TRANSACTIONS ON DATA PRIVACY 5 (2012) 403–432

Comparison of Remote Analysis with Statistical Disclosure Control for Protecting the Confidentiality of Business Data

Christine M. O'Keefe*, **Natalie Shlomo****

*CSIRO Mathematics, Informatics and Statistics, GPO Box 664, Canberra ACT 2601 AUSTRALIA

**Southampton Statistical Sciences Research Institute, University of Southampton, Southampton SO17 1BJ UK.

E-mail: Christine.OKeefe@csiro.au, N.Shlomo@soton.ac.uk

Abstract. This paper is concerned with the challenge of allowing statistical analysis of confidential business data while maintaining confidentiality. The most widely-used approach to date is statistical disclosure control, which involves modifying or confidentialising data before releasing it to users. Newer proposed approaches include the release of multiply imputed synthetic data in place of the original data, and the use of a remote analysis system enabling users to submit statistical queries and receive output without direct access to data. Most implementations of statistical disclosure control methods to date involve census or survey microdata on individual persons, because existing methods are generally acknowledged to provide inadequate confidentiality protection to business (or enterprise) data.

In this paper we seek to compare the statistical disclosure control approach with the remote analysis approach, in the context of protecting the confidentiality of business data in statistical analysis. We provide an example which enables a side-by-side comparison of the outputs of exploratory data analysis and linear regression analysis conducted on a sample business dataset under these two approaches, and provide traditional unconfidentialised results as a standard for comparison. There are certainly advantages and disadvantages in the remote analysis approach and it is unlikely that remote analysis will replace statistical disclosure control methods in all applications. If the disadvantages are judged too serious in a given situation, the analyst may have to seek access to the unconfidentialised dataset. However, our example supports the conclusion that the advantages may outweigh the disadvantages in some cases, including for some analyses of unconfidentialised business data, provided the analyst is aware of the output confidentialisation methods and their potential impact.

Keywords. Confidentialised output, Output checking, Noise addition, Attribute disclosure, Data utility

1 Introduction

This paper addresses the challenge of balancing the competing objectives of allowing statistical analysis of confidential data and maintaining confidentiality.

In addition to restricting access to confidential data, custodian agencies often release less than the full original dataset or alter the data before release to analysts, in order to provide enhanced confidentiality protection. First, identifying attributes such as name and

address are usually removed, as well as other sensitive attributes or observations. Often, this is followed by the application of *statistical disclosure control* methods such as aggregation of geographic classifications, rounding, swapping or deleting values, and adding random noise to data. In this case, the data are first confidentialised then analysed, as shown diagrammatically in Figure 1. We will call this the *confidentialised input* approach.

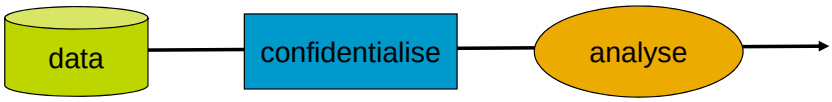


Figure 1: *Confidentialised input:* statistical disclosure control approach of confidentialising data before release for analysis

Unfortunately, statistical disclosure control methods result in information loss and/or biased estimation, and it can be extremely difficult to quantify the level of protection achieved. For more information on statistical disclosure control methods, see for example [1, 5, 6, 7, 8, 17, 30].

Motivated by the drawbacks associated with statistical disclosure control, Rubin [26] suggested the alternative of generating and releasing *synthetic data*, see also [14, 24]. In this approach, the data custodian fits a model to the original data then repeatedly draws from the model to generate multiple synthetic datasets which are released for analysis.

A *remote analysis* system accepts a query from an analyst, runs it on data held in a secure environment, then returns results to the analyst. In particular, the analyst does not have direct access to the data at all. In designing a remote analysis system to deliver useful results with acceptably low risk of a confidentiality breach, restrictions can be imposed on the queries, the analysis itself can be modified and the results can be modified. In this approach, the data are first analysed and only the results are confidentialised, as is shown diagrammatically in Figure 2. We will call this the *confidentialised output* approach.

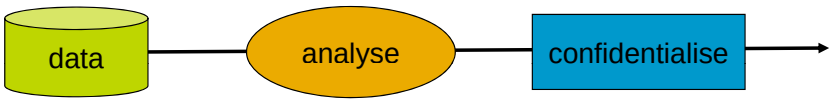


Figure 2: *Confidentialised output:* remote analysis system approach of analysing data then confidentialising only the analysis outputs

A system could combine the confidentialised input and confidentialised output approaches, by implementing a remote analysis system on confidentialised underlying data, see Section 2.1 for an example implemented by the Australian Bureau of Statistics.

A remote analysis system could be fully automated, or could involve some manual checking of queries or outputs, for example. In fact, many of the methods are similar to those implemented in national statistical agencies for manual checking of data laboratory outputs, see for example [25]. For more information on remote analysis, see for example [11, 22, 27].

A 2006 survey of OECD countries [2] found that most of the countries' statistical offices are using statistical disclosure control to make microdata available, and a number of countries provide or are experimenting with remote access and/or remote analysis solutions. (Here, remote access refers to secure direct access to data through an online interface, not through

an analysis server.) Most of these implementations were found to involve household or individual microdata. In the case of business data, there have been some implementations of the synthetic data approach, see for example [9, 13]. The authors conclude that remote access to microdata which are held in national or international institutions " ... may allow for the full exploitation of microdata analysis while limiting the risk of disclosing confidential information."

To date, the evidence supports the conclusion in the case of microdata about individual persons, see [11, 18, 20, 27].

1.1 In this paper

It is the purpose of this paper to contribute to the examination of the potential use of remote analysis in addressing the balance between microdata access and confidentiality protection, in the context of business data. The evidence is found to be promising and merits further investigation.

In Section 2 we discuss the characteristics of business data, and provide a discussion of current practice for enabling statistical analysis of business microdata in some national statistical agencies. Section 3 provides details of an example business dataset and two approaches for enabling statistical analysis without revealing commercially sensitive information. In Sections 4, 5 and 6 we provide a comprehensive example of these two approaches, with reference to the unconfidentialised approach, for univariate exploratory data analysis, bivariate exploratory data analysis and linear regression, respectively. The example enables a side-by-side comparison of analysis results generated under each of the three approaches. Section 7 gives a discussion and conclusion.

2 Business data

In general, business survey data are different in nature from household or individual-level survey data. Business data often exhibit some or all of the following characteristics, which highlight this difference from household or individual data.

1. Business survey data exhibit a characteristic pattern in inclusion probabilities, which heighten the confidentiality issues for large businesses:
   (a) Large businesses are always sampled. Thus business data often include a census of large businesses.
   (b) Medium-sized businesses are frequently sampled.
   (c) Small businesses are seldom sampled.
2. There are generally few variables.
3. Most variables are continuous rather than discrete.
4. The distributions of many variables are highly skewed.
5. Business data commonly include enterprises which are outliers on each of many variables. These are the large businesses in the industry sector or sectors sampled.

Business survey data can be highly commercially sensitive, and businesses have a keen interest in the confidentiality of their data. Data custodian agencies therefore have a responsibility to protect the confidentiality of business data as well as household and individual data, although for different reasons.

Interestingly, the particular characteristics of business data mean that it is more challenging to protect business data confidentiality. A 2006 survey of OECD countries [2] found that "Only a limited number of countries permit some form of access to business microdata; illustrating the practical difficulties inherent in preserving confidentiality of individual businesses." The survey analysts also found this to be particularly true in smaller economies where large businesses are more prominent. They conclude that: "The increased difficulty and the risks associated with disclosure of business microdata have so far stopped some countries from moving forward in this domain."

2.1 Examples of current practice for enabling statistical analysis of business microdata

In this section we provide an overview of the approaches currently used by the Australian Bureau of Statistics, the United Kingdom Office For National Statistics and the United States Census Bureau for enabling research analysis of business microdata.

Australian Bureau of Statistics

The Australian Bureau of Statistics (ABS) releases business microdata as Confidentialised Unit Record Files (CURFs) on CD-ROM, as well as through its Remote Access Data Laboratory and its On-site Data Laboratory, see [3].

The ABS CURFs contain data from ABS surveys in the form of unit records, and represent the most detailed statistical information available from the ABS for researchers and analysts to run statistical analyses. The unit record files are confidentialised by removing name and address information, by controlling and limiting the amount of detail available, and by very slightly modifying or deleting data where it is likely to enable identification of individuals or businesses.

Basic level CURFs, which are the least detailed, are available on CD-ROM for use on researchers' own computers. Each CURF is released for an individual's specified statistical purposes and for a stated period to the nominated Responsible Officer and Individual Authorised Users. Both the Responsible Officer and Individual Authorised Users are required to sign and agree to be bound by a legal Undertaking which if breached can result in a fine or imprisonment or both. The Responsible Officer and Individual Authorised Users consequently have an obligation to ensure that the CURF and any copies of the CD-ROM remain secure. More detailed Expanded CURFs may only be accessed via the ABS Remote Access Data Laboratory (RADL) and the most detailed Specialist CURFS may be accessed through the ABS Data Laboratory (ABSDL). The Remote Access Data Laboratory (RADL) is a secure online data query service that clients can access via the ABS web site, and the ABS Data Laboratory (ABSDL) is an on-site facility offering a high level of data analysis of specialist level CURFs.

