Journal of Privacy and Confidentiality (2013) **5**, Number 1, 107–135

On Regression-Tree-Based Synthetic Data Methods for Business Data

Joo Ho Lee*, In Yong Kim†, Christine M. O'Keefe‡

1 Introduction

The challenge of balancing the competing objectives of allowing statistical analysis of confidential data and maintaining confidentiality is of great interest to national statistical agencies and other data custodians seeking to make their data available for research. This balance is often characterised as a trade-off between disclosure risk and data utility, where *disclosure risk* attempts to capture the probability of a data release resulting in a disclosure, while *data utility* attempts to capture some measure of the usefulness of the released data, see [6]. To date, most of the literature on addressing this balance has focussed on data about individuals, however, the same problem arises in the context of data about businesses and enterprises. It is the purpose of this paper to provide an empirical evaluation of existing methodology for individual data being applied to business data.

Several national statistical agencies have started to explore the option of releasing only *synthetic* data comprising original records with sensitive values replaced with synthetic values, see [21] and [12]. As discussed by [5], the model used to generate the synthetic values is fundamental to the performance of synthetic data. Current practice for generating synthetic data typically employs sequential modelling strategies based on parametric or semi-parametric models similar to those for imputation of missing data in [16]. Unfortunately, the process of estimating effective statistical models needed for the synthesis can be a difficult and labour-intensive task. To avoid this difficulty, it has recently been proposed that statistical model estimation could be replaced by easily-implemented methods from machine learning, as has been done in missing data imputation (see [11] and [3]).

[5] have conducted a simulation study to compare four data synthesisers based on machine learning algorithms, using a subset of the 2002 Uganda census public use file. Among these, they found the CART synthesiser to be best suited as a general-purpose, low-cost approach to generating partially synthetic data with good utility and acceptable disclosure risk.

The different characteristics of business data compared with population census and survey data give rise to the possibility that synthesisers may have a different effect. In this paper we explore the applicability of the CART synthesiser for synthetic business

*Department of Statistics, Korea University, mailto:joe.jooholee@gmail.com.
†Department of Statistics, Korea University, mailto:kiykk@korea.ac.kr.
‡Commonwealth Scientific and Industrial Research Organization (CSIRO), Canberra, Australia, mailto:Christine.OKeefe@csiro.au.

data through an empirical evaluation of the use of two variants of it for an example of a business data set. We compare the synthesisers' performance on both disclosure risk and data utility in comparison with traditional statistical disclosure control and remote analysis methods.

In particular, we give a detailed example enabling a comparison of the outputs of exploratory data analysis and linear regression under two variants of the CART synthesiser, as well as an evaluation with respect to analysis of the original data. We use the same sample business data set and analyses used by [15], which then enables a direct comparison of traditional statistical disclosure control (SDC), remote analysis, and synthetic data approaches for confidentiality protection.

While it is certainly true that any method designed to protect confidentiality introduces error, and may indeed give misleading conclusions, our analysis of the results for synthesisers based on CART models has provided some evidence that this error is not random but is due to the particular characteristics of business data. We conclude that more careful analysis needs to be done in applying these methods and end users certainly need to be aware of possible discrepancies.

The remainder of the paper is structured as follows. In Section 2 we describe the particular characteristics of business data and give an overview of the approaches currently implemented by national statistical agencies making business data available for research. Section 3 describes the sample data set we will use and outlines our construction of synthetic versions of it. Sections 4, 5, and 6 contain the main results of the paper. They provide a comparison of exploratory data analysis and linear regression results for the original data and the two synthetic data sets. Section 6 also reproduces some of the results of [15] for linear regression under the SDC and remote approaches, for easier comparison. In Section 7 we explore the impact of different algorithm stopping criteria on the results, and we conclude with a discussion in Section 8.

2 Business Data and Confidentiality

Before discussing confidentiality of business data in particular, we introduce some terminology to be used in the paper.

Traditional approaches to balancing data use with confidentiality protection include *statistical disclosure control (SDC)* methods such as rounding, swapping, or deleting values and adding random noise to data, see [7] and [10]. Under the synthetic approach, as in Section 1, no actual microdata but only synthetic data are released. A remote analysis system accepts a query from an analyst, runs it on data held in a secure environment, then returns confidentialised results, see for example, [9], [18], and [23].

2.1 Privacy, confidentiality and commercial sensitivity

Much of the literature on statistical disclosure control, synthetic data, and remote analysis assumes the context of census or survey data about individual persons. In this case,

the data custodian usually has an obligation to keep the data confidential, in order to protect the individuals' privacy. Here *privacy* can be understood as the interest a person has in controlling the dissemination of information about themselves, whereas *confidentiality* can be understood as the expectation on a data custodian not to disseminate this information.

In this paper, we are concerned with business data, which are often developed to facilitate analysis at the microeconomic level of a range of policy issues based around business growth and performance. Such data usually contain commercially sensitive information about individual businesses, for example, production, employment, customs, financial, and tax data. Although the concept of privacy is not relevant, the businesses have an interest in controlling the dissemination of information about themselves, and the data custodian usually has an obligation to keep the data confidential.

Thus, data custodian agencies usually have a responsibility to protect the confidentiality of business data as well as household and individual data, although for the different reason of commercial sensitivity rather than individual privacy. In practice the objective is the same—not to reveal confidential information in a given data set—and therefore the same range of approaches is available. On the other hand, business data are quite different in nature from census and survey data, and the particular characteristics of business data mean that it is more challenging to protect business data confidentiality.

