-dim	3-dim(1 var redesign)	3-dim(2 var redesign)	All
3	11.62	13.53	11.26	12.44	11.29	12.85	12.11
5	14.35	17.56	11.77	13.45	14.38	13.93	14.49
10	16.02	20.83	22.80	17.09	19.32	15.97	18.52

Table 4.4.1: Percentage difference in residuals deviances between actual and disclosure controlled tables for separate minimum frequencies and table designs. Optimal secondary suppression used.

Design / Min Freq	2-dim	2-dim(1 var redesign)	2-dim(2 var redesign)	3-dim	3-dim(1 var redesign)	3-dim(2 var redesign)	All
3	10.85	14.39	23.45	19.52	16.69	16.80	16.75
5	14.20	18.75	23.45	21.80	23.35	17.36	20.66
10	17.64	20.21	22.80	27.03	24.71	16.70	22.41

Table 4.4.2: Percentage difference in residuals deviances between actual and disclosure controlled tables for separate minimum frequencies and table designs. Hypercube secondary suppression used.

Table 4.3.1 and 4.3.2 appear to confirm the assumption that as the minimum safe frequency increases the utility of the table in the further analysis decreases using this data. This is the case for all the table designs for both secondary suppression techniques except for the case when hypercube secondary suppression is used on the 3-dimension tables with 2 variables redesigned where when 10 is used as the minimum safe frequency the average percentage difference between the two residual deviances is lower than when 3 or 5 is the minimum safe frequency. The results are however very close to what is expected and illustrates nicely the trade off to be made between data utility and potential disclosure.

4.4.2 Choice of Rounding Base

It is also intuitively obvious that if rounding is chosen as the disclosure control method that adjusting the rounding base will result in a change in the final disclosure controlled table. It would seem clear that the pattern should follow that of changing the minimum safe frequency and that as the rounding base increases the difference in residual deviance between the model from the disclosure controlled table and the model from the actual table will get larger. This makes intuitive sense since the larger the base the further away the rounded value is likely to be from the value in the actual table. The rounding bases applied to the tables from this data set were 3, 5 and 10. These choices give a reasonable spread of potential options since 3 is as small a rounding base as would ever be reasonable and 10 is a large rounding base, especially for this data. *Table 4.3.3* shows the average, over each table of the specified design, of the percentage difference between the residual deviance from the log linear model from the disclosure controlled table and the residual deviance from the log linear

model from the actual table for tables produced using all three rounding bases (the rows).

Design / Base	2-dim	2-dim(1 var redesign)	2-dim(2 var redesign)	3-dim	3-dim(1 var redesign)	3-dim(2 var redesign)	All
3	0.001	0.004	0.003	0.019	0.005	0.006	0.007
5	0.027	0.013	0.016	0.029	0.010	0.003	0.015
10	0.057	0.057	0.252	0.057	0.038	0.096	0.069

Table 4.4.3: Percentage difference in residuals deviances between actual and disclosure controlled tables for separate rounding bases and table designs.

Table 4.3.3 shows that, as expected, as the rounding base increases the effectiveness of the table in further analysis appears to decrease when using this data. Once again this is the case for all designs of tables except for where rounding base 5 gives a slightly more accurate residual deviance than rounding base 3 in the 3-dimension tables with two variables redesigned on this data set but the difference is extremely small (0.003%). Once again these results illustrate the trade off between data utility and potential disclosure however it is also interesting to note that that the percentages are extremely small with the highest average percentage difference being 0.252% and even this is for a 2-dimension table with two variables redesigned and a rounding base of 10 which is a scenario which is unlikely to occur in practice.

4.4.3 Choice of Secondary Suppression Technique in All Tables

When cell suppression is chosen as the disclosure control method there is a less intuitive decision for the statistician to make with regards the secondary suppression

method. This is due to the fact that there is a selection of secondary suppression methods which can be used in a variety of scenarios. Four different secondary suppression methods are described in Section 2.2.2 however only two of the methods (hypercube and optimal) are applicable to all of the tables produced from this data set whilst one of the methods (modular) is only applicable to the tables containing Variable 3 since it requires the table to contain a hierarchical variable. The optimal secondary suppression technique should, as the name would suggest, give the most effective solution to the suppression problem. However it is computationally complex and for large tables an attempt to find an optimal solution may be time consuming and at times near impossible. In these situations a time limit is set and Tau-Argus will compute a near optimal solution in the specified time frame. Fortunately for this data there was only one occasion (Sex x Deprivation Decile x Geog Indicator) where the optimal solution was not found in 3 minutes and a near optimal solution was used. It would also be hoped that the modular secondary suppression method would be effective on the tables which contain a hierarchical variable since it has been designed specifically for this scenario. The hypercube secondary suppression method requires relatively little computing power compared to the optimal and modular methods making it quick to run whilst still being reasonably effective. There are two potentially interesting comparisons that are of primary interest here. Firstly the comparison between the optimal and hypercube secondary suppression methods to investigate whether the increased computing power required for the optimal method results in a significant improvement in the effectiveness of the final table in further analysis. Secondly the comparison between the modular method and the hypercube and optimal methods on the tables containing a hierarchical variable to investigate

whether the case specific method is more effective than the standard methods in producing tables which are more useful in further analysis.

Table 4.3.4 shows the average, over each table of the specified design, of the percentage difference between the residual deviance from the log linear model from the disclosure controlled table and the residual deviance from the log linear model from the actual table for tables produced using both the optimal and hypercube secondary suppression methods for the three chosen minimum frequencies.

Design / Supp Method	Min Freq	2-dim	2-dim(1 var redesign)	2-dim(2 var redesign)	3-dim	3-dim(1 var redesign)	3-dim(2 var redesign)	Total
O	3	11.62	13.53	11.26	12.44	11.29	12.85	12.11
H	3	10.85	14.39	23.45	19.52	16.69	16.80	16.75
O	5	14.35	17.56	11.77	13.45	14.38	13.93	14.49
H	5	14.20	18.75	23.45	21.80	23.35	17.36	20.66
O	10	16.02	20.83	22.80	17.09	19.32	15.97	18.52
H	10	17.64	20.21	22.80	27.03	24.71	16.70	22.41

Table 4.4.4: Percentage difference in residuals deviances between actual and disclosure controlled tables for separate secondary suppression methods, minimum frequencies and table designs. O-Optimal Method H-Hypercube Method.