As noted in [28], "It is ABS policy that no information will be released that compromises the undertaking of confidentiality we have made with providers. In practice this means that:

- Aggregated data will not be published or released at a fine level if
  - The major proportion is from one business, or
  - There are fewer than three businesses contributing.

  Data suppression occurs in these instances.
- When releasing unit record information,
  - Any identifying information is removed (i.e. name, address etc),

- Units that are spontaneously identifiable are removed (such as very large businesses in certain industries who will be recognisable from other information on the dataset) and
- Some data perturbation occurs to maintain both the confidentiality and structure of the dataset."

The ABS acknowledges that: "It is primarily the impact that the confidentiality policy has on the release of information from ... large businesses that is of concern."

Through the ABS, the Business Longitudinal Database for 2004-05, 2005-06 and 2006-07 is available as an Expanded CURF, while the Business Longitudinal Survey for 1994-95 to 1997-98 is available as a Basic CURF.

United Kingdom Office for National Statistics

The United Kingdom Office For National Statistics (ONS) Business Data Linking (BDL) Project provides access to business data only via its secure on-site Microdata Lab, where academic researchers can carry out statistical analyses, see [16]. This data is confidential, therefore access is tightly restricted.

The restrictions can be summarised as:

1. Only researchers fully employed at bona fide academic or charitable research institutes, or civil servants, may have access. There is no facility at the moment for PhD students.
2. The employer is required to sign an agreement taking collective responsibility for the actions of all its researchers. Researchers are required to agree to standard secondment contract terms. There is no access without signed agreements.
3. Projects must be of academic value and demonstrate (a) a clear interest for ONS in the results (b) the specific need for the datasets requested.
4. Access is only granted through BDL's secure microdata lab on site at ONS premises.

A research project must specify which dataset it wants to use, why it wants to use it, and why the data cannot be found elsewhere. Additional datasets may be contributed or linked by researchers. The procedures BDL uses to ensure efficient and safe access to data include the signing of relevant contracts.

Business data sets made available through the ONS BDL include: Annual Respondents Database (1970-2001), New Earnings Survey (1986-2002), Business Enterprise R&D (1994-2000) and Capital Stock (1980-2001).

United States Census Bureau

The Census Bureau's Center for Economic Studies (CES) allows special research projects using microdata files, under strictly controlled confidentiality rules, at Census Research Data Centers (RDC), see [29]. Researchers with approved RDC projects gain restricted access to selected internal microdata from the Census Bureau and other statistical agencies for statistical purposes only. Approved researchers are sworn for life to protect the confidentiality of the data they access, and projects must provide benefit to Census Bureau programs, demonstrate scientific merit and pose no risk of disclosure.

Restricted use business data includes: Economic Census files (1977-2000), Longitudinal Business Database (1977-2005) and Annual Survey of Manufacturers (1973-2006).

In addition, the CES research program develops public-use business data products by combining and enhancing existing data. Examples include: Business Dynamics Statistics, Quarterly Workforce Indicators, Synthetic OnTheMap (data on where workers live and work), Synthetic Survey of Income and Program Participation and Synthetic Longitudinal Business Database.

2.2 Potential applicability of the remote analysis approach to business data

Large businesses can be included in analyses processed by the remote analysis system and the analysis results then represent all businesses. This is in marked contrast to the current situation in which released business datasets only cover small and medium sized enterprises, and conclusions can only be drawn for this restricted part of the business sector, see Figure 2.

However, remote analysis also has restrictions and disadvantages, including restrictions on allowable data transformations, data subsetting and new variable definitions. For example, it is common practice in economics to use information like the variable area to identify small, medium and large firms in order to distinguish differences between these groups. Such an analysis would not automatically be available through a remote analysis system, unless the custodian would be willing to make the three relevant data subsets available through the system. There are potentially many such useful analyses which would not be possible through a remote analysis system, and the user would need to seek an alternative data access mode.

More serious are the drawbacks associated with outlier treatment, which will impact the treatment of large enterprises in business data. Analysts are not permitted to view outliers (since these present confidentiality risks) and so cannot make their own removal or treatment decisions. Instead, the remote analysis system uses robust regression methods to minimise the influence of outliers and only removes outliers in the presented results. The analyst can be alerted that outliers have been removed from the presented results. If these disadvantages are judged too serious in a given situation, the analyst may have to seek access to the unconfidentialised dataset. For a more detailed discussion of exploratory data analysis in a remote analysis system, see [19].

The important question which we address in this paper is whether the output confidentialisation process in remote analysis is similar to the standard approach of confidentialising the input, in terms of the usefulness of the output and whether it may lead to incorrect inferences.

3 Example business dataset and confidentialisation approaches

In this paper we provide a detailed example which enables a side-by-side comparison of outputs generated under the statistical disclosure control approach and the remote analysis approach, for the common tasks of exploratory data analysis and linear regression on business data.

In this section we describe the dataset itself and the analyses to be performed, as well as the confidentialisation measures applied under the statistical disclosure control and remote analysis approaches.

3.1 The Sugar Farms data

We will use the *Sugar Farms* data from a 1982 survey of the sugar cane industry in Queensland, Australia [4]. The survey was carried out annually by the Queensland Sugarcane Growers Association, and during the 1980's quotas were used to control the amount of cane grown. The sample frame was all members of the Queensland Sugarcane Growers Association, that is, all commercial cane growers in Queensland at the time. The dataset

corresponds to a sample of 338 Queensland sugar farms, where the sample was stratified by cane growing region and size of quota and within each stratum a simple random sample was selected.

The dataset has one nominal categorical variable: Cane Growing Region (region) and five continuous variables: Sugar Cane Area (area), Sugar Cane Harvest (harvest), Receipts (receipts), Costs (costs) and Profit (profit). Note that profit is calculated as the difference between receipts and costs. There are no missing values.

The Sugar Farms data display many of the characteristics of business data, as follow:

1. There are few variables - in this case only six, with one of them (profit) derived as the difference of another two (receipts minus costs).
2. Four of the five variables are continuous.
3. The distributions of many variables are highly skewed.
4. The five farms with receipts over (over $300K) are outliers on most of the continuous variables.

For the purpose of this paper, we assume that the sample design has led to a pattern of equal inclusion probabilities within strata. We are interested in conducting exploratory data analysis of the Sugar Farms data, and then running a linear regression.

3.2 Confidentialised input approach: statistical disclosure control on the Sugar Farms data

Under this approach, the data are first confidentialised with statistical disclosure control techniques and then analysed. Our method is the same as the general approach used by the ABS, as described in Section 2.1, which is to remove or limit identifying information, suppress spontaneously identifiable units such as very large businesses, and use data perturbation. In this section we give details of the statistical disclosure control techniques that we applied to the Sugar Farms data.

First, the records for the five large farms with receipts over $300K were deleted. The variable region is not disclosive, and was not confidentialised. The variable area was determined to be a key identifying variable because of the risk of matching area values to public registers of farm size and thereby re-identifying farms. It is common practice to reduce the risk of matching to external databases by coarsening the key identifying variables, so we categorised area into six groups, namely up to 29, 30-39, ..., 60-79 and 80 and over. The categorisation of area was chosen so that the cross-classification of area with region has at least 3 farms in each cell (see Section 3.4).

In official statistics datasets such as the Sugar Farms Data, it is important to preserve the additivity constraint in the dataset, and we therefore used Gaussian additive noise. Note that the removal of the large farms avoids the problem of needing excessive additive noise. Each of the target survey variables harvest, receipts, costs and profit was perturbed by the addition of random noise generated from a Multivariate Normal Distribution. This noise was chosen to preserve the mean and covariance structure of the target survey variables, as well as ensuring the edit constraint of profit being equal to receipts minus costs for each farm [15]. However, it is important to note that the process preserves the properties of the dataset *without the large farms* as these are removed prior to the addition of noise.