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 sample inclusion probabilities, which heighten the confidentiality issues for large businesses:
   (a) An individual large business would be always included. Thus business data often include a census of large businesses.
   (b) An individual medium-sized business would be frequently included.
   (c) An individual small business would be seldom included. Note that there is a very large number of small businesses so the sample contains a sufficient number of them, so as to not raise confidentiality concerns.
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 have the characteristics of being outliers on each of many variables. These are the large businesses in the industry sector or sectors sampled.

The consequences of this difference were highlighted in a 2006 survey of OECD countries [1], which 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

concluded 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."

It is the purpose of this paper to investigate whether the differences in data set characteristics lead to differences in the applicability of regression-tree-based synthetic data methods to business data compared with census data.

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

In this section we provide a brief overview of 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. This overview is provided only as examples of policy options, since none of the data we use in this paper comes from any of these agencies. We use only publicly-available information found on the agencies' websites, which has led to some inconsistency between the depth of information provided in the following three subsections.

2.2.1 Australian Bureau of Statistics

The Australian Bureau of Statistics (ABS) releases confidentialised 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.[1] The ABS CURFs contain confidentialised 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 conduct statistical analysis.

As noted in [24], "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 data set) and
  - Some data perturbation occurs to maintain both the confidentiality and structure of the data set."

[1] See `www.abs.gov.au`.

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."

2.2.2 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.[2] 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 and (b) the specific need for the data sets requested.
4. Access is only granted through BDL's secure microdata lab on site at ONS premises.

A research project must specify which data set it wants to use, why it wants to use it, and why the data cannot be found elsewhere. Additional data sets 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.

2.2.3 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).[3] Researchers with approved RDC projects gain restricted access to selected internal microdata from the Census Bureau and other statistical agencies for statistical purposes only. In addition, the CES research program develops public-use business data products by combining and enhancing existing data. The US Department of Labor Bureau of Labor Statistics (BLS) also has an onsite researcher program at the BLS national office in Washington, D.C.[4] The program allows access to confidential microdata to eligible researchers for approved statistical analysis.

To protect the confidentiality of the data, researchers are not permitted to store confidential data files on their own computer equipment. The BLS provides an IBM-compatible personal computer (PC) subject to the following four conditions:

[2]See the website at `www.statistics.gov.uk`.
[3]See the website at `www.census.gov`.
[4]See the website at `www.bls.gov/bls/blsresda.htm`.

1. The PC is either stand-alone or connected to an internal server where the confidential data files are stored.
2. Each PC is equipped with PC SAS. The server is equipped both with SAS and Stata statistical software. On a case-by-case basis, a researcher may be permitted to load his or her own statistical software onto the PC or server provided that licensing agreements are observed.
3. BLS PCs do not provide access to e-mail or the Internet.
4. BLS staff review all printouts, disks, reports, and other project outputs to ensure that data confidentiality is protected. Although researchers are allowed to bring microdata into the BLS national office for merging to confidential BLS microdata, any resulting merged microdata sets (with or without identifiers) are considered confidential and must remain onsite at the BLS national office.

Visiting researchers may bring their own laptops to conduct outside work and to connect to the internet.

2.3 Potential applicability of the synthetic data approach to business data

As discussed above, it is primarily the large businesses that cause confidentiality concerns because large business records are usually outliers on each of many variables, and so the business itself is recognisable and identifiable from the data. Most SDC approaches do not provide sufficient protection to outliers, so all outliers are removed from the data, and this is indeed how large businesses are often treated in business data (see Section 2.2.1). Therefore, released business data sets often only cover small and medium sized enterprises, and conclusions can only be drawn for this restricted part of the business sector.

It has been suggested that remote analysis servers may overcome this difficulty, since analysis is conducted on the original data which includes the large businesses. On the other hand, remote analysis server outputs cannot be assumed to be non-disclosive and so need to also undergo confidentialisation including removal of outliers from plots (see, for example, [13, 14, 23]. A comparison of the impact of SDC and remote analysis methods on exploratory data analysis and linear regression conducted on a business data set is found in [15]. The results suggest that remote analysis may provide generally more accurate exploratory data analysis and regression results than SDC, provided the analyst understands the output confidentialisation methods and their potential impact.

It is the purpose of the current paper to continue the effort to understand the impact on statistical analysis of business data of different confidentialisation treatments, and in particular to expand the investigation to include partially synthetic data constructed from the same original data set. For the comparison, we retain all the large businesses in the partially synthetic data, and synthesise all sensitive variables.

Region	No. of farms
1	123
2	38
3	89
4	88

Table 1: Number of farms in each region in the Sugar Farms data.

3 Sugar Farms Data

In this section we describe the original data set we will use, and outline our construction of synthetic versions of it. We provide a brief overview of the application of SDC and remote analysis methods to the data as in [15]. We show that the different confidentialisation approaches have similar values on an approximate measure of disclosure risk, so that comparisons of their utility can reasonably be made.

3.1 Original Sugar Farms data

We will use the *Sugar Farms* data from a 1982 survey of the sugar cane industry in Queensland, Australia, see [4]. The data set 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 data set 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). The variable profit is calculated as the difference between receipts and costs and is not considered further in this paper. There are no missing values.

The sugar farms are distributed across the regions as shown in Table 1. The variation in the number of farms across the regions will become important in Section 3.5.

The Sugar Farms data set displays many of the characteristics of business data, as discussed in Section 2.1, namely: it has few variables, most of which are continuous with highly skewed distributions, and it has five large farms, all of which are outliers on most of the variables.