It would appear that, given the values in *table 4.3.4,* as expected the optimal secondary suppression method does in general produce a more effective table with which to carry out further analysis than the hypercube secondary suppression method for this data set. This can be seen by the fact that almost every set of tables produced

by the optimal method give a lower percentage difference between the residual deviance from the actual table and the disclosure controlled table than when the hypercube method is used. It is interesting to note that in the 3-dimension tables there seems to be a larger difference between the optimal and hypercube methods and in fact in three cases in the 2-dimension tables (2-dimension no redesign min freq 3, 2-dimension no redesign min freq 5, 2-dimension one variable redesign min freq 10) the hypercube method does on average produce tables from which the residual deviance of the log linear model are closer to that given by the actual table than the optimal method whereas in the 3-dimension tables there are no cases where the hypercube method performs more effectively than the optimal method. It would appear, from the results from these data, that when time constraints and computing power allow the optimal method should be preferred over the hypercube method as the secondary suppression method of choice in all cases. The difference appears to have a more significant effect in the more complex higher level tables.

4.4.4 Choice of Secondary Suppression Technique in Tables Containing a Hierarchical Variable

Table 4.3.5 again shows the average, over each table of the specified design, of the percentage difference between the residual deviance from the log linear model from the disclosure controlled table and the residual deviance from the log linear model from the actual table for tables containing variable 3 (the hierarchical variable) produced using the three secondary suppression methods and using the three different minimum frequencies.

Design / Supp Method	Min Freq	2-dim	2-dim(1 var redesign)	2-dim(2 var redesign)	3-dim	3-dim(1 var redesign)	3-dim(2 var redesign)	Total
O	3	14.15	16.13	4.51	13.27	12.76	12.00	13.03
H	3	12.86	17.45	28.88	23.44	21.58	18.87	20.50
M	3	13.68	16.84	4.51	15.52	13.15	12.00	13.71
O	5	15.83	18.56	4.51	14.52	16.16	12.00	15.12
H	5	14.50	23.12	28.88	26.22	29.60	18.87	25.11
M	5	15.23	18.43	4.51	17.80	16.33	12.00	15.78
O	10	18.16	23.67	26.56	19.17	21.80	16.03	20.56
H	10	19.48	23.86	26.56	32.59	30.76	15.91	26.53
M	10	17.98	23.13	26.56	20.40	21.67	16.03	20.67

Table 4.4.5: Percentage difference in residuals deviances between actual and disclosure controlled tables for separate secondary suppression methods, minimum frequencies and table designs. O-Optimal method H-Hypercube method M-Modular method.

One of the interesting comparisons in *table 4.3.5* is between the percentage difference of the residual deviances from the log linear model from the actual table and from the disclosure controlled table using the modular method and the optimal method at the various levels. This comparison is of interest since, as noted above, the optimal method appears to be more effective than the hypercube method in retaining the effectiveness of the table for further analysis in all cases whilst the modular method is used specifically for dealing with hierarchical variables. However a potential problem of analysing this table in too much detail is that some of the cells contain very few observations. This is due to the fact that there is only one hierarchical variable in the

data set and there is a limit to the number of tables that contain potential disclosure that can be created. This may result in some slightly skewed data, however it is hoped that any trends established would be replicated by the use of more tables.

The first observation to make about table *4.3.5* is that in most cases the optimal method and the modular method produce tables which are far more effective in the further analysis than the hypercube method on these data. The only exception to this is in the 2-dimension tables with no variable redesign where the hypercube method produces tables more effective for the further analysis when the minimum safe frequency is three and five and the effectiveness of the tables produced by the three methods is similar when the minimum safe frequency is ten. This observation is not unexpected and leads to the more interesting comparison between the optimal secondary suppression method and the modular secondary suppression method. There appears to be very little to choose between the two methods with regards to the percentage difference between the residual deviances of the actual model and the disclosure controlled model. The total average does suggest that over all types of model the optimal method may provide on average slightly better tables from which to carry out further analysis however in many of the cells the values are the same suggesting the same tables were produced by both methods. It may be suggested that using modular secondary suppression may be a worthwhile alternative to using the optimal method since it is less time consuming and requires less computing power. However to draw any real conclusions on this more tables would certainly have to be investigated.

4.4.5 Choice between Rounding and Cell Suppression

The decision made by the statistician that has the most affect on the make up of the table is the choice of disclosure control methods. This decision will affect both the aesthetic qualities of the table and the ability for a user to carry out further analysis using the table. Tau-Argus offers three techniques for removing potential disclosure in the tables. These are rounding, cell suppression and table redesign. As has been seen earlier, table redesign can be combined with either of the other two techniques. The decision about whether table redesign is a valid disclosure control technique to use in each individual situation is more often governed by issues other than the statistical implications of a redesigned table. These issues include whether there are groupings which the separate categories of the variables can be collapsed into and whether the end user has use for tables which have been redesigned or if they require the table with the original categories. It is generally believed however that if table redesign is reasonable and doing so protects the table and retains the utility for the user this technique should be applied. These issues can also have a bearing on whether the rounding or cell suppression method is used since the end user may prefer a table which has been rounded rather than one which has had cells suppressed and vice versa. However it is useful to be able to investigate the statistical implications of each method and how they affect the effectiveness of the tables in terms of further analysis. The table (*Table 4.3.5*) shows the average, over each table of the specified design, of the percentage difference between the residual deviance from the log linear model from the disclosure controlled table and the residual deviance from the log linear model for tables which have been both optimal suppressed and controlled rounded. Optimal suppression was chosen as the secondary suppression technique for

the cell suppression method since it has been shown earlier that in almost all cases it is at least equally as an effective method for allowing accurate further analysis as either of the other secondary suppression techniques.