The details follow. Denote the variables by harvest h, receipts r, costs c and profit p, where $p = r - c$, and the generated multivariate random noise by $(\epsilon_h, \epsilon_r, \epsilon_c, \epsilon_p)^T \sim N(\mu', \mathbf{\Sigma})$, where Σ is the covariance matrix of the original data and the superscript T denotes the

transpose. For each quintile, let μ' be the vector of corrected means of the four variables, based on a noise parameter $\delta = 0.7$. In order to preserve sub-totals and limit the amount of noise, the random noise was generated within quintiles of receipts (note that for ease of notation we drop the quintile index in the following). Let $d_1 = \sqrt{(1-\delta^2)}$ and $d_2 = \sqrt{\delta^2}$, so that the corrected means are

$$\mu'^T = \left(\mu'_{\mathbf{h}}, \mu'_{\mathbf{r}}, \mu'_{\mathbf{c}}, \mu'_{\mathbf{p}}\right) = \left(\frac{(1-d_1)\mu_{\mathbf{h}}}{d_2}, \frac{(1-d_1)\mu_{\mathbf{r}}}{d_2}, \frac{(1-d_1)\mu_{\mathbf{c}}}{d_2}, \frac{(1-d_1)\mu_{\mathbf{p}}}{d_2}\right),$$

where $\mu^T = (\mu_{\mathbf{h}}, \mu_{\mathbf{r}}, \mu_{\mathbf{c}}, \mu_{\mathbf{p}})$ is the vector of means of the original variables (after removing the large farms). For each variable, we calculate a linear combination of the original variable and the random noise generated above, for example, the variable harvest h for record i would be perturbed to $h'_i = d_1 h_i + d_2 \epsilon_{h_i}$. Note that $\delta = 0.7$ means that we construct a composite estimator where the weight is 0.7 of the true value plus 0.7 of the noise. The amount of noise introduced into the dataset was chosen to ensure that no single individual can be identified by their original values, in order to ensure a fair comparison with the remote analysis server which is based on this premise The impact of the added noise on the variable receipts can be seen by comparing the univariate exploratory data analysis results in Figures 7 and 9, see Section 4.2.4. In particular, the percentages of records with perturbed value differing from the original value by less than 20% are: harvest 76%, receipts 84%, costs 68% and profit 48%. Profit has more perturbation due to the additivity constraint. The approach of perturbing with correlated noise is similar to the synthetic approach assuming normally distributed variables but with a noise parameter of $\delta = 1$. The mean vector and the covariance matrix remain the same as the original data with the five large farms removed, and the additivity constraint is exactly preserved. The resulting dataset is said to be *confidentialised*.

The results of the analyses of the confidentialised dataset are shown in Sections 4, 5 and 6. It will be important to remember that the confidentialised dataset preserves the properties of the original data minus the large farms, because that will have an impact on the results. For example, there will be an impact on the mean according to the number and values of the large farms, and variances are likely to be reduced because large farms were removed.

3.3 Confidentialised output approach: remote analysis of the Sugar Farms data

Under this approach, the data are first analysed then the results are confidentialised. One of the main ways that disclosures of information about variable values can occur is through the existence of small numbers of data cases with a given combination of values (this is the problem of so-called *small cells* in tabular data). Therefore many of the measures taken to confidentialise analysis output simply ensure that each combination of variable values has sufficient data cases represented, through data winsorising or aggregation, and by rounding or smoothing of the results. Additional disclosure risk associated with influential large outliers can be reduced by using robust methods.

In this paper we will confidentialise the results of exploratory data analysis and linear regression conducted on the Sugar Farms data, using the general methods proposed in [27], see also [20] for details.

In the case of exploratory data analysis, the confidentialisation measures include:

- Suppress all output if data set is too small
- Suppress or amalgamate all low frequency variable values or ranges

- Remove outliers in graphical output
- Round or replace exact data values (such as median and quantiles) in output
- Replace tables with correspondence analysis plots
- Replace scatter plots by (confidentialised) parallel boxplots
- Replace Q-Q and P-P plots with robust regression lines

Note that the remote analysis system does not enable the output of user-defined tables. The problem of confidentialising tabular output has been extensively studied in the literature and several good solutions exist, see for example [12, 30] and the *Tau Argus* home page [10].

For linear regression, confidentialisation measures include:

- Queries are run on a data subset - the same subset each time the same query is run
- Only restricted transformation of variables are permitted, for example, log
- At most 2-way variable interactions are permitted
- If model error is too small, parameter estimates are not returned
- Parameter estimates are rounded
- Diagnostic plots of residuals are confidentialised

There is an important difference between the treatment of the five large farms in the Sugar Farms data under the confidentialised input and the confidentialised output approaches. Recall that in the confidentialised input approach, the five large farms are removed from the dataset before perturbation is applied. These large farms may or may not be outliers. In the confidentialised output approach for exploratory data analysis, the outliers removed from plots are likely to include the five large farms.

However, in the confidentialised output approach for linear regression, outliers are removed from diagnostic plots such as plots of regression residuals, and it is not guaranteed that these outliers will include points which correspond to the five large farms.

3.4 Disclosure risk in the confidentialised input and confidentialised output approaches

In Sections 4, 5 and 6 below, we provide a comparison of the confidentialised input approach with the confidentialised output approach, for some common statistical analyses. We provide a discussion of the usefulness of the results obtained under the two approaches, with reference to the unconfidentialised results obtained by running the traditional analyses on the original dataset. In order for the comparison between the two approaches to be meaningful in assessing utility, the two approaches should have comparable disclosure risk. In this section we quantify the disclosure risk under the two approaches. If the two approaches have approximately the same disclosure risk, then we can assume that they provide approximately the same level of confidentiality protection. In that case, our comparison of the usefulness of the outputs is valid.

When dealing with statistical microdata, there are two types of disclosures which should be considered:

- *Identity disclosure* where an intruder re-identifies a data subject represented in the microdata, normally through learning the value of identifying key variables in the dataset.

- *Attribute disclosure* where an intruder learns some new information and attributes it to a data subject represented in the microdata, normally through learning the value of sensitive target variables in the dataset.

We emphasize that business data differ from microdata arising from social surveys, where typically protection against identity disclosure is enough to be able to release the microdata. Business data are never released by agencies unless highly perturbed by, for example, removal of all large businesses and application of other disclosure limitation techniques or by replacement of the entire dataset with a synthetic dataset (see [21] and [23]). This is because of the typical skewed distributions and the likelihood that values of sensitive variables are released through publically available sources. For this reason, we need to protect the Sugar Farms business dataset against both identity and attribute disclosure.

The confidentialised input approach seeks to reduce the risk of identity disclosure by ensuring that there are no small counts in the cross-classifications of identifying key variables. In the case of the Sugar Farms data, the identifying key variables are region and area, since these are likely to be publically available on external administrative datasets. The discrete variable region has only four categories, and was not altered in the confidentialisation process for the confidentialised input approach. The continuous variable area was coarsened into six categories as described in Section 3.2, where the categories were chosen to ensure that the cross-classification of region and area does not have any cell with a count of three or less. The confidentialised output approach seeks to reduce the risk of identity disclosure by ensuring that each combination of variable values has sufficient data cases represented, through results suppression, data winsorising or aggregation, and by rounding or smoothing of the results. In this way, the two approaches provide a comparable level of protection against the risk of identity disclosure.

Attribute disclosure risk is relevant for the sensitive variables harvest, receipts and costs. In the confidentialised input approach, we propose to approximate a measure of attribute disclosure risk by calculating the sum of the relative absolute differences between the variable values in the confidentialised dataset and the variable values in the original dataset. Since the variable region was not confidentialised, the attribute disclosure risk measure is calculated for each of the four regions separately. In the confidentialised output approach, there are no individual confidentialised variable values to use in a comparable measure for attribute disclosure risk. Instead, we generated estimates from the exploratory data analysis results, where the results for receipts are shown in Figures 8 and 11(b) and results for costs are shown in Figure 11(c). The method for generating the estimates was as follows. Each single box plot for the sensitive variable on a region provides an estimate of the values: lower whisker, first quartile, median, third quartile and the upper whisker. We assume that 25% of the values of the sensitive variable lie in the interval between the lower whisker and the 1st quartile, 25% of the values lie in the interval between the 1st quartile and the median, 25% of the values lie in the interval between the median and the 3rd quartile, and finally, 25% of the values lie in the interval between the 3rd quartile and the upper whisker. Next, since we have no other information, we assume that the values of the sensitive variable are equally distributed across the intervals of the box plot and we carry out an interpolation to obtain individual estimated values. We then estimate the attribute disclosure risk measure for each sensitive variable, on each region, by computing the sum of the relative absolute difference between the estimated value based on the interpolation and the original value. The table in Figure 3 provides a comparison of the attribute disclosure risk measure values for the sensitive variables harvest, receipts and costs under the two confidentiality approaches. To make the comparison, the large farms that were removed in

the confidentialised input approach are not included in either of the risk measures.

Variable		Attribute Disclosure Risk	
		Confidentialised Input Approach	Confidentialised Output Approach
Region 1	Harvest	16.80	12.94
	Receipts	14.39	14.50
	Costs	21.28	8.55
Region 2	Harvest	3.82	2.04
	Receipts	3.28	5.59
	Costs	5.66	3.63
Region 3	Harvest	14.82	5.80
	Receipts	9.58	7.84
	Costs	20.33	5.73
Region 4	Harvest	12.20	4.16
	Receipts	9.19	4.78
	Costs	16.56	6.51

Figure 3: Attribute disclosure risk for sensitive variables harvest, receipts and costs within regions (not including the large farms) for the two confidentiality approaches

In Figure 3, the attribute disclosure measures are generally smaller for the confidentialised output approach based on interpolation of individual values than for the confidentialised input approach based on the addition of noise. However, the two approaches give measures that are similar in magnitude for most cases and therefore we can assume that both confidentiality approaches have comparable levels of confidentiality protection.

4 Univariate exploratory data analysis of the Sugar Farms data

In this section we give a comparison of univariate exploratory data analysis outputs under the confidentialised input and confidentialised output approaches, with reference to the unconfidentialised analysis.