3.2 Statistical disclosure control on the Sugar Farms data

In this subsection we give a very brief introduction to the SDC approach to confidentialising the Sugar Farms Data, for details see [15]. Our method 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.

In official statistics data sets such as the Sugar Farms Data, it is important to preserve additivity constraints in the data set, 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, see [22]. However, it is important to note that the process preserves the properties of the data set *without the large farms*, as these are removed prior to the addition of noise.

3.3 Remote analysis on the Sugar Farms data

In this subsection we give a brief introduction to the remote analysis approach to confidentialising the Sugar Farms Data, for details see [15].

Under remote analysis, 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.

3.4 Synthetic Sugar Farms data

We use Classification and Regression Trees (CART) to construct partially synthetic Sugar Farms data, see [5], [19], and [2]. CART is a tool for estimating the conditional distribution of a univariate outcome given multivariate predictors. It is a non-parametric recursive decision tree learning technique, which partitions the predictor space so that subsets of units formed by the partition have relatively homogeneous outcomes. CART induces a binary split of the predictor space at each step, and if the splits are represented in a tree then a strictly binary tree is generated with leaves corresponding to the subsets of units formed by the partition. In the synthesiser, the values in each leaf represent the conditional distribution of the outcome variable for units with predictor variable values satisfying the partitioning criteria that define the leaf.

Let $D = (Y_1, \ldots, Y_d)$, where $Y_i = (Y_{1i}, \ldots, Y_{ni})$, denote a data set with observations

Y_{ij} on n units and d variables. In order to synthesise all values of one variable Y_i in a data set given values of all the other variables Y_{-i}, we first fit a tree TREE_i for outcome Y_i and predictors Y_{-i}. If Y_i is categorical then each split is chosen to minimise the Gini index in the child nodes [8], while if Y_i is continuous then each split is chosen to minimise the deviance of Y_i in the child nodes. Splitting ceases when the Gini index or deviance at a node is less than some custodian-specified threshold or the number of units in a child node is less than some other custodian-specified threshold k. These two thresholds enable the custodian to control the trade-off between disclosure risk and data utility.

Once the tree TREE_i is fitted, for any record in the data set we can follow the split criteria down through the nodes until we find that record's terminal leaf. Let L_w be the wth terminal leaf in TREE_i, and let $Y^i_{L_w}$ be the n_{L_w} values of Y_i in leaf L_w. For the jth record in the data set, $j = 1, \ldots, n$, we find the corresponding terminal leaf L_w and generate a synthetic value for Y_{ij} by drawing from $Y^i_{L_w}$ with selection probabilities generated by the Bayesian bootstrap [20]. The synthesising process is repeated m times to generate m data sets with synthetic values of Y_i.

In order to synthesise all values of two or more variables, choose an ordering so that the variables to be synthesised are $Y_1, \ldots, Y_s$. First use the above algorithm for a single variable on the data set comprising the outcome Y_1 and predictors $Y_{s+1}, \ldots, Y_d$, to obtain synthetic values Y'_1 for Y_1. Next, use the algorithm for a single variable on the data set comprising the outcome Y_2 and predictors $Y'_1, Y_{s+1}, \ldots, Y_d$, to obtain synthetic values Y'_2 for Y_2. Continue this process until all values of $Y_1, \ldots, Y_s$ have been replaced by synthetic values $Y'_1, \ldots, Y'_s$.

In the case of the Sugar Farms data, we used the above procedure to synthesise all values of the sensitive variables: area, harvest, costs, and receipts. Region is not sensitive and was not synthesised, giving rise to partially synthetic data. We chose the threshold $k = 5$ for the minimum number of records in each leaf, as it provided sufficient accuracy and reasonably fast running time, as well as protecting confidentiality by ensuring that at least 5 units contribute to any synthetic value. Further details about this choice, including a comparison with results for $k = 2$ and $k = 7$, is provided in Section 7. We also imposed a maximum threshold of 10 records per leaf, in order to ensure the synthetic data would be a good fit to the original data. Further, we ordered the variables using the method described in [19] and constructed a number m synthetic data sets. We used $m = 10$ data sets for the statistical analyses, however, we constructed $m = 100$ synthetic data sets to investigate the approximate disclosure risk in Section 3.5 and the significance of regression coefficient for the variable area in Section 6.1.

In this data set, the variable region was not considered sensitive and has not been synthesised. On the other hand, farm location could be expected to have an impact on farm harvest, costs, and receipts, and even possibly farm area, so we decided to also investigate the potential role of the variable region on the data set syntheses. We therefore constructed two types of synthetic data sets. In the first, we separated the units into four sub-data sets by region and synthesised the sub-data sets separately before reassembling the full data set. In the second, we conducted the synthesis on the whole data set. Since the two methods gave different estimates and inferences, we

provide the results of both for comparison.

3.5 Approximate disclosure risk

In the process of confidentialising a data set, original values are changed to confidentialised or protected values. One measure of the overall difference between the confidentialised data set and the original one is the sum of the relative absolute differences between the variable values in the confidentialised data set and the corresponding variable values in the original data set. The larger the value of this measure, the more different is the confidentialised data set overall from the original one, and the more difficult it would be expected to be, overall, to extract confidential information from it. We therefore propose to use this measure as an approximation to disclosure risk, see also [15].