Design / Method	Min Freq /Base	2-dim	2-dim(1 var redesign)	2-dim(2 var redesign)	3-dim	3-dim(1 var redesign)	3-dim(2 var redesign)	Total
O	3	11.62	13.53	11.26	12.44	11.29	12.85	12.11
R	3	0.001	0.004	0.003	0.019	0.005	0.006	0.007
O	5	14.35	17.56	11.77	13.45	14.38	13.93	14.49
R	5	0.027	0.013	0.016	0.029	0.010	0.003	0.015
O	10	16.02	20.83	22.80	17.09	19.32	15.97	18.52
R	10	0.057	0.057	0.252	0.057	0.038	0.096	0.069

Table 4.4.6: Percentage difference in residuals deviances between actual and disclosure controlled tables for separate suppression methods and table designs. O-Optimal secondary suppression method R-Controlled rounding.

The main observation that can be made from *table 4.3.6* is that for all the specific table designs and minimum frequencies/rounding bases in this data set the effectiveness of the rounded tables in further analysis is much greater than for tables with cells suppressed. The difference between the two methods is so large that it is fair to say that controlled rounding is a more effective technique than cell suppression for controlling disclosure in tables in terms of retaining the usefulness of the table for further analysis. However, as mentioned earlier, this does not necessarily mean that controlled rounding should always be used on this type of data. There may be occasions where the user can get the information they need from a table containing

suppressed cells which would not have been possible had all the cells in the table been rounded. It does appear from *table 4.3.6* that if further analysis is to be carried out on the table then using controlled rounding provides an effective technique for retaining the usefulness of the data.

4.4.6 Choice between Rounding and Barnardisation

It was mentioned in section 4.2 that ISD currently use a technique known as barnardisation to protect tables from potential disclosure. This technique is not available on Tau-Argus however the tables used in these comparisons were protected using a barnardisation program written on another package to allow a comparison between the current method of ISD with potential new methods such as rounding and cell suppression. Due to the nature of the barnardisation technique (i.e. it is a perturbative method) it seems reasonable to compare the results from the tables produced using barnardisation to those from the tables produced using rounding. *Table 4.3.5* shows the average, over each table of the specified design, of the percentage difference between the residual deviance from the log linear model from the disclosure controlled table and the residual deviance from the log linear model from the tables which have been rounded using the three different bases and the tables produced by barnardisation.

Design / Base	2-dim	2-dim(1 var redesign)	2-dim(2 var redesign)	3-dim	3-dim(1 var redesign)	3-dim(2 var redesign)	All
3	0.001	0.004	0.003	0.019	0.005	0.006	0.007
5	0.027	0.013	0.016	0.029	0.010	0.003	0.015
10	0.057	0.057	0.252	0.057	0.038	0.096	0.069
B	0.004	0.004	0.013	0.009	0.005	0.007	0.006

Table 4.4.7: Percentage difference in residuals deviances between actual and disclosure controlled tables for separate rounding bases, the barnardisation technique and table designs. B-Barnardisation.

Table 4.3.7 appears to show that on these data, barnardisation provides tables which are more effective in terms of handling further analysis than rounding with a base 5 or 10. There appears to be little difference between the effectiveness of barnardisation and rounding to base 3. It appears from these results that barnardisation is a good choice when it comes to removing potential disclosure from a table and in fact in the correct circumstance may provide a useful tool in terms of disclosure control. However it may be that since barnardisation adjusts the marginal values of the tables only slightly that it may be subject to a threat from record linkage where an intruder can use a combination of a number of different tables from the same area and combine the data to pose an identification threat.

4.5 Observations on Comparisons of Different Methods

It is clear, from chapter 2, that there are many different methods which can be used to detect potential disclosure and protect against it. The number of potential methods compared in this chapter was restricted due to two factors, namely the data provided

and the limitations of the Tau-Argus package. The data provided by ISD could only be used to produce frequency tables. This is the case with the majority of the work carried out by ISD so the need to investigate the effect of the various methods on this style of table was much greater than an investigation into the efficiency of the methods on magnitude tables. There are numerous methods to detect potential disclosure in magnitude tables whereas there are a rather limited number of methods to detect disclosure in frequency tables. Therefore the lack of magnitude data results in a lack of investigation into differing methods for detecting potential disclosure. Furthermore Tau-Argus clearly has a limit to the number of methods it can carry out. In fact for detecting potential disclosure in frequency tables the only test on offer is the minimum frequency test. Also when protecting the table against the potential disclosure Tau-Argus does not offer any option to add random noise to the data but concentrates on controlled rounding and cell suppression whilst also allowing table redesign to occur. In spite of these restrictions there are plenty of methods available for comparison given the data available and the comparison of these produced some interesting results.

A problem encountered was how best to compare the effectiveness of the different disclosure control methods. Tau-Argus offers a value for information loss in the table summary option. This calculates the information lost as a result of the disclosure control process following the rules given in section 2.3. This representation of the information lost due to the disclosure control process could legitimately be used to compare the effectiveness of different variants on the same method for example when comparing between two separate secondary suppression methods or when comparing two separate rounding bases. However since the information loss for the different

methods is calculated using different techniques these values of information loss cannot be used to compare the different methods therefore another measure of effectiveness of the techniques was required. It was decided to use a measure of how the disclosure controlled table would perform in further analysis compared to the actual table. This was done by calculating the difference between the residual deviances of the models from the two tables and dividing by the residual deviance from the model from the actual table and multiplying by one hundred thus giving a percentage difference between the residual deviance from a model from the actual table and the residual deviance from a model from the disclosure controlled table. This was calculated for a variety of methods to produce results relating to the average effectiveness of the various methods.

Some initial observations from the various disclosure control processes were the speed with which Tau-Argus computed the disclosure controlled tables and the effect of table redesign on some of the tables. Tau-Argus produces the tables from the raw data and then implements the various disclosure control processes almost instantaneously. This is extremely useful if the user wishes to try a selection of disclosure control methods on the table. Also if the user is wishing to redesign the table this process is also made simple by Tau-Argus. Table redesign was implemented on the tables used in this study. It is often the case that if table redesign is applicable to the table and is acceptable to the end user and in doing so the statistician can make the table safe from potential disclosure this method will be implemented since no information is adjusted or suppressed. Four tables from these data were made safe from potential disclosure by using table redesign and a further table was made safe from potential disclosure for minimum frequency 5 and 10.

However if the redesigned table requires further disclosure control methods to make it safe it is often the case that the redesigning of the table provides no real advantage to the disclosure control process.