We focus on the variables area and receipts, in Sections 4.1 and 4.2 respectively, as representative of the variables present. The comparison for the non-disclosive variable region would be the same under the confidentialised input and confidentialised output approaches, so it is not given. The comparisons for harvest, costs and profit are similar to that for receipts, so are also omitted.

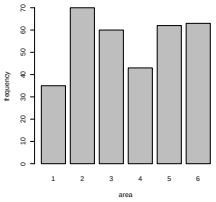
Recall that in the confidentialised input approach, the continuous variable area is categorised into six groups and the variable receipts has noise added.

4.1 The variable area

4.1.1 Confidentialised input approach

area		
	Category	Frequency
1	Up to 29	35
2	30 – 39	70
3	40 – 49	60
4	50 – 59	43
5	60 – 79	62
6	80 and over	63
	TOTAL	333

(a) Frequency Table



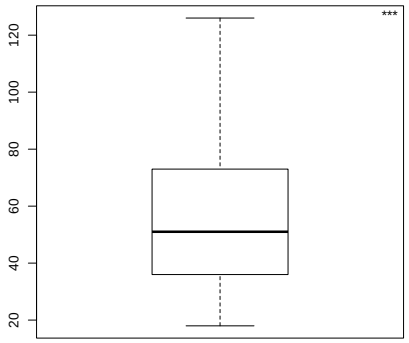
(b) Bar Chart

Figure 4: Univariate exploratory data analysis output for the variable area in the confidentialised Sugar Farms data

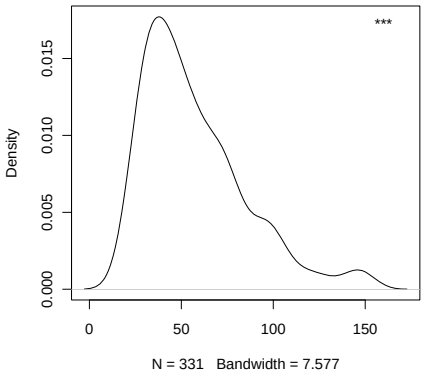
4.1.2 Confidentialised output approach

Sugar Cane Area	
1st Quartile	35
Median	50
Mean	60.25
3rd Quartile	70
Standard Deviation	36

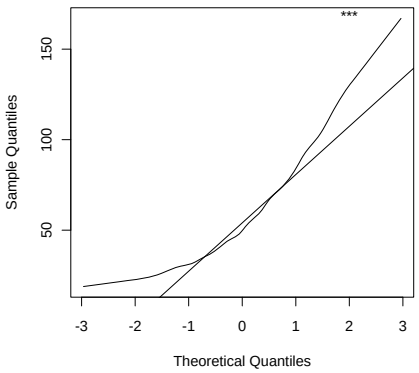
(a) Confidentialised Summary Statistics



(b) Confidentialised Box Plot



(c) Confidentialised Density Estimate



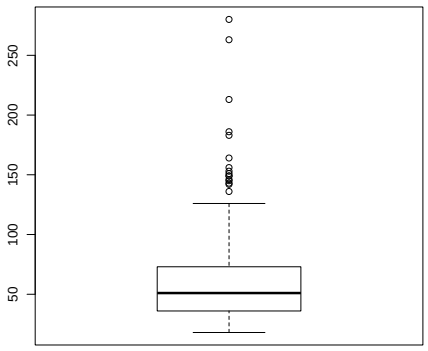
(d) Confidentialised QQ-Plot

Figure 5: Confidentialised output from exploratory data analysis of variable area in the Sugar Farms data. The symbol $***$ is used in a figure to indicate that outliers have been removed for plotting.

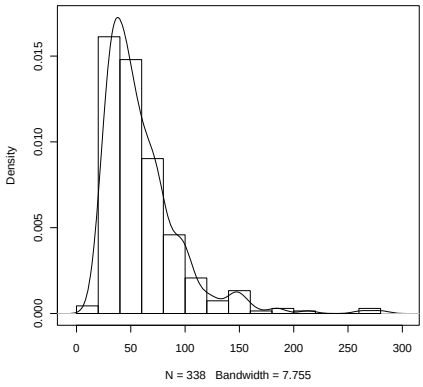
4.1.3 Unconfidentialised analysis

Sugar Cane Area	
Minimum	18
1st Quartile	36
Median	51
Mean	60.25
3rd Quartile	73
Maximum	280
Standard Deviation	35.61062

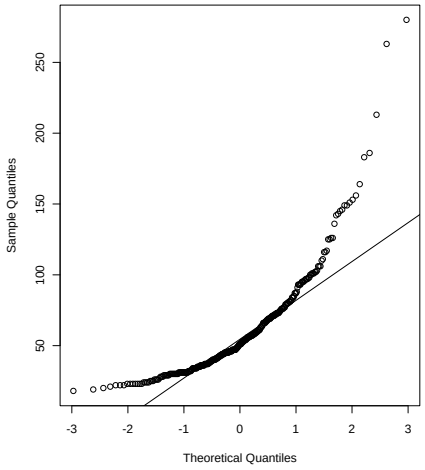
(a) Summary Statistics



(b) Box Plot



(c) Histogram and Density



(d) Normal QQ-plot

Figure 6: Unconfidentialised univariate exploratory data analysis output for the variable area in the Sugar Farms data

4.1.4 Comparison and discussion

From the table in Figure 4(a), the analyst knows which group contains the median and the quartiles, which probably gives reasonable estimates based on linear interpolation. However, it would be harder for the analyst to be confident about estimating the mean or standard deviation and so reproduce the information in Figure 5(a). The confidentialised output approach provides rounded quartiles, mean and standard deviation. Neither of the confidentialised approaches provides information about the sample minimum and maximum, which does appear in the unconfidentialised results in Figure 6(a).

The bar chart in Figure 4(b) does not capture the shape of the data seen in the confidentialised boxplot in Figure 5(b) and the confidentialised density in Figure 5(c). The confidentialised boxplot and density appear quite close to the unconfidentialised output in Figures 6(b) and 6(c), except that large farms are not represented or included.

The confidentialised input results in Figure 4 give no equivalent to the Q-Q plot shown in

Figure 5(d). An analyst provided with the confidentialised input would be unaware that the variable area is skewed, however this is indicated under the confidentialised output approach and confirmed in the unconfidentialised results in Figure 6(d). It is worth noting that the normal distribution in the confidentialised output and unconfidentialised results have the same mean and variance.

In comparing the three sets of results, it is important to look carefully at the differences in the scales. The removal of large farms in the confidentialised input approach and the removal of plot outliers in the confidentialised output approach may both result in a compression of the plot scales in comparison with the unconfidentialised results.

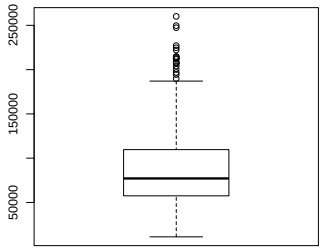
In summary, the categorisation of area into six groups, which was necessary to avoid identity disclosure, has led to a significant deterioration in the information made available to the analyst in the confidentialised input case, in comparison with the confidentialised output case. The information provided under the confidentialised output approach appears to give a good indication of the characteristics of the original dataset with the large farms removed.

4.2 The variable receipts

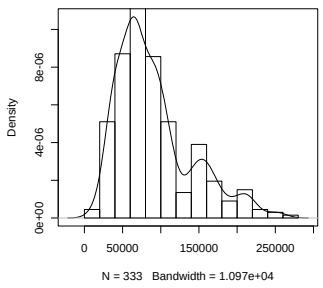
4.2.1 Confidentialised input approach

	Receipts
No. observations	333
Minimum	11140
1st Quartile	57473
Median	77144
Mean	90643
3rd Quartile	109637
Maximum	260098
Standard Deviation	49214.06

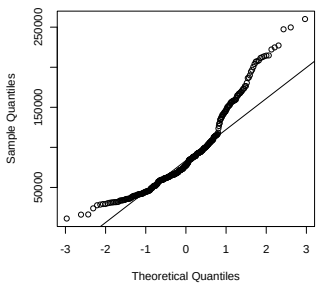
(a) Summary Statistics



(b) Box Plot



(c) Histogram and Density



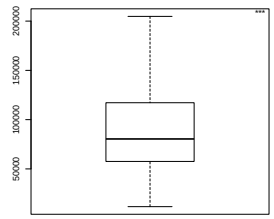
(d) Normal QQ-plot

Figure 7: Univariate exploratory data analysis output for the variable receipts in the confidentialised Sugar Farms data

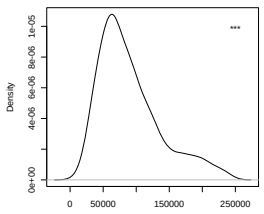
4.2.2 Confidentialised output approach

	Receipts
1st Quartile	57600
Median	80400
Mean	96000
3rd Quartile	117100
Standard Deviation	61600

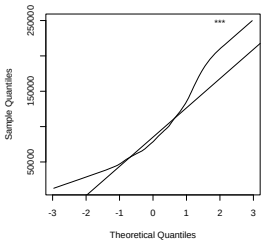
(a) Confidentialised Summary Statistics



(b) Confidentialised Box Plot



(c) Confidentialised Density Estimate



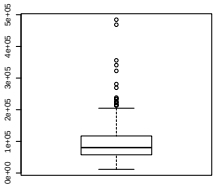
(d) Confidentialised QQ-Plot

Figure 8: Confidentialised output from exploratory data analysis of variable receipts in the Sugar Farms data. The symbol $***$ is used in a figure to indicate that outliers have been removed for plotting.