Since the variable region was not confidentialised, the disclosure risk measure is calculated for each of the four regions separately. There is a slight complication regarding the treatment of the large farms. The large farms were removed from the data under the SDC approach, for reasons explained in Section 2.3, which means that they were not included in the calculation of the disclosure risk measure. In order to ensure comparability of the disclosure risk measure across the approaches, the large farms were also omitted from the calculation of the disclosure risk measure for the remote analysis and synthetic data approaches, even though they would be included in the analysis itself. The process for calculating approximate disclosure risk for a variable v is as follows:

First, remove the large farms from the data and for $r = 1, \ldots, 4$, let N_r be the number of farms remaining in Region r. The approximate disclosure risk for variable v in Region r is given by the formula:

$$\sum_{i=1}^{N_r} |cv_i - v_i| / v_i$$

where v_i is the value of variable v for farm i in the original data and cv_i is the value of variable v for farm i in the confidentialised data.

Table 2 provides a comparison of the approximate disclosure risk measures for the SDC, remote analysis, and synthetic approaches. Given that the approximate disclosure risk measure is a sum of 333 relative absolute values, it is noticeable that the values across the regions, variables, and approaches all fall within a small range of values from around 2 to around 21. In each region, the variability is reduced further, with Region 1 values ranging from around 10 to around 21, Region 2 values ranging from around 2 to around 7, Region 3 values ranging from around 7 to around 20 and Region 4 values ranging from around 5 to around 17. The difference in the ranges in each region is most likely due to the different numbers of farms in each region. For example, Region 1 has about three times as many farms as Region 2, and the values of the disclosure risk measures are also larger. This investigation suggests that the approaches under the four measures of SDC, remote analysis, and synthetic data with region pooled or separate all give measures that are similar in magnitude within a region and therefore we can

		Approach to Confidentiality Protection			
		SDC	Remote Analysis	Synthetic data	
				Region pooled	Region separate
Region 1	Harvest	16.80	12.94	10.80	11.30
	Receipts	14.39	14.50	14.45	11.36
	Costs	21.28	8.55	17.25	16.27
Region 2	Harvest	3.82	2.04	2.99	3.23
	Receipts	3.28	5.59	3.74	4.18
	Costs	5.66	3.63	5.77	6.75
Region 3	Harvest	14.82	5.80	7.67	9.38
	Receipts	9.58	7.84	8.76	8.79
	Costs	20.33	5.73	15.29	20.83
Region 4	Harvest	12.20	4.16	7.03	5.95
	Receipts	9.19	4.78	8.13	6.40
	Costs	16.56	6.51	13.40	13.60

Table 2: Approximate disclosure risk for sensitive variables harvest, receipts and costs within regions (not including large farms) for the different confidentiality approaches

assume that the confidentialised data sets constructed under the different approaches provide similar confidentiality protection.

Recall that the values in Table 2 are the averages of the values on $m = 100$ data sets. To give a better indication of the distribution of the disclosure risk values, histograms of the individual values are provided in Figure 1.

4 Univariate exploratory data analysis of the Sugar Farms data

In this section we provide a comparison of univariate exploratory data analysis results for the original data and the two types of synthetic Sugar Farms data, in Figure 2. The results for the synthetic data are based on a random selection from the $m = 10$ generated data sets. We focus on the variable area, as it is representative of the sensitive variables, and note that the results are directly comparable with those for SDC and remote analysis found in [15]. To be precise, Figures 2(a) and 2(b) for the original data are identical to

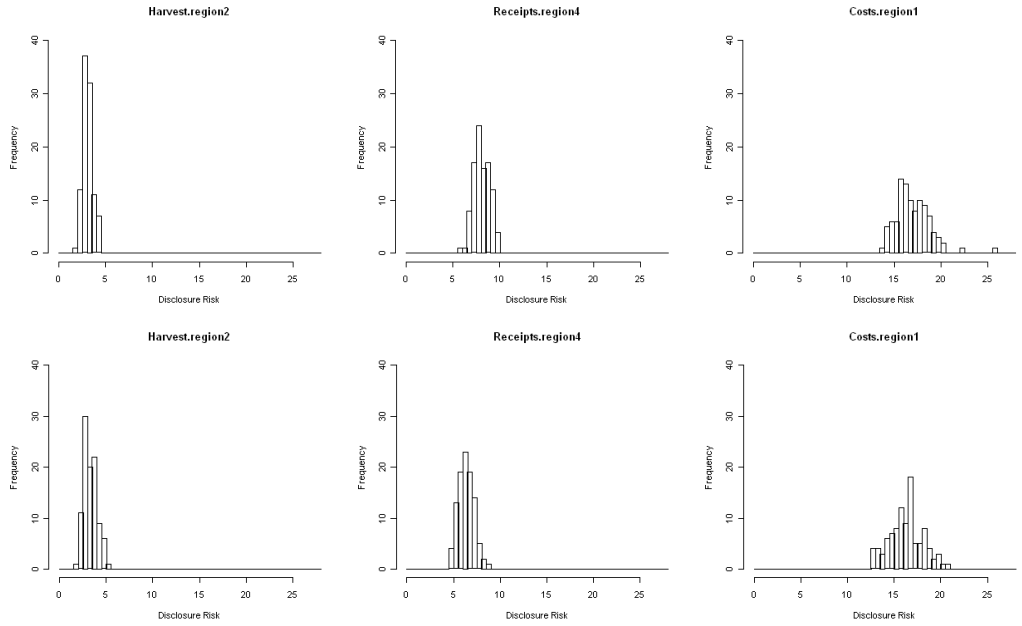


Figure 1: Histograms of approximate disclosure risk values for sensitive variables harvest on Region 2, receipts on Region 4 and costs for Region 1, for synthetic data with region pooled (upper row) and separate (lower row)

O'Keefe and Shlomo [15, Figures 6(c) and 6(d)], the corresponding results for SDC are in O'Keefe and Shlomo [15, Figure 4] and the corresponding results for remote analysis are in O'Keefe and Shlomo [15, Figure 5].