In section 4.4 different options available to the statistician in the disclosure control process were considered and a comparison made of the options. The options compared were:

1. Choice of minimum safe frequency
2. Choice of rounding base
3. Choice of cell suppression method (across all table)
4. Choice of cell suppression method (across table containing a hierarchical variable)
5. Choice of controlled rounding or cell suppression
6. Choice of barnardisation or controlled rounding

Three separate minimum safe frequencies were compared in the study. These were three, five and ten. It was found that, as expected, in general as the minimum safe frequency increased the percentage difference between the residual deviance from a model from the actual table and the residual deviance from a model from the disclosure controlled table increased implying that the table was less useful in further analysis when the minimum safe frequency was larger. This is unsurprising since as the minimum safe frequency increased either the number of cells which had to be suppressed or the rounding base, depending on the disclosure control method, would increase so the disclosure controlled table would be further from the actual table.

When using rounding as the disclosure control method and the minimum frequency rule to detect potential disclosure the whole table must be rounded if any of the cells is deemed unsafe by the disclosure detection rule. This can sometimes result in large tables with only one safe cell being rounded. Three separate rounding bases for the controlled rounding method were compared in the study. These were three, five and ten. It was also found that, again as expected, in general as the rounding base increased the percentage difference between the residual deviance from a model from the actual table and the residual deviance from a model from the disclosure controlled table increased implying that the table was less useful in further analysis when the rounding base was larger. Again this is unsurprising since a larger rounding base will result in many of the cell values being further away from their value in the actual table than if a smaller rounding base was used. This is not the case for every cell but over a whole table the difference will usually become significant.

There are four methods for secondary suppression available in Tau-Argus. Of these only optimal suppression and hypercube suppression were applicable to all tables in this study whilst modular suppression was only applicable to the tables containing geographical indicator (the only hierarchical variable). It made sense to firstly compare the optimal and hypercube methods across all the tables. Doing this showed that the optimal secondary suppression method appeared to in general give a lower percentage difference between the residual deviance from a model from the actual table and the residual deviance from a model from the disclosure controlled table than the hypercube secondary suppression method over all the tables. This difference appeared to be larger for 3-dimension tables than 2-dimension tables. This result is

not surprising although it should be remembered that the optimal method requires more computing power and in some occasions more time to run than the hypercube method. If the statistician has access to sufficient computing power they should endeavour to use the optimal method over the hypercube method.

The modular secondary suppression method is a technique for dealing specifically with tables which contain a hierarchical variable. Optimal and hypercube secondary suppression methods can also deal with hierarchical variables in the table so it was of interest to compare these techniques with the modular method. The comparison of these methods over the tables with a hierarchical variable showed that both the modular and optimal secondary suppression methods in general gave a lower percentage difference between the residual deviance from a model from the actual table and the residual deviance from a model from the disclosure controlled table than the hypercube secondary suppression method. Furthermore there was very little to choose in terms of percentage difference between the residual deviance from a model from the actual table and the residual deviance from a model from the disclosure controlled table between the modular method and the optimal method. Although there was some indication that on average the optimal method may have given slightly smaller percentage differences than the modular method. It must be remembered however that there were a limited number (20) of tables spread amongst all the separate forms of table and that to confirm these findings further analysis involving more tables would be required.

The most influential decision made by a statistician in the disclosure control process is whether to employ rounding or cell suppression as the disclosure control method.

As mentioned Tau-Argus employs controlled rounding as the rounding option whilst there is a choice between four different secondary cell suppression techniques. It has already been noted that for all tables optimal secondary suppression appears on average to produce more efficient tables than the other secondary suppression methods on this data set therefore it appears sensible to compare rounding with optimal suppression to give an idea of which is the most efficient disclosure control method. The results from this study showed that controlled rounding produced tables with a much lower percentage difference between the residual deviance from a model from the actual table and the residual deviance from a model from the disclosure controlled table than optimal suppression. For example for 3-dimension tables with one variable redesigned and a minimum safe frequency of 5 the average percentage difference using optimal suppression was 14.38 whilst the average percentage difference using controlled rounding was 0.01. This difference is very large and if there are no mitigating circumstances the recommendation would be to use controlled rounding as opposed to cell suppression however, as has already been mentioned there are many cases where the construction of the disclosure control process is dependent on other variables such as the end user's views on the aesthetics of the table and other such concerns. It is though interesting to note the size of the difference between the results collected for the two methods as this may have serious consequences in any disclosure control policy created by a statistical organisation.

As mentioned ISD policy currently involves using a technique called barnardisation. Barnardisation is a perturbative disclosure control method which adds noise (usually +1, 0, -1) to the cell values less than five and adjusts the marginals accordingly. The natural method to compare with barnardisation was controlled rounding as both are

perturbative techniques. It was found that barnardisation on average produced tables with a lower percentage difference between the residual deviance from a model from the actual table and the residual deviance from a model from the disclosure controlled table than controlled rounding with base five or ten and showed very similar results to controlled rounding with base three. This would suggest that barnardisation is a very efficient technique for protecting against potential disclosure although there may be issues with barnardisation in particular with linked tables if an intruder has auxiliary data they may be able to unpick a table protected by barnardisation table with greater ease than a table protected by controlled rounding or cell suppression. Also the choice between disclosure control methods may again be influenced by the needs of the end user.

4.6 Tabular Summary of Comparisons of Techniques on ISD Data

Table 4.6.1 provides a summary of the results of the comparisons of different techniques on this particular ISD data set. It must however be noted that the observations given in *table 4.6.1* are very crude and more detailed information can be found in sections 4.4 and 4.5.

Choice Made	Observations
Choice of Minimum Safe Freqeuncy	As minimum safe frequency increased the utility of the protected table decreased.
Choice of Rounding Base	As rounding base increased the utility of the protected table decreased.
Choice of Seconday Supression Technique (All Tables)	The optimal method produced protected tables with greater utility than the hypercube method.
Choice of Seconday Supression Technique (Hierarchical Tables)	The optimal and modular methods produced protected tables with similar utility and this utility was greater than the protected tables produced by the hypercube method.
Choice between Rounding and Cell Suppression	Controlled rounding produced protected tables with significantly greater utility than any cell suppression method.
Choice between Rounding and Barnardisation	Barnardisation produced protected tables with greater utility than controlled rounding with base 5 or 10 and similar utility to the protected tables produced by controlled rounding with base 3.