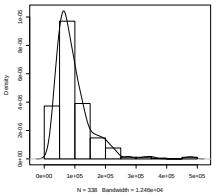
4.2.3 Unconfidentialised analysis

	Receipts
Minimum	11703
1st Quartile	57607
Median	80391
Mean	95965
3rd Quartile	117062
Maximum	484346
Standard Deviation	61609.105256

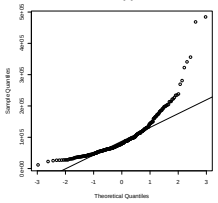
(a) Summary Statistics



(b) Box Plot



(c) Histogram and Density



(d) Normal QQ-plot

Figure 9: Unconfidentialised univariate exploratory data analysis output for the variable receipts in the Sugar Farms data

4.2.4 Comparison and discussion

Comparing the Summary Statistics in Figures 7(a), 8(a) and 9(a), we see that the values for 1st Quartile are quite close across the three approaches. The median, mean 3rd Quartile and standard deviation are smaller in the confidentialised input approach than in the other two approaches. The minimum in the confidentialised input approach is quite close to the unconfidentialised, however the maximum is significantly smaller due to the removal of the large farms. As seen previously in Section 4.1.4, minimum and maximum are missing from the confidentialised output results. These observations are not surprising as in the confidentialised input approach large farms were removed and the added noise preserved the first two moments of the dataset without the large farms. In summary,

The box plots in Figures 7(b), 8(b) and 9(b) have similar boxes, once the difference in the scales is noted, except that the confidentialised input box plot has a smaller upper quartile (this was noted in the Summary Statistics). The confidentialised input box plot in Figure 7(b) shows values above the upper whisker, but the maximum is smaller in comparison with the unconfidentialised box plot in Figure 9(b). The confidentialised output box plot in Figure 8(b) shows no values above the upper whisker.

In the confidentialised output in Figure 8(c), the histogram has been omitted and the tail of the density has been truncated. However, the density is quite close to the unconfidentialised density in Figure 9(c). The confidentialised input histogram and density in Figure 7(c) appear a bit different from the unconfidentialised histogram and density in the upper tail. Again, there is a difference in scales due to removal of dataset large farms in the confidentialised input approach and the removal of plot outliers in the confidentialised output approach.

The confidentialised input and confidentialised output Q-Q plots in Figures 7(d) and 8(d) are quite similar shapes, though both are truncated with respect to the unconfidentialised plot in Figure 9(d).

In summary, the confidentialised output results appear to give good information about the original data. The confidentialised input results, particularly those provided as 3rd Quartile, standard deviation, maximum, histogram and density, give good information about the sub-population of farms which does not include the large farms.

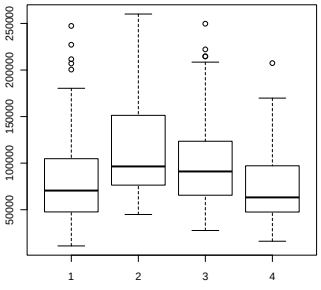
5 Bivariate exploratory data analysis of the Sugar Farms data

5.1 Bivariate area, receipts and costs with region

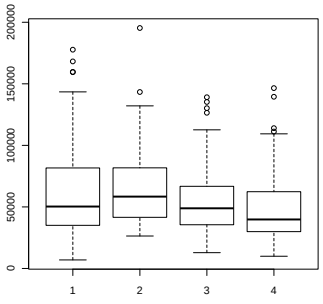
5.1.1 Confidentialised input approach

		region			
		1	2	3	4
area	Up to 29	11	7	3	14
	30 – 39	18	7	16	29
	40 – 49	20	7	16	17
	50 – 59	18	6	9	10
	60 – 79	24	5	22	11
	80 and over	32	6	19	6

(a) area by region



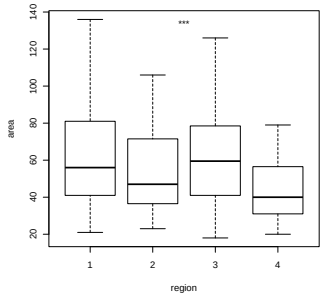
(b) receipts by region



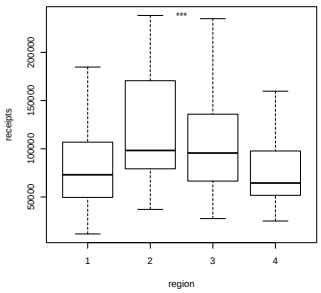
(c) costs by region

Figure 10: Bivariate exploratory data analysis for area, receipts and costs with the discrete variable region in the confidentialised Sugar Farms data

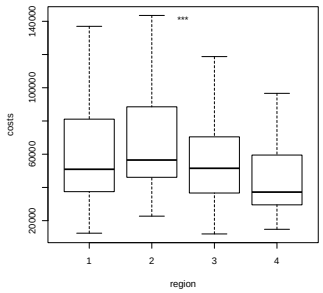
5.1.2 Confidentialised output approach



(a) Box plots for area by region



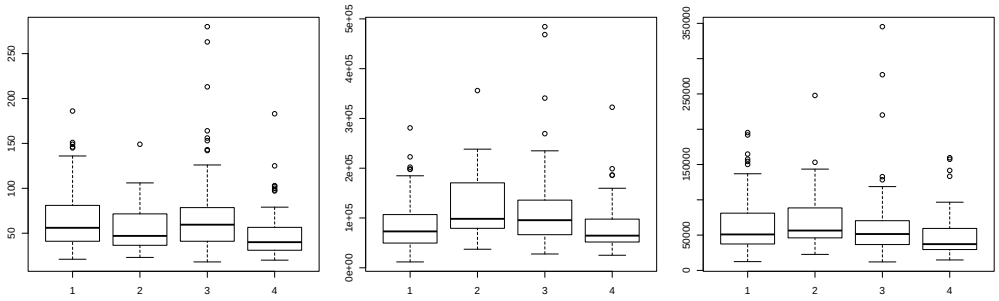
(b) Box plots for receipts by region



(c) Box plots for costs by region

Figure 11: Confidentialised bivariate exploratory data analysis for each continuous variable area, receipts and costs with the discrete variable region. The symbol $***$ is used in a figure to indicate that outliers have been removed for plotting.

5.1.3 Unconfidentialised analysis



(a) Box plots for area by region (b) Box plots for receipts by region (c) Box plots for costs by region

Figure 12: Unconfidentialised bivariate exploratory data analysis for each continuous variable area, receipts and costs with the discrete variable region

5.1.4 Comparison and discussion

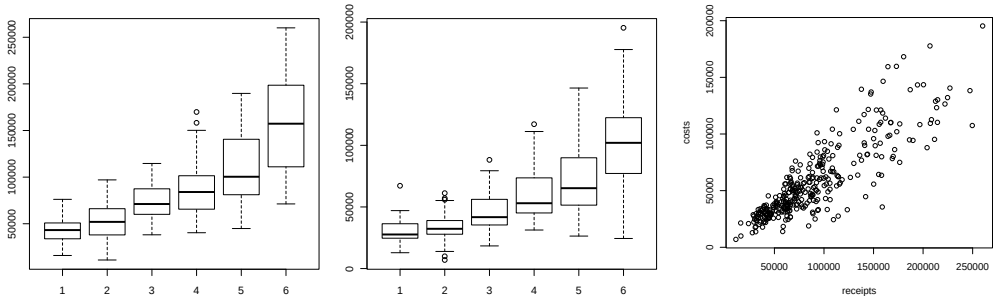
Comparing the bivariate exploratory data analysis output for area by region in Figures 10(a), 11(a) and 12(a), we see that the confidentialised input approach leads to a clear loss of information in comparison with the confidentialised output and unconfidentialised results. The analyst could obtain reasonable estimates of the minimum, median and 1st Quartiles from the table in Figure 10(a), but estimates of 3rd quartile and maximum from the table would be too low. The confidentialsed output box plots in Figure 11(a) do not show values beyond the ends of the whiskers, and the scale is compressed in comparison with the unconfidentialised box plots in Figure 12(a), due to the removal of plot outliers.

The box plots for receipts by region in Figures 10(b), 11(b) and 12(b) and costs by region in Figures 10(c), 11(c) and 12(c) are quite similar across the confidentialised input, confidentialised output and unconfidentialised results, except that the confidentialised output in Figures 11(b) and 11(c) do not show values beyond the extremes of the whiskers, and display compressed scales due to the removal of plot outliers. The scales on the confidentialised input box plots in Figures 10(b) and 10(c) are also compressed, due to the removal of large farms. The compression is particularly noticeable for the 3rd quartile, extent of upper whisker and maximum.

In summary, the categorisation of area into six groups has led to a significant deterioration in the information made available to the analyst in the confidentialised input case, in comparison with the confidentialised output case. The information provided under the confidentialised output approach appears to give a good indication of the characteristics of the original dataset with the large farms removed.