The histograms and densities for the synthetic data sets in Figures 2(c) and 2(e) give the analyst quite good information about the shape of the variable distribution, when compared with the original data in Figure 2(a).

Similarly, the normal Q-Q plots for the synthetic data sets in Figures 2(d) and 2(f) both indicate a departure from normality similar to that observed for the original data in Figure 2(b). Each plot is curved with slope increasing from left to right, implying that each distribution is skewed to the right. This indicates that a transformation of the variable, such as log-transformation, would give a better result in analyses that assume normality.

In summary, the synthetic data sets provide good information about the distribution of the variable area, in comparison with the original data.

5 Bivariate Exploratory Data Analysis of the Sugar Farms Data

In this section we provide a comparison of bivariate exploratory data analysis results for the original data and the two synthetic Sugar Farms data sets. The same sample synthetic data set as in Section 4 is used for the synthetic data boxplots and scatter plots. For the correlation analysis on the synthetic data in Section 5.2, we used $m = 10$ synthetic data sets and applied the methods of [17] to estimate each correlation statistic as the average of the estimates calculated with standard procedures on each of the $m = 10$ data sets. The corresponding significance is calculated using the underlying distribution of the average of estimates.

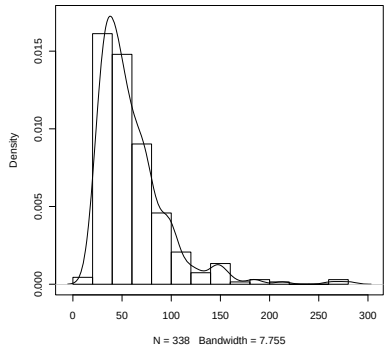
Again, the results are directly comparable with those for SDC and remote analysis found in [15].

5.1 Bivariate: Area and costs with region

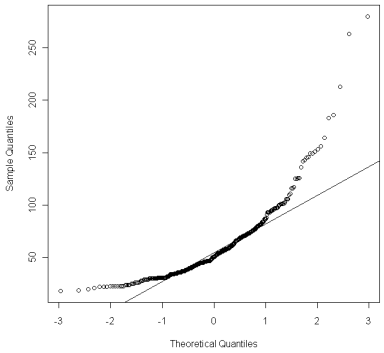
Output from bivariate exploratory data analysis for area with region and costs with region is shown for the original and synthetic data sets with region separate and pooled in Figure 3. For comparison, results for area and costs with region under SDC and remote analysis are found in O'Keefe and Shlomo [15, Figures 10 and 11], respectively.

For both area and costs, the plots in Figure 3 are quite similar. For area, the boxplots for synthetic data with region separate (Figure 3(c)) mostly have fewer outliers than those for the region pooled synthetic data (Figure 3(e)). For costs, the synthetic data with region pooled (Figure 3(f)) has some undesirable outliers in Region 1.

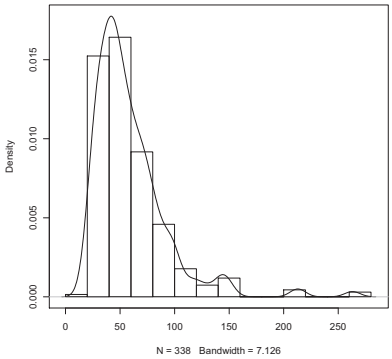
In summary, the boxplots for the synthetic data with region pooled provide good



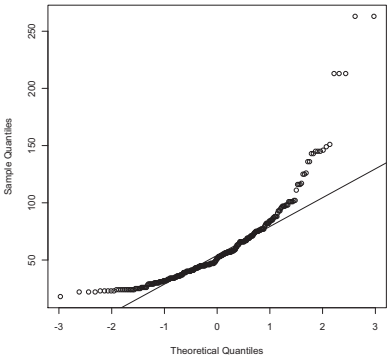
(a) Histogram and Density - original



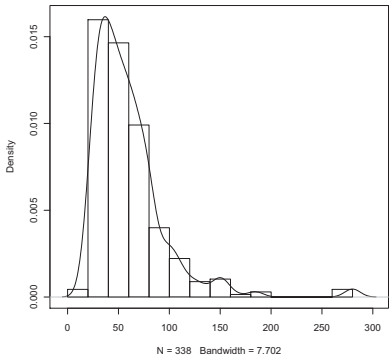
(b) Normal QQ-plot - original



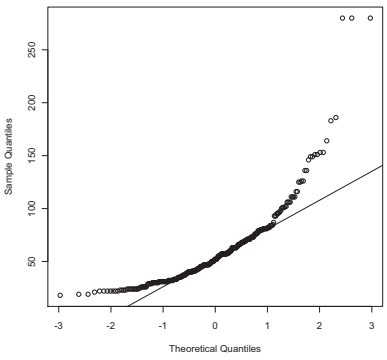
(c) Histogram and Density - synthetic, region separate