Table 4.6.1: Summary of comparisons of techniques on ISD data.

5. Conclusions and Further Study

5.1 Recommendations on Statistical Disclosure Control Social Issues

Statistical disclosure control is an extremely progressive subject. Only in the relatively recent past have statisticians become involved in the debate and development of techniques to deal with the problem. This means that the thoughts on what is regarded as best practice are also evolving rapidly. However also evolving, at an extremely fast rate thanks to the exponential increase in the amount of data available over the World Wide Web, is the ability of potential intruders to acquire confidential information using the data supplied by the statistical organisation along with any auxiliary data they possess. Furthermore the increase in awareness of the ability of intruders to acquire information on subjects has led to an increase in the number of laws pertaining to the confidentiality of subjects. These laws are themselves constantly being reviewed due to the increases in technology and disclosure control methodology. The result of this is that statistical organisations must have in place effective disclosure control policy which can be adapted to both differing data sets and the ever changing face of statistical disclosure control methodology and legislation. In general the problem encountered in statistical disclosure control is to ensure the confidentiality of subjects in a study whilst retaining as much of the data utility as possible.

There are numerous social issues pertaining to the topic of statistical disclosure control. Firstly an important issue is defining what exactly constitutes disclosure. Clearly there are many ways in which a subject can be identified in a study. It may be

that the intruder has knowledge from auxiliary data or may for example have prior knowledge as to the identities of the subjects in a particular study. Furthermore it may be that a person has rare characteristics making them easily identifiable. If disclosure results from the data produced by a statistical organisation in any of the above scenarios it would be regarded as a failing on the part of the disclosure control policy of the statistical organisation. It has also been suggested that a subject being able to identify themselves in a study may technically be termed as disclosure from the data. However most organisations and laws would regard protecting against this level of disclosure unreasonable and would not consider this form of disclosure a failing on the part of the disclosure control policy.

Almost all the data given to statistical organisation is given under the promise of confidentiality of the subject. This is crucial since this promise encourages the subject to give reliable and truthful data about themselves. If the situation arises where the publishing of a statistical study results in disclosure of information about a subject, especially if this disclosure is uncovered or reported by the media, there may develop a lack of trust in not only the erring statistical organisation but in the statistical community in general. This potential lack of trust would most likely result in unreliable and incomplete data being provided by subjects who do not wish to risk any of their private information being disclosed. This issue has been brought into focus with the implementation of the Freedom of Information Act 2000 which came into force at the beginning of 2005 instructing statistical organisations to release information to anyone who requests it unless the request falls into one of the limited number of exempt circumstances. To deal with this statistical organisations must

have in place a stringent disclosure control policy which can be applied to all the data they hold.

There is an issue in statistical disclosure control that involves whether certain variables should be regarded as more sensitive as others and therefore be subject to more stringent disclosure control techniques. It is argued that different disclosure control techniques should be used on the data involving sensitive variables than on the data involving non-sensitive variables since the risk associated to the respective potential disclosures is different. The problem with this is that there are differing opinions on what data should be deemed sensitive and in fact this is often just a personal opinion. This leaves the statistician or their organisation in the unenviable position of having to make a decision on which subjects the public would deem as sensitive. However as a rule the statistician should err on the side of over protecting any information they feel may potentially cause harm or distress to the subject if disclosure occurred.

It is has been proposed that statistical organisations could restrict the access to the data to only a selected group of users. Those given access to the data would be required to sign a confidentiality agreement and then would be given unrestricted access to the raw data. This would appear to be a reasonable way of allowing access to complete data to those who use it for the good of society however it is believed that not releasing data builds up a mystique around the statistical community which may result in distrust. As has already been noted any lack of trust in the statistical community may lead to poor quality data being supplied by subjects. Furthermore it is believed that subjects are more likely to make an effort to take part in studies if they

can view the results and understand how the use of their time and information has resulted in a collection of data that is of use to both them and society in general. For these reasons most statisticians believe that statistical organisations should endeavour to, where at all possible, release the results they collect to the general public in some disclosure controlled form rather than restrict access to the data.

5.2 Recommendations on Statistical Disclosure Control Methods and Findings

There are many statistical techniques involved in the disclosure control process. These techniques can be split into two categories; techniques for detecting potential disclosure in the data and techniques for removing potential disclosure in the data. Due to the nature of the data that ISD generate it was decided that this thesis would concentrate on the techniques used on tabular data and in particular detecting and removing potential disclosure in an existing table. Furthermore almost all the tables produced by ISD are in the form of frequency tables therefore most of the analysis in this thesis is produced on frequency tables. However chapter 2 gives an overview of techniques for detecting and removing potential disclosure in both magnitude and frequency tables. The techniques reviewed were on the whole standard techniques that have been tried and tested in statistical disclosure control processes in the past. Mention is also made of both the theoretical and practical rationale behind the techniques and the advantages and disadvantages of each method are discussed.

There are a number of computing programs available for carrying out statistical disclosure control on tables. However the major package available in Europe which is being developed by a group working within the EU, CASC (Computational Aspects of Statistical Confidentiality), is Tau-Argus. The major aim of the program is to give the statistician the ability to carry out a number of disclosure control techniques without the problems caused by complex and time consuming calculations. The programme is evolving rapidly and the aim is to both offer more statistical techniques and make the program more user friendly. The techniques available in Tau-Argus to detect potential disclosure in tables are the minimum frequency rule, the dominance rule and the p% rule. Of these only the minimum frequency rule can be used to detect potential disclosure in a table of frequency data whilst all three techniques can be used to detect potential disclosure in a table of magnitude data. To remove potential disclosure in the tables Tau-Argus offers controlled rounding, table redesign and cell suppression (with four secondary suppression techniques; optimal, hypercube, modular, network). The experience gained from using the programme in this thesis suggests that it was initially difficult run the programme smoothly and some of the workings take time to get used to. However once one has a full understanding of how the programme operates it is extremely simple to carry out complex techniques on some large tables without any problems. Another advantage of the programme is that (on the tables investigated in this thesis) the techniques are carried out extremely quickly allowing the statistician to try a number of different disclosure control techniques without worrying about the issue of time consumption. However it must be remembered that Tau-Argus is not designed to replace the role of the statistician but is simply a time saving tool to aide the statistician in the disclosure control process. It is always prudent for the statistician to do background research and have

an idea of potential disclosure control techniques they wish to implement before they start the process.