5.2 Bivariate pairs from area, receipts and costs

5.2.1 Confidentialised input approach



(a) receipts by area (b) costs by area (c) receipts by costs

	area	receipts	costs
area	1	0.7487469 p-value < 2.2e-16	0.7145667 p-value < 2.2e-16
receipts		1	0.8594960 p-value < 2.2e-16
costs			1

(d) Pearson Correlation Coefficients

Figure 13: Bivariate exploratory data analysis for pairs of variables receipts, area and costs in the confidentialised Sugar Farms data

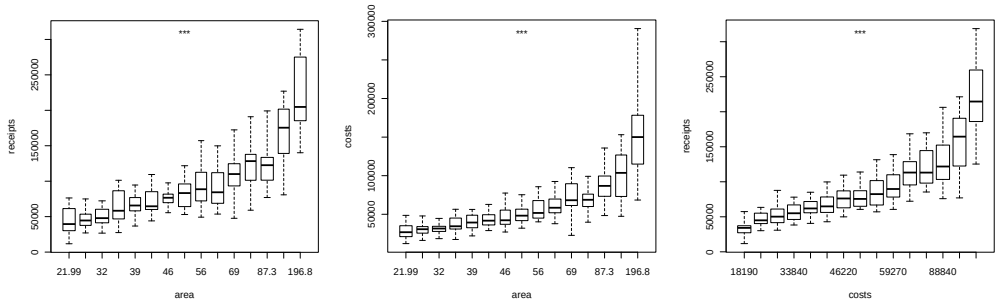
Area	Receipts (thousands)						
	Less than 40	40-59.9	60.0-79.9	80.0-99.9	100.0-139.9	140.0 and over	Total
Less than 29	14	13	8	0	0	0	**35**
30-39	20	25	17	8	0	0	**70**
40-49	3	11	25	12	9	0	**60**
50-59	0	7	14	11	6	5	**43**
60-79	0	2	12	16	16	16	**62**
80 and Over	0	0	1	10	12	44	**63**
Total	**37**	**58**	**77**	**57**	**43**	**61**	**333**

$\chi^2 = 268.5 \qquad p < 0.0001 \qquad C.V. = 0.4015$

Figure 14: Chi-Square test and Cramer's V (C.V.) for area with receipts in the confidentialised Sugar Farms data

Normally a "1" or a "2" in a cell in a cross-tabulation is considered to be a confidentiality concern. However, in this case the data have been confidentialised with statistical disclosure control processes, so only confidentialised values are revealed. Therefore, the small counts in this table are not a confidentiality concern.

5.2.2 Confidentialised output approach



(a) receipts vs area (b) costs vs area (c) receipts vs costs

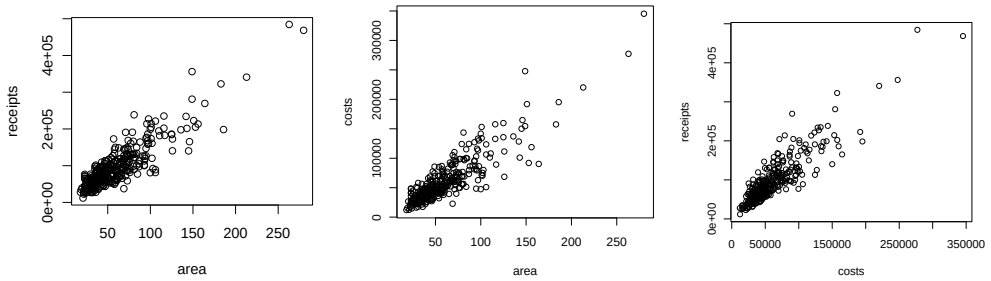
	area	receipts	costs
area	1	0.8877 ***	0.8868 ***
receipts		1	0.9010 ***
costs			1

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(d) Pearson Correlation Coefficients

Figure 15: Confidentialised bivariate exploratory data analysis output for the variables area, receipts and costs in the Sugar Farms data. The symbol $***$ is used in a figure to indicate that outliers have been removed for plotting.

5.2.3 Unconfidentialised analysis



(a) receipts by area (b) costs by area (c) receipts by costs

	area	receipts	costs
area	1	0.8876671 p-value < 2.2e-16	0.8867933 p-value < 2.2e-16
receipts		1	0.90096490 p-value < 2.2e-16
costs			1

(d) Pearson Correlation Coefficients

Figure 16: Unconfidentialised bivariate exploratory data analysis for pairs of continuous variables area, receipts and costs in the Sugar Farms data

Area	Receipts (thousands)						
	Less than 40	40-59.9	60.0-79.9	80.0-99.9	100.0-139.9	140.0 and over	Total
Less than 29	18	9	8	0	0	0	**35**
30-39	13	31	19	6	1	0	**70**
40-49	0	13	23	19	5	0	**60**
50-59	0	7	13	11	9	3	**43**
60-79	1	1	11	15	24	10	**62**
80 and Over	0	0	1	8	15	44	**68**
Total	**32**	**61**	**75**	**59**	**54**	**57**	**338**

$\chi^2 = 346.8$ $p < 0.0001$ $C.V. = 0.4530$

Figure 17: Unconfidentialised Chi-Square test and Cramer's V (C.V.) for area with receipts in the Sugar Farms data

5.2.4 Comparison and discussion

In the confidentialised input approach, area has been transformed into a discrete variable with six groups, so only the box plots in Figures 13(a) and 13(b) are provided in place of the scatter plots in Figures 16(a) and 16(b). The scatter plot in Figure 13(c) is provided in place of the scatter plot in Figure 16(c). General relationships between the variables are still observed in the confidentialised input plots, however the scales of the plots are compressed due to the removal of large farms. The confidentialised output approach provides

box plots as in Figures 15(a), 15(b) and 15(c). The box plot intervals are narrower than in the confidentialised input box plots in Figures 13(a) and 13(b). The general relationships between the variables are still observed in the confidentialised output plots, however the scales of the plots are compressed due to the removal of plot outliers. As before, the scales of the confidentialised input and confidentialised output plots are not the same, and both are compressed in comparison with the unconfidentialised plots.

The confidentialised input Pearson correlation coefficients in Figure 13(d) are smaller than in the confidentialised output in Figure 15(d) and the unconfidentialised version in Figure 16(d), so the confidentialised input approach underestimates the correlations between the variables. This may be because the large farms are the most highly correlated observations, and they are omitted in the confidentialised input approach. In contrast, the confidentialised output correlations are rounded from the true unconfidentialised versions, so correlations are accurately reported.

The confidentialised output does not include a Chi-Square test, since the confidentialisation of Chi-Square output when one or both variables is continuous requires some additional system functionality. The issue is in the specification of the categories for continuous variables, when it leads to small cells. Even if the table is not provided to the researcher, there would need to be an indication of whether expected cell counts are too small to obtain a valid chi-square distribution. The confidentialised input table in Figure 14 is reasonably similar to the unconfidentialised version in Figure 17, with small differences due to the data set confidentialisation process. The confidentialised input Chi- square value and Cramer's V (C.V.) are smaller than the unconfidentialised approach at the same significance level. For the chi-square test, the confidentialised input approach provides the same decision in terms of rejecting the null hypothesis of independence.

In summary, the confidentialised input and confidentialised output approaches seem to provide fairly similar bivariate information about non-transformed pairs of variables, except that the confidentialised input approach underestimates the correlations amongst the variables. The confidentialised output approach does not allow the discretising of continuous variables in order to perform a chi-square test as would be possible in the confidentialised input approach but will enable a chi-square test for two pre-specified categorical variables.

6 Regression analysis

In this section, we compare the results of conducting a linear regression analysis on the Sugar Farms data under the different approaches discussed in this paper. We are interested in modelling receipts as the response variable, with explanatory variables region, area, harvest and costs. Since profit is a derived variable calculated as the difference between receipts and costs, we omit it from the model to avoid collinearity. The exploratory data analysis conducted suggests that it may be appropriate to transform the variables receipts, harvest and costs using the log function, so our model has log(receipts) as response, with region, area, log(harvest) and log(costs) as explanatory variables.

We note that area can be included directly as a continuous variable in the confidentialised output and unconfidentialised regression models. In the confidentialised input regression, we include area as a continuous variable with a scale of 1-6. We note that the confidentialised output approach implements a robust regression method to minimise the impact of dataset outliers such as large farms.

The following sections show the regression results and provide a discussion.