(d) Normal QQ-Plot - synthetic, region separate

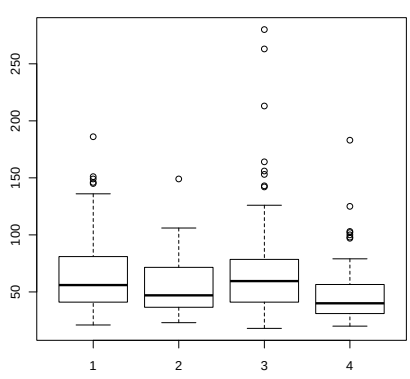


(e) Histogram and Density - synthetic, region pooled

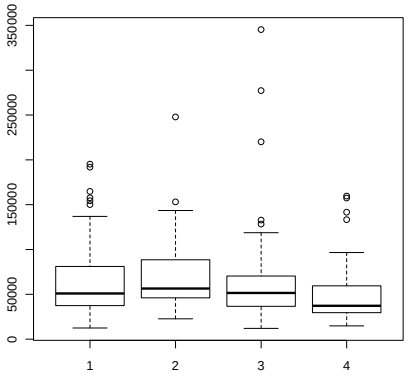


(f) Normal QQ-Plot - synthetic, region pooled

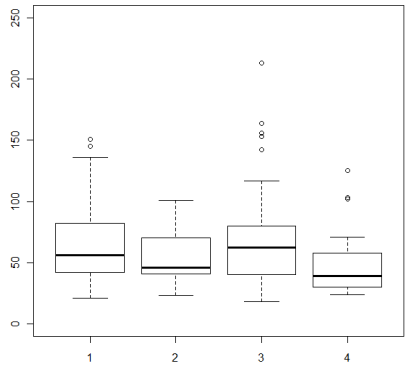
Figure 2: Univariate exploratory data analysis output for the variable area in the Sugar Farms data, for the original data, synthetic data with region separate, and synthetic data with region pooled.



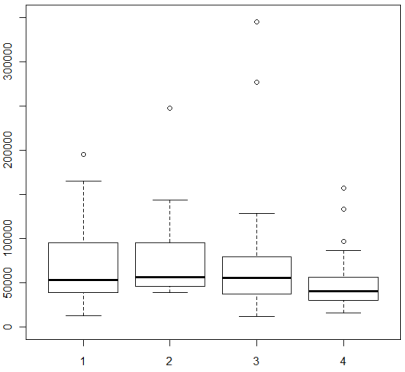
(a) Box plots for area by region - original



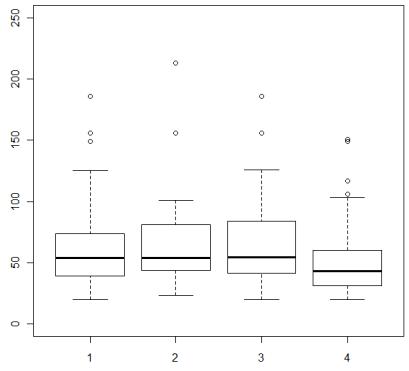
(b) Box plots for costs by region - original



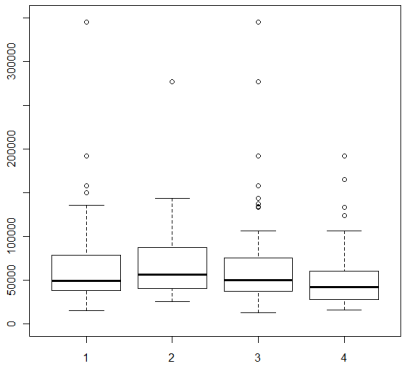
(c) Box plots for area by region - synthetic, region separate



(d) Box plots for costs by region - synthetic, region separate



(e) Box plots for area by region - synthetic, region pooled



(f) Box plots for costs by region - synthetic, region pooled

Figure 3: Bivariate exploratory data analysis output for each variable area and costs with region in the Sugar Farms data, for the original data, synthetic data with region separate, and synthetic data with region pooled.

		Original	Synthetic data	
			Region pooled	Region separate
receipts vs area	Pearson p-value	0.888 $< 2.2e^{-16}$	0.882 $< 2.2e^{-16}$	0.773 $< 2.2e^{-16}$
	Chi-sq p-value	347 < 0.0001	398 < 0.0001	348 < 0.0001
	Cramer V	0.453	0.485	0.454
costs vs area	Pearson p-value	0.887 $< 2.2e^{-16}$	0.895 $< 2.2e^{-16}$	0.722 $< 2.2e^{-16}$
	Chi-sq p-value	361 < 0.0001	428 < 0.0001	429 < 0.0001
	Cramer V	0.462	0.503	0.504

Table 3: Pearson Correlation Coefficients, chi-square statistics, and Cramer's V statistic values for receipts and costs with area and for the original Sugar Farms data and the synthetic Sugar Farms data with region separate or pooled.

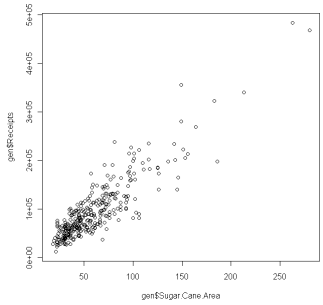
central information about the variable distributions but tend to show more outliers.

5.2 Bivariate: Pairs from area, receipts and costs

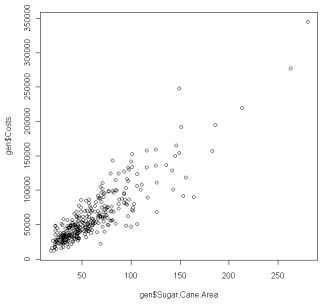
Output from bivariate exploratory data analysis for receipts with area and costs with area is shown for the original and synthetic data sets with region separate and pooled in Figure 4. For comparison, results for pairs from area, receipts, and costs under SDC and remote analysis are found in O'Keefe and Shlomo [15, Figures 13 and 15], respectively.