In chapter 4 an attempt was made to compare the effectiveness of various disclosure control techniques on a data set of interest to ISD. The data used to compare the various disclosure control techniques were in fact provided by ISD. It came from a study into diabetes in children across the whole of Scotland. Data were collected on 365 diabetic children and 499428 control children. The variables selected were variables which may have had a causal effect on a child's chances of developing diabetes. All the variables were categorical therefore all the tables contain frequency data. It was decided to carry out disclosure control on both 2 and 3 dimensional tables although it would have been possible to carry out the disclosure control procedure on tables with higher dimensions. The data provided produced three tables of 2 dimensions and six tables of 3 dimensions which contained cells with a risk of potential disclosure. Since all the tables were made up from variables which had categories which could be collapsed into larger categories it was decided to carry out redesign on all the tables and follow the redesign of the table with further disclosure control methods if necessary. This provided an opportunity to consider how redesigning the table would affect both further analysis and the role of the further disclosure control techniques.

The statistical disclosure control procedure was carried out using Tau-Argus. The aim was to compare the effect of different disclosure control methods on the information lost in the table when each method was applied to the table. Before proceeding with the comparisons it was important to have a technique to quantify the

loss of information. Tau-Argus provides a tool for quantifying information loss but this was not used since the technique was different for the different methods used for removing potential disclosure in the table. To allow a comparison to be made between different disclosure control methods requires a technique which is constant for all methods. The technique used to allow these comparisons was based on constructing a log linear model for both the actual and disclosure tables and comparing the residual deviances.

There are many choices available to the statistician in the disclosure control process. These choices often have a large bearing on the final outcome of the table and therefore all options should be thoroughly considered before any disclosure control procedure is confirmed. The choices investigated in this thesis were:

1. Choice of minimum safe frequency
2. Choice of rounding base
3. Choice of cell suppression method (across all tables)
4. Choice of cell suppression method (across tables containing a hierarchical variable)
5. Choice of controlled rounding or cell suppression
6. Choice of barnardisation or controlled rounding

The findings are summarised in the remainder of this section.

Three different minimum safe frequencies (3,5 and 10) were compared. As expected as the minimum safe frequency was increased the percentage difference between the

residual deviance from a model of the actual table and the residual deviance from a model of the disclosure controlled table increased implying that the table was less useful in further analysis when the minimum safe frequency was larger. This is unsurprising since as the minimum frequency increased the number of cells below the minimum safe frequency will have increased. Clearly the choice of the minimum safe frequency is initially determined by how large the statistician feels the cell should be to ensure there is a small enough risk of potential disclosure, however these results show that the statistician should endeavour to keep the minimum safe frequency as low as possible.

Controlled rounding is the only rounding option offered by Tau-Argus. This causes no problem since it is generally accepted that controlled rounding is the most efficient rounding technique. Three different rounding bases were compared (3, 5 and 10). As expected as the rounding base was increased the percentage difference between the residual deviance from a model of the actual table and the residual deviance from a model of the disclosure controlled table increased implying that the table was less useful in further analysis when the rounding base was larger. This is unsurprising since when the rounding base is increased there is a chance that the rounded cell value will be further away from the actual value than when a smaller rounding base is used. As with the minimum frequency it is clear that the rounding base should be kept as low as possible whilst providing adequate protection against potential disclosure.

Of the four secondary suppression techniques available in Tau-Argus only two (optimal suppression and hypercube suppression) were applicable to all the tables in this study. Of these two techniques it appeared that the optimal secondary

suppression method gave a lower percentage difference between the residual deviance from a model of the actual table and the residual deviance from a model of the disclosure controlled table than the hypercube secondary suppression method over all the tables and that the difference appeared to be larger for 3-dimension tables than 2-dimension tables. This was not a surprising result but in some cases the optimal method required two or three minutes longer to run than the hypercube method. Clearly this length of time is no problem but if the table became more complex there may become issues with time constraints. However if the statistician has the time and computing power to carry out an optimal secondary suppression procedure they should endeavour to do so.

The modular secondary suppression method is a technique used to deal with the secondary suppression problem in hierarchical tables. It was of interest to compare the modular method with the two universal secondary suppression techniques, (optimal suppression and hypercube suppression) when protecting against potential disclosure in hierarchical tables. It appeared that both the modular suppression method and the optimal suppression method in general gave a lower percentage difference between the residual deviance from a model of the actual table and the residual deviance from a model of the disclosure controlled table than the hypercube secondary suppression method whilst there was very little to choose in terms of percentage difference between the residual deviance from a model of the actual table and the residual deviance from a model of the disclosure controlled table between the modular method and the optimal method. There was some indication that the optimal suppression technique gave slightly smaller percentage differences than the modular suppression technique however there were only 20 tables and the difference in results

from each method was small. It would appear from these results that the modular secondary suppression technique offers little advantage to using the optimal suppression technique. If there are hierarchical variables with more hierarchies or categories the modular suppression method may prove more efficient.

The choice between either controlled rounding or cell suppression is the most influential made in the options given by Tau-Argus since the decision has a large bearing on the final appearance of the table. The comparison made in chapter 4.4 was between controlled rounding and cell suppression with optimal secondary suppression used since it has been shown to be more (or at least equivalently) efficient to the other secondary suppression methods in all the cases. It was shown that controlled rounding produced tables with a much lower percentage difference between the residual deviance from a model of the actual table and the residual deviance from a model of the disclosure controlled table than optimal suppression. The difference between the efficiency of the two methods was so large that the statistician should if possible use controlled rounding as opposed to cell suppression however, there are often other circumstances (such as an end users view on the aesthetics of the table) which may lead to the statistician rejecting the controlled rounding option.

ISD policy currently employs a technique known as barnardisation. Barnardisation is a perturbative technique for removing potential disclosure in a table. Since the technique is perturbative it is sensible to compare this to controlled rounding. It was found that barnardisation on average produced tables with a lower percentage difference between the residual deviance from a model of the actual table and the residual deviance from a model of the disclosure controlled table than controlled

rounding with base five or ten and showed very similar results to controlled rounding with base three. This would suggest that barnardisation is an effective technique for removing potential disclosure in tables. However doubts have been raised as to the effectiveness of barnardisation with regards to protecting data in linked tables and whether the protection offered is sufficient.