6.1 Summary results

	Confidentialised Input	Confidentialised Output	Un-confidentialised
Intercept	3.627253	3.06	2.7060226
p-value	< 2e-16		< 2e-16
significance	***	***	***
Factor(region)2	0.192557	0.205	0.1814301
p-value	2.97e-15		< 2e-16
significance	***	***	***
Factor(region)3	0.187611	0.244	0.2390758
p-value	< 2e-16		< 2e-16
significance	***	***	***
Factor(region)4	0.091021	0.117	0.1184681
p-value	1.91e-7		< 2e-16
significance	***	***	***
area	0.031205	0.0004	0.0000792
p-value	4.81e-6		0.773
significance	***		
harvest	0. 831541	0.883	0.8655644
p-value	< 2e-16		< 2e-16
significance	***	***	***
costs	0. 063136	0.0823	0.1309820
p-value	0.0147		4.05e-8
signficance	*	***	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 18: Comparison of coefficient estimates and significance levels for linear regression of log(receipts) dependent on region, area, log(harvest) and log(costs) under the confidentialised input, confidentialised output and unconfidentialised approaches

The most important difference between the results in Figure 18 is that the confidentialised input approach incorrectly concludes that area is a significant explanatory variable, and underestimates the significance of costs by two significance levels. In contrast, the confidentialised output significance levels are all correct. It is possible that this impact is partly or entirely due to the discretization of the variable area. However as this is part of the input confidentialisation process we do not need to distinguish the source of the incorrect conclusion.

Overall, the parameter estimates for the confidentialised output approach are more similar to the original model and therefore we are able to provide a more accurate interpretation of the effects of the explanatory variables on the response variable receipts.

6.2 Overall goodness-of-fit statistics

	Confidentialised Input	Confidentialised Output	Un-confidentialised
Residual standard error	0.1151	0.08	0.09024
degrees of freedom	326	314	331
Multiple R squared	0.9554	0.97	0.974
Adjusted R squared	0.9546	0.97	0.9735
F-statistic	1164	2100	2067
degrees of freedom	6 and 326	6 and 331	6 and 331
p-value	< 2.2e-16	-	< 2.2e-16
significance	***	***	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

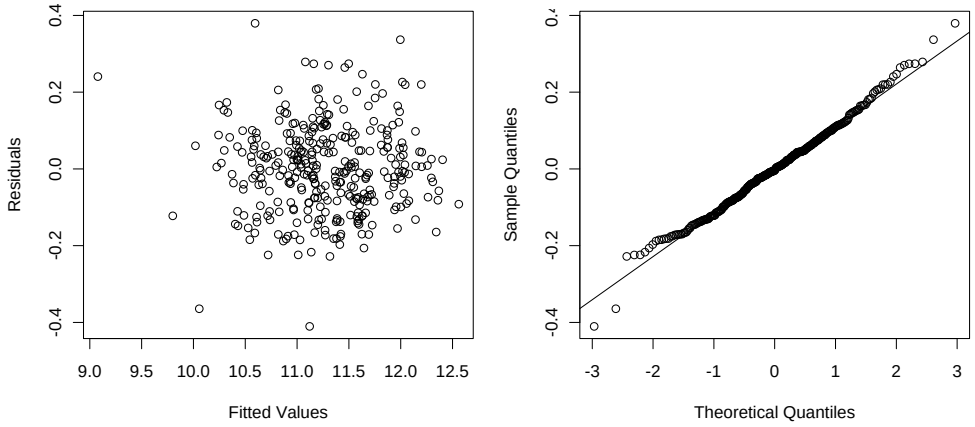
Figure 19: Comparison of goodness of fit statistis for linear regression of log(receipts) dependent on region, area, log(harvest) and log(costs) under the confidentialised input, confidentialised output and unconfidentialised approaches

From Figure 19, the residual standard error is greater for the confidentialised input regression than for the confidentialised output and unconfidentialised regressions. Because the residual standard error is the square root of the sum of the squares of the difference between the observed and predicted values divided by the degrees of freedom, this means that the confidentialised input predicted values will be further from the observed values than the confidentialised output and unconfidentialised predicted values. The difference in residual standard error and degrees of freedom between the confidentialised output and unconfidentialised results is due to the robust regression procedure implemented.

The R squared and adjusted R squared are smaller for the confidentialised input regression than for the unconfidentialised regression. The confidentialised output R squared and adjusted R squared values are the rounded unconfidentialised values.

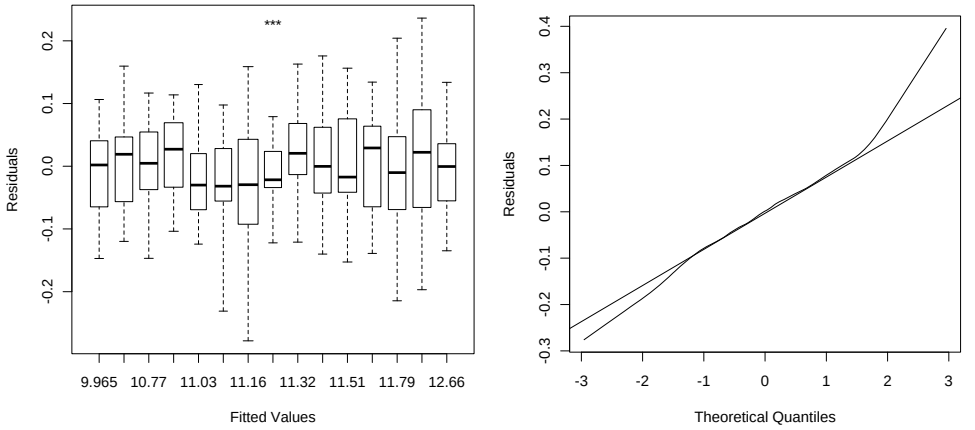
The F statistic is smaller for the confidentialised input regression than for the unconfidentialised regression. but the significance level is correct. The confidentialised output R squared values are the rounded unconfidentialised values.

6.3 Model diagnostics



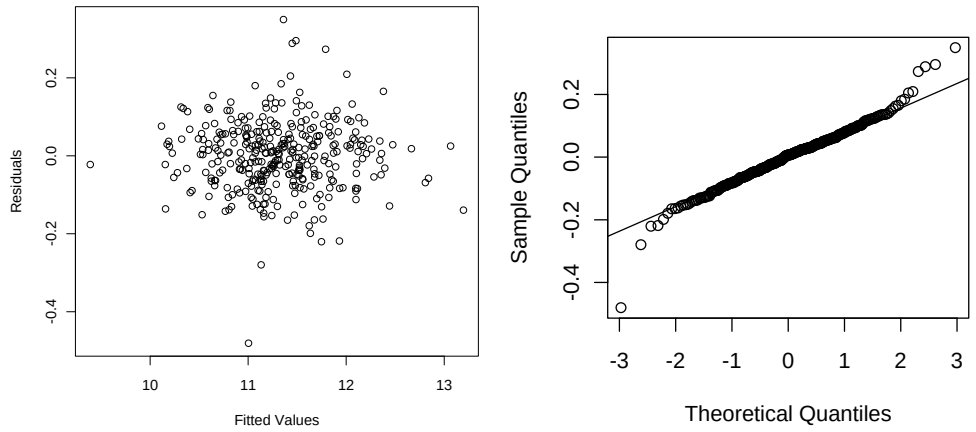
(a) Residuals vs Fitted Values (b) Normal Q-Q Plot of Residuals

Figure 20: Diagnostics for linear regression of log(receipts) dependent on region, area, log(harvest) and log(costs) in the confidentialised Sugar Farms data



(a) Residuals by fitted values (b) Normal QQ Plot of Residuals

Figure 21: Confidentialised diagnostics for linear regression of log(receipts) dependent on region, area, log(harvest) and log(costs) in the original Sugar Farms data. The symbol $***$ is used in a figure to indicate that outliers have been removed for plotting.



(a) Residuals vs Fitted Values

(b) Normal Q-Q Plot of Residuals

Figure 22: Unconfidentialised diagnostics for linear regression of log(receipts) dependent on region, area, log(harvest) and log(costs)

The confidentialised input residual scatterplot in Figure 20(a) shows a compressed spread for higher fitted values in comparison with the unconfidentialised scatterplot in Figure 22(a). The lower fitted values show an increased spread, and the residuals are also slightly more spread for both higher and lower values. The confidentialised output plot provides side-by-side box plots as in Figure 21(b). Both the fitted values and the residuals are compressed to a smaller scale than the unconfidentialised plot in Figure 22(a). The box plots in the confidentialised output plot provide more detailed information about the distribution of the residuals. For example the wave-like behaviour above and below the zero value exhibited by the median values of the box plots is not apparent in either residual scatter plot. It is possible that a meaningful pattern not visible in a residual scatter plot may be revealed in such residual box plots.

The two normal Q-Q plots in Figures 21(b) and 22(b) comparing the distribution of model residuals with the normal distribution give similar information, even though the confidentialised output plot has been smoothed and plot outliers have been removed. The normal Q-Q plot under the confidentialised input approach, Figure 20(b), does not accurately show the features present in the tails of the distribution.

In all cases it is necessary to look carefully at the plot scales. There are differences in the scales, due to the removal of large farms in the confidentialised input plots and the removal of plot outliers in the confidentialised output plots. There is a general compression of plot scales in the confidentialised input and confidentialised output plots in comparison with the unconfidentialised plots.

7 Discussion and conclusions

In this paper we explored the statistical disclosure control (*confidentialised input*) approach and the remote analysis (*confidentialised output*) approach to data confidentiality, in the context of protecting the confidentiality of business data in statistical analysis. In particular, we discussed a detailed example enabling a side-by-side comparison of the outputs of exploratory data analysis and linear regression analysis conducted on the Sugar Farms

dataset under these two approaches, and provided traditional unconfidentialised results as a standard for comparison. The paper therefore contributes to developing an understanding of the potential use of remote analysis in addressing the balance between microdata access and confidentiality protection, in the context of business data.