The Pearson Correlation Coefficients, chi-square test statistics, and values of Cramer's V statistic for receipts and costs with area and for the different approaches are shown in Table 3. For comparison, the corresponding results under SDC are found in O'Keefe and Shlomo [15, Figures 13(d) and 14]. Corresponding Pearson Correlation coefficients for remote analysis are found in O'Keefe and Shlomo [15, Figure 15(d)] while chi-square statistics and values of Cramer's V are not provided under remote analysis.

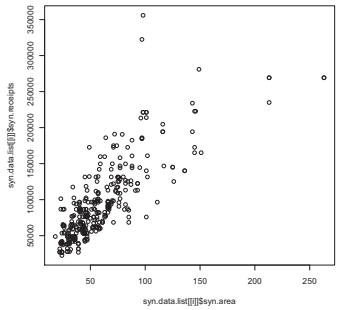
The relationships between the variables receipts and costs with area in the original data, as observed in Figures 4(a) and 4(b), are also generally observed in the two synthetic data sets in Figures 4(c) and 4(d), and 4(e) and 4(f), respectively. However, comparing the Pearson Correlation Coefficients in Table 3 shows that the synthetic data



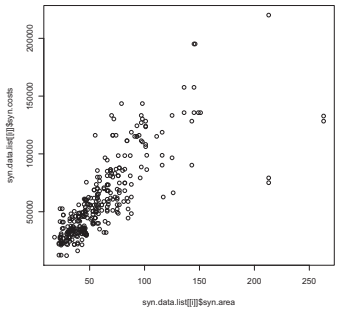
(a) receipts by area - original



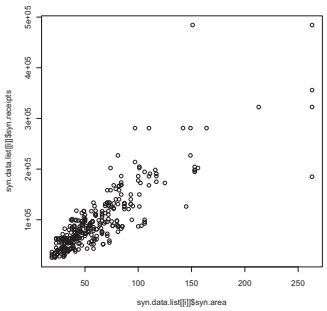
(b) costs by area - original



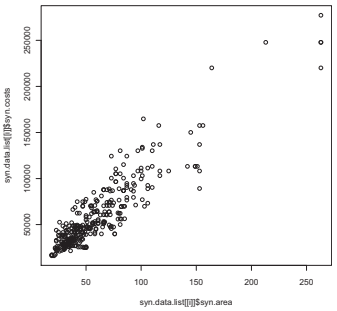
(c) receipts vs area - synthetic, region separate



(d) costs vs area - synthetic, region separate



(e) receipts vs area - synthetic, region pooled



(f) costs vs area- synthetic, region pooled

Figure 4: Bivariate exploratory data analysis output for receipts and costs with area on the Sugar Farms data, for the original data, synthetic data with region separate, and synthetic data with region pooled.

set with region pooled produces more highly correlated variables than the synthetic data set with region separate, and so is more similar to the original data set in that respect. This observation might be because the trees constructed by CART on separate regions may not be deep enough to adequately represent the correlations. This could potentially be an issue for other data sets which display the same characteristic. For the chi-square test, all approaches provide the same decision in terms of rejecting the null hypothesis of independence at a given significance level.

In summary, both the original and synthetic data seem to provide fairly similar bivariate information about these pairs of variables, except that the synthetic with region separate CART underestimates the correlation amongst the variables.

6 Regression Analysis

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 is 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.

The regression is conducted in the traditional manner on the original Sugar Farms data, and under the SDC and remote analysis approaches. For the regression analysis on the synthetic data in Section 5.2, we used $m = 10$ synthetic data sets and applied the methods of [17] to estimate each correlation statistic as the average of the estimates calculated with standard procedures on each of the $m = 10$ data sets. The corresponding significance is calculated using the underlying distribution of the average of estimates.

As for the exploratory data analysis results, the output is directly comparable with the output for SDC and remote analysis found in [15], some of which is reproduced in the tables below for comparison.

6.1 Summary results

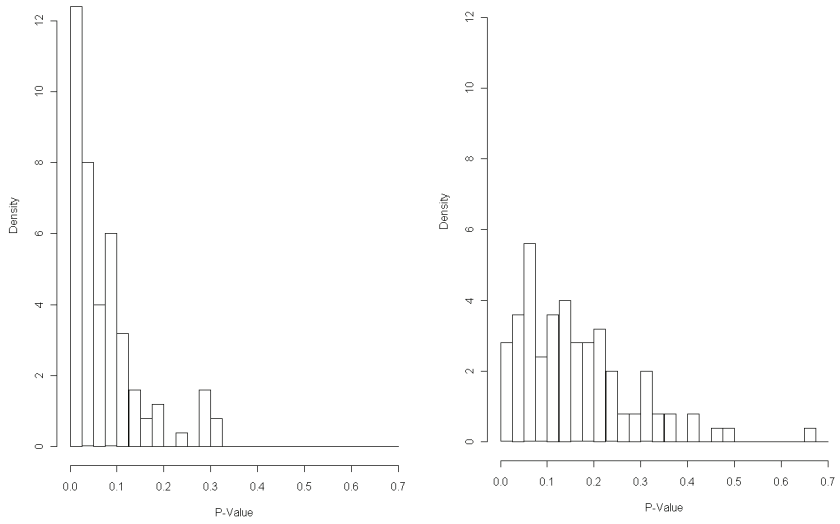
Table 4 shows the regression coefficients and their significance values for the SDC, remote analysis approaches, as well as the original and two synthetic data sets.

The most important observation is that the synthetic data with region pooled underestimates the significance of each of region(2), region(4), and costs. On the other hand, the synthetic data with region separate incorrectly indicates that area is significant, an error also seen in the SDC-generated data set. There are also differences observed in the values of the coefficients, with the synthetic data with region pooled underestimating the coefficients for each region and log(costs), while the synthetic data with region separate significantly underestimates the coefficient of log(harvest).

	SDC	Remote analysis	Original	Synthetic data (m = 10)	
				Region pooled	Region separate
Intercept	3.63	3.06	2.71	2.98	2.99
p-value	< 2e-16		< 2e-16	< 2e-16	< 2e-16
significance	***	***	***	***	***
Factor(region)2	0.193	0.205	0.181	0.0880	0.215
p-value	2.97e-15		< 2e-16	1.56e-03	1.59e-10
significance	***	***	***	*	***
Factor(region)3	0.188	0.244	0.239	0.0771	0.249
p-value	< 2e-16		< 2e-16	1.32e-04	< 2e-16
significance	***	***	***	***	***
Factor(region)4	0.091	0.117	0.118	0.0536	0.142
p-value	1.91e-7		< 2e-16	1.01e-02	1.57e-10
significance	***	***	***	*	***
area	0.0312	0.0004	0.0000792	0.000572	0.00113
p-value	4.81e-6		0.773	1.78e-01	1.36e-02
significance	***				*
harvest	0. 832	0.883	0.866	0.903	0.733
p-value	< 2e-16		< 2e-16	< 2e-16	< 2e-16
significance	***	***	***	***	***
costs	0. 0631	0.0823	0.131	0.0817	0.197
p-value	0.0147		4.05e-8	2.19e-02	1.84e-07
signficance	*	***	***	.	***

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

Table 4: Coefficient estimates and significance levels for linear regression of log(receipts) on region, area, log(harvest), and log(costs) for the approaches: SDC, remote analysis, original data, and synthetic data with region separate or pooled.



(a) p-values for area in synthetic data, region separate

(b) p-values for area in synthetic data, region pooled

Figure 5: Histogram of p-values for the coefficient of area from 100 replications

The observed differences in the regression results occur probably because the CART models underestimate the strong correlation structure between the variables in the data set. The extent of the underestimation seems to depend mostly on the minimum number k of observations in a leaf allowed during the generation of trees for the CART model, and we will investigate the impact of different choices for k in Section 7.

Due to the randomness in the synthetic data generation methods, variation is observed in the values of the coefficients and statistics across the $m = 10$ data sets. This was particularly evident in the p-values for the area coefficient. In order to assist with understanding the potential impact of this effect, in Figure 5 we have provided a histogram of p-values for the coefficient of area from $m = 100$ replications of the process, for synthetic data with region separate or pooled.

We observe that the synthetic data with region separate (Figure 5(a)) has generally smaller p-values and higher significance levels for the coefficient of area than with region pooled (Figure 5(b)). We do not know why this would be the case.

6.2 Overall goodness-of-fit statistics

Values of overall goodness-of-fit statistics for the regression are shown in Table 5.

The R squared and adjusted R squared are smaller for the synthetic data approaches

	SDC	Remote analysis	Original	Synthetic data	
				Region pooled	Region separate
Residual standard error	0.115	0.08	0.0902	0.137	0.142
degrees of freedom	326	314	331	331	331
Multiple R squared	0.955	0.97	0.974	0.940	0.936
Adjusted R squared	0.955	0.97	0.974	0.939	0.934
F-statistic	1160	2100	2070	881	818
degrees of freedom	6 and 326	6 and 331	6 and 331	6 and 331	6 and 331
p-value	< 2.2e-16	-	< 2.2e-16	< 2.2e-16	< 2.2e-16
significance	***	***	***	***	***

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

Table 5: Goodness-of-fit statistics for linear regression of log(receipts) on region, area, log(harvest), and log(costs) approaches: SDC, remote analysis, original data, and synthetic data with region separate or pooled.

than for the other approaches and original data, indicating that the proportion of total variability that is accounted for by the models is smaller for the synthetic data approaches, but the values still indicate high explanatory power for the models.

The F statistics for the synthetic data are smaller even than under the SDC approach, indicating that the level of significance of the model could be smaller, but the corresponding p-values indicate that the models are still significantly valid.

The residual standard errors are greater for the synthetic data approaches than the other approaches and original data, and are greater for the SDC approach than for the original data. The difference between the results for remote analysis and the original data is due to the robust regression procedure implemented in the remote analysis approach. 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 in the synthetic data approaches the predicted values will be further from the synthetic observed values than in the other approaches.

The observed differences in goodness-of-fit statistics across the approaches is probably due to the fact that the SDC and synthetic data approaches introduce more randomness into the data and so weaken the sorts of relationships that are investigated with regression analysis. The implication is that the analyst needs to be aware that relationships in the original data may be stronger than is indicated in the synthetic data, for example, in model selection. In a sense, the values of the goodness-of-fit statistics quantify these differences.

6.3 Model diagnostics

For the regressions, plots of the residuals by the fitted values and normal Q-Q plots of the residuals are provided for the original data and the synthetic data with region separate or pooled in Figure 6. Again, these are the average of the values on the $m = 10$ individual synthetic data sets.

The synthetic data residuals in Figures 6(c) and 6(e) are more spread than the original data residuals in Figure 6(a), consistent with the the smaller R squared values observed in Section 