From the techniques available in Tau-Argus it would appear that on this data set the tables produced were most efficient in further analysis when the minimum frequency rule with a minimum safe frequency of three was the technique used to detect potential disclosure in the table and controlled rounding with a rounding base of three was the technique used to remove potential disclosure from the table. Using this technique would potentially open the issue of whether three is a large enough minimum safe frequency but in this situation when the nature of the data (i.e. diabetes not being an extra 'sensitive' issue) is considered it is probably fair to say that as long as each cell has three subjects the table has a low enough disclosure risk to publish. Clearly this decision will be made on each case individually and the responsibility will lie with the statistician. It is important to remember that the utility, although extremely, important is not the only issue that governs the required disclosure control technique. It is the responsibility of the statistician to have a dialog with the end user of the table (or if the end user is the general public the statistician must consider what the wishes of the general public would be) to determine what they require from the published table and attempt to construct a disclosure process that allows these requirements to be met.

5.3 Potential for Further Study

This thesis has given a general overview of statistical methods used to prevent potential disclosure in tabular data. The theory behind and the practicalities of a number of disclosure control techniques have been investigated fully and many of these techniques have been investigated in practical application. The thesis has attempted to give a broad overview of various techniques and applications however there are certain areas in which the work can potentially be furthered.

Due to both the nature of the data and time constraints it was only possible to carry out the disclosure control process on a limited number of tables from a single data set. It would be prudent to carry out disclosure control on tables produced from different sets of data. This would both allow more confidence in the results produced and ensure that the trends found in this data set in terms of effectiveness of the various disclosure control techniques are the same as the trends found in other data sets. This may also give a greater insight into the different occasions in which certain disclosure control methods are more efficient. Furthermore it may be interesting to investigate the efficiency of the different disclosure control techniques when linked tables are present. Two tables are linked if they have at least one common explanatory variable and the variable which gives the cell values (e.g. frequency) is common. Linked tables create a whole new dimension for the problem of disclosure control and the process of protecting linked tables against potential disclosure offers further challenges. Tau-Argus does in fact give the option to the user to protect linked tables against potential disclosure and an investigation into this option would be of interest.

This thesis has concentrated on the statistical disclosure control methods with particular application to healthcare (and in particular ISD) scenarios. This resulted in the comparisons of the different disclosure control methods being carried out on real ISD data. These data were of the type usually released by ISD meaning that the data used to carry out the comparisons was solely frequency data. Many of the disclosure control techniques, in particular those used to detect potential disclosure, are only applicable when the table contains magnitude data. It would be interesting to make similar comparisons, as done in chapter 4, but using magnitude data allowing different techniques for detecting potential disclosure to be investigated. It would also be interesting to investigate how the efficiency of the different techniques used for removing potential disclosure differed when the table consisted of magnitude data rather than frequency data.

In this thesis it was decided to use the percentage difference between the residual deviance from a log linear model of the actual table and the residual deviance from a log linear model of the disclosure controlled table to investigate the efficiency of the tables produced using each disclosure control method. The reason that this measure was chosen was that it gave a means to investigate each disclosure control process using the same measurement. There are countless other ways in which the efficiency of the table could have been measured although I believe that the measurement should be associated with the effect the disclosure control method has on the ability of someone not involved in process to carry out reasonable further analysis on the table and produce results similar to those for the actual table. It would be interesting if the conclusions drawn about the efficiency of each method were the same had the efficiency been measured in a different but still sensible way.

Since ISD in general carry out the disclosure process after the tables have been produced this thesis has concentrated on the techniques associated with this form of disclosure control. There are also techniques used to prevent potential disclosure by adjusting the microdata before the tables have been produced. This form of disclosure control includes the use of techniques such as global recoding, local suppression and the post randomisation method. These techniques are very different in theory from the techniques used to protect the table after it has been produced. This may result in large differences in the disclosure controlled tables produced. Mu-Argus, a package produced by the CASC project alongside Tau-Argus, allows these methods to be carried out in a simple computing package. It would be interesting to investigate the effects of using the methods available in Mu-Argus on the disclosure controlled table and comparing the efficiency of the various disclosure control methods available in both Tau-Argus and Mu-Argus.

References

[1] M. Elliot, A. Hundepool, E.S. Nordholt, J. Tambay and T. Wende. (2003). *Glossary on Statistical Disclosure Control.* CASC Project.

[2] *Report of the Task Force on Disclosure.* (1995). Government Statistical Service. Available at http://www.statistics.gov.uk/downloads/theme_other/GSSMethodology_No_04_v2.pdf

[3] G.E. Scherff. (1952). *Die Rechtsgrundlagen der Statistik, Eine international vergleichende Darstellung unter besondere Vorhebung der Volkszählungen und der statistischen Geheimhaltungspflicht.* Baden Wurttenberg, Stuttgart.

[4] *United States of America Privacy Act of 1974.* Available at http://www.epic.org/privacy/laws/privacy_act.html.

[5] *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data.* Available at http://www.oecd.org/document/18/0,2340,en_2649_34255_1815186_1_1_1_1,00.html

[6] *Council of Europe Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data,* Available at http://conventions.coe.int/Treaty/EN/Treaties/Html/108.htm

[7] Sharpe Research Ltd. (2004). *Public Attitudes to the Development of Surveillance Techniques in Public Places.* Information Commissioner's Office.

[8] Holvast, J. (1999). *Statistical Confidentiality at the European Level.* Proceedings of the Eurostat Work Session on Statistical Data Confidentiality, Greece.

[9] Cox, L. H. (1992). *Solving Confidentiality Protection Problems in Tabulations Using Network Optimization: A Network Model for Cell Suppression in U.S. Economic Censuses.* Proceedings of the International Seminar on Statistical Confidentiality, Dublin.

[10] *Data Protection Act 1998.* Available at http://www.opsi.gov.uk/acts/acts1998/19980029.htm

[11] Hundepool, A. and Willenborg, L. (1999). *Argus: Software from the SDC Project.* Proceedings of the Eurostat Work Session on Statistical Data Confidentiality, Greece.

[12] Fienberg, S. E. (2000). *Confidentiality and Data Protection Through Disclosure Limitation: Evolving Principles and Technical Advice.* The Philippine Statistician.

[13] Cox, L. H. (2003). *Overview of Statistical Disclosure Limitation.* DIMACS Working Group on Privacy/Confidentiality of Health Data.

[14] Als, G. (1993). *Organization of Statistics in the Member Countries of the European Community, Volume I: Essays on the 12 statistical institutes, Comparative study.* Luxembourg: Office for Official Publications of the European Communities.

[15] Freedom of Information Act 2000.
Available at
http://www.opsi.gov.uk/acts/acts2000/20000036.htm

[16] Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice.* Springer-Verlag, New York.

[17] Gates, G. (2005). *Public Perceptions of Risks to Privacy.* Proceedings of the International Symposium on Confidentiality, Privacy and Disclosure in the 21st century, Manchester.
Available at
http://www.ccsr.ac.uk/capri/symposium/documents/Gates.ppt

[18] Sasse, A. (2005). *Effective Security and Privacy - Proctecting People Not Just Data.* Proceedings of the International Symposium on Confidentiality, Privacy and Disclosure in the 21st century, Manchester.
Available at
http://www.ccsr.ac.uk/capri/symposium/documents/Sasse.ppt

[19] Nordholt, E. S. (2002). *Applications of Statistical Disclosure Control at Statistics Netherlands.* Lecture Notes In Computer Science; Vol. 2316: Inference Control in Statistical Databases, From Theory to Practice. Springer-Verlag, London.

[20] Australian Census and Statistics Act 1905.
Available at
http://scaleplus.law.gov.au/html/pasteact/1/580/top.htm

[21] Australian Statistics Determination 1983 – List of Regulations.
Available at http://scaletext.law.gov.au/html/pastereg/0/414/top.htm

[22] Website of the U.S. Census Bureau.
Available at
http://www.census.gov/privacy/files/data_protection/002825.html

[23] Privacy Principles of the U.S. Census Bureau.
Available at
http://www.census.gov/privacy/files/data_protection/002822.html

[24] Zayatz, L., Hawala, S. and Rowland, S. (2003). *American Factfinder: Disclosure Limitation for Census 2000 Tabular Data.* Paper compiled for Department of Commerce, Bureau of the Census, Statistical Research Division.

[25] Camden, M., Daish, K. and Krsinich, F. (2003). *The Noise Method for Tables – Research and Applications at Statistics New Zealand.* Proceedings of the Joint ECE/Eurostat work session on statistical data confidentiality, Luxembourg.

[26] National Statistics Code of Practice: Protocol on Release Practices.
Available at
http://www.scotland.gov.uk/stats/bulletins/national_stats.pdf

[27] National Statistics Code of Practice: Statement of Principles.
Available at
http://www.statistics.gov.uk/about/national_statistics/cop/downloads/StatementRD.pdf

[28] Website of CASC Group.
Available at
http://neon.vb.cbs.nl/casc/

[29] EPROS European Plan for Research in Official Statistics. September 1999.
Available at
http://europa.eu.int/en/comm/eurostat/research/fp5/documents/eprosen1.pdf

[30] Giessing, S. and Hundepool, A. (2001). *The CASC Project: Integrating Best Practice Methods for Statistical Confidentiality.* [presented at] New Techniques and Technologies for Statistics: Exchange of Technology and Know-how.

[31] Hundepool, A. (2002). *The CASC Project.* Lecture Notes In Computer Science; Vol. 2316: Inference Control in Statistical Databases, From Theory to Practice. Springer-Verlag, London.

[32] Website of ISD (Scotland).
Available at http://www.isdscotland.org/

[33] Personal Health Information in ISD Scotland
Available at
http://www.isdscotland.org/isd/files/InfoGuideForPatients.pdf

[34] Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control.* Springer-Verlag, New York.

[35] Hundepool, A. et al. (2004). *Tau-Argus Version 3.0 User Manual.* CASC Project.
Available at
http://neon.vb.cbs.nl/casc/

[36] Loeve, J. A. (2001). *Notes on sensitivity measures and protection levels.* Research paper

no. 0129. Methods and Informatics Department, Statistics Netherlands, Voorburg.

[37] Nordholt, E. S. (2003). *Applications of Statistical Disclosure Control Methods.* Proceedings of the Statistics Canada International Symposium.

[38] Westlake, A. (2003). *Security and Disclosure for Statistical Information.* Survey and Statistical Computing, London.

[39] Cuppen, M. and Willenborg, L. (2000). *Source Data Perturbation in Statistical Disclosure Control.*
Available from CASC website at
http://neon.vb.cbs.nl/casc/RelatedPapers.html

[40] Lowthian, P. and Merola, G. (2004). *The Application of Controlled Rounding for Tabular Data with Particular Reference to the Tau-Argus Software.* Proceedings of the Methods for Statistics for UK Countries and Regions Conference, London.

[41] Evans, T., Zayatz, L. and Slanta, J. (1998). *Using Noise for Disclosure Limitation of Establishment Tabular Data.* Journal of Official Statistics, 14(4):537—551.

[42] Repsilber R. D. (1994). *Preservation of Confidentiality in Aggregated data.* paper presented at the Second International Seminar on Statistical Confidentiality, Luxembourg, 1994.

[43] Massel, P. (2002). *Optimization Models and Programs for Cell Suppression in Statistical Tables.* Discussion paper produced by U.S. Census Bureau.
Available at
http://www.census.gov/srd/sdc/Massell.JSM2002.v4.pdf

[44] Fischetti, M. and Salazar Gonzalez, J .J. (2000). *Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints.* JASA, v.95.

[45] de Wolf, P. P. (2002). *HiTaS: a heuristic approach to cell suppression in hierarchical tables.* Inference Control in Statistical Databases, Springer-Verlag, Heidelberg.

[46] de Wolf, P. P. (1999). *A heuristic approach to cell-suppression in hierarchical tables.*
Paper produced for the CASC project.
Available at
http://neon.vb.cbs.nl/casc/Related/99wol-heu-r.pdf

[47] Domingo-Ferrer, J. and Torra, V. (2004). *Privacy in Statistical Databases.* Lecture Notes in Computing Science; 3050. Vol. 3050., Springer, London.