The main relevant features of the confidentialised input procedure, and their main consequences, are:

- The deletion of large farms at the beginning of the process, as in the ABS method of releasing data only for small or medium businesses. This means that the confidentialised dataset models the properties of the original data minus the large farms, resulting in similar medians but reduced means and variances in the confidentialised dataset.
- The categorisation of area into six groups to reduce identity disclosure risk.
  - In exploratory data analysis, this leads to a significant deterioration in the information made available to the analyst.
  - In regression analysis, there is a real risk of incorrect conclusions regarding signficance of variables, as was seen in our example.

The main relevant feature of the confidentialised output procedure, and its main consequence, is:

- The smoothing and trimming of displayed results. This means that the information presented to the analyst does not exactly correspond to the analysis as it was carried out. For example, the removal of outlying points from residual plots is indicated with three asterisks on the plot. The analyst will therefore know that the model has outlying residual values, but will have no information about their magnitude or impact. Another example is restricting Chi-Square tests to only pairs of categorical variables.

A major difference between the two approaches is that in the confidentialised input approach the five largest farms were deleted before the analyses were conducted. In contrast, in the confidentialised output approach the five largest farms were included in the analyses, however the display of results under this approach were confidentialised to remove outlying points, and indicate this removal with three asterisks. The main consequences of this difference are:

- The two approaches may display results with different groups of farms removed. The analyst is alerted to the removal of large farms in the confidentialised input approach, and plot outliers in the confidentialised output approach.
  - In exploratory data analysis there is minimal or no data processing, so the five largest farms deleted under the confidentialised input approach are likely to be not shown in the confidentialised output approach, although their presence is indicated by three asterisks at the top of the plot. We found this to be true, though the confidentialised output approach often deleted additional outlying farms.
  - In regression analysis there is significant data processing, and in fact robust regression methods are used to reduce the influence of outliers through downweighting them. The outliers in the confidentialised output approach, are likely to include the large farms but may also include other farms that were determined to be influential by the model.

- Plots produced under the two approaches may have significantly different scales. The removal of large farms in the confidentialised input approach and the removal of plot outliers in the confidentialised output approach may both result in a compression of the plot scales in comparison with the unconfidentialised results. It is not guaranteed that the amount of compression in each case is the same. We remark that putting the two plots on the same scale would reveal information about the magnitude of the outliers including large farms.
- The confidentialised input approach may underestimate dataset correlations, because large farms typically contribute more to the correlation in the dataset.
- The overall goodness-of-fit measures would be expected to be more accurate under the confidentialised output approach than the confidentialised input approach.

The comparison provided in this paper is only for one specific SDC approach and one dataset, leaving open the possibility that there might be alternatives that provide higher data utility for the illustrative utility evaluations on the dataset given in the paper. However, we have minimised this possibility by using a dataset which has the common characteristics of business data, and by modelling the confidentialised input approach on published descriptions of the current practices at the Australian Bureau of Statistics and state-of-the-art perturbation methods that ensure statistical properties.

In general, the confidentialised input approach to protecting business microdata involves creating highly noisy variable values (including synthetic values) and recoding many variables, which leads to significant information loss. There are certainly advantages and disadvantages in the remote analysis approach and it is unlikely that remote analysis will replace statistical disclosure control methods in all applications. If the disadvantages are judged too serious in a given situation, the analyst may have to seek access to the unconfidentialised dataset. However, our example supports the conclusion that the advantages may outweigh the disadvantages in some cases, including for some analyses of unconfidentialised business data, provided the analyst is aware of the output confidentialisation methods and their potential impact. For example, we believe that the remote analysis system provides analysts with a good way of developing their research strategies and obtaining preliminary indicative results prior to gaining full access to licensed detailed data in an on-site data laboratory. It may also be suitable for the general public interested in some simple statistics.

Acknowledgements:

We thank Ray Chambers for making the Sugar Farms data available to us. We also thank the anonymous referees for their attention to detail, which has resulted in an improved paper.

References

[1] N.R. Adam and J.C. Wortmann. Security-control methods for statistical databases: A comparative study. *ACM Comput Surv*, 21:515–556, 1989.

[2] N. Ahmad, K. De Backer, and Y. Yoon. An OECD perspective on microdata access: Trends, opportunities and challenges. *Statistical Journal of the IAOS*, 26:57–63, 2009–2010.

[3] Australian Bureau of Statistics. `www.abs.gov.au`.

[4] R.L. Chambers and R. Dunstan. Estimating distribution functions from survey data. *Biometrika*, 73:597–604, 1986.

[5] J. Domingo-Ferrer and E. Magkos, editors. *Privacy in Statistical Databases*, volume 6344 of *Lect Notes Comp Sci*. Springer, 2010.

[6] J. Domingo-Ferrer and Y. Saygin, editors. *Privacy in Statistical Databases*, volume 5262 of *Lect Notes Comp Sci*. Springer, 2008.

[7] J. Domingo-Ferrer and V. Torra, editors. *Privacy in Statistical Databases*, volume 3050 of *Lect Notes Comp Sci*. Springer, 2004.

[8] P. Doyle, J.I. Lane, J.J.M. Theeuwes, and L. Zayatz, editors. *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. North-Holland: Amsterdam, 2001.

[9] J. Drechsler. New data dissemination approaches in old europe - synthetic datasets for a german establishment survey. *J Appl Stat*, 39:243–265, 2011.

[10] ESSnet-project. τ-ARGUS home page.

[11] S. Gomatam, A.F. Karr, J.P. Reiter, and A. Sanil. Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access systems. *Stat Sci*, 20:163–177, 2005.

[12] A. Hundepool. The casc project. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases: From Theory to Practice*, volume 2316 of *Lecture Notes in Computer Science*. Springer-Verlag, 2002.

[13] S.K. Kinney, J.P. Reiter, A.P. Reznek, J. Miranda, R.S. Jarmin, and J.M. Abowd. Towards unrestricted public use business microdata: The synthetic longitudinal business database. Working Paper 11–04, Center for Economic Studies, 2011.

[14] R. Little. Statistical analysis of masked data. *J Off Stat*, 9:407–426, 1993.

[15] N.Shlomo and T. De Waal. Protection of micro-data subject to edit constraints against statistical disclosure. *J Off Stat*, 24:1–26, 2008.

[16] Office for National Statistics. `www.statistics.gov.uk`.

[17] Office of Information and Regulatory Affairs. Statistical policy working paper 22 - report on statistical disclosure limitation methodology. Subcommittee on Disclosure Limitation Methodology, Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget., 1994.

[18] C.M. O'Keefe. Remote analysis in action - design and implementation of a demonstration remote analysis system. In *Proc New Techniques and Technologies in Statistics NTTS 2011, Brussels*. Eurostat, 22–24 Feb 2011. Available at www.ntts2011.eu.

[19] C.M. O'Keefe. Confidentialising exploratory data analysis output in remote analysis, 2012. Preprint.

[20] C.M. O'Keefe and N.M. Good. Regression output from a remote analysis system. *Data Knowl Eng*, 68:1175–1186, 2009.

[21] T.E. Raghunathan, J.P. Reiter, and D.R. Rubin. Multiple imputation for statistical disclosure limitation. *J Off Stat*, 19:1–16, 2003.

[22] J.P. Reiter. Model diagnostics for remote-access regression systems. *Stat Comput*, 13:371–380, 2003.

[23] J.P. Reiter. Releasing multiply imputed, synthetic public-use microdata: An illustration and empirical study. *J Roy Stat Soc A Sta*, 168:185–205, 2005.

[24] J.P. Reiter. Using cart to generate partially synthetic public use microdata. *J Off Stat*, 21:441–462, 2005.

[25] F. Ritchie. Disclosure detection in research environments in practice. In *Joint UNECE/Eurostat*

work session on statistical data confidentiality, number WP. 37 in Topic (iii): Applications, Manchester, UK, 17–19 December 2007. United Nations Statistical Commission and Economic Commission for Europe Conference of Europe Statisticians, European Commission Statistical Office of the European Communities (Eurostat).

[26] D.B. Rubin. Discussion: Statistical disclosure limitation. *J Off Stat*, 9:462–468, 1993.

[27] R. Sparks, C. Carter, J. Donnelly, C.M. O'Keefe, J. Duncan, T. Keighley, and D. McAullay. Remote access methods for exploratory data analysis and statistical modelling: Privacy-preserving Analytics^{TM}. *Comput Meth Prog Bio*, 91:208–222, 2008.

[28] P. Sutcliffe, M. Caruso, and H. Teasdale. Issues associated with producing a longitudinal dataset of businesses. Research Paper, Methodology Advisory Committee 1352.0.55.062, Australian Bureau of Statistics, Statistical Services Branch, Canberra, 2004. 32pp.

[29] United States Census Bureau. website. `http:www.census.gov`.

[30] L. Willenborg and T. de Waal. *Elements of Statistical Disclosure Control*, volume 155 of *Lecture Notes in Statistics*. Springer, 2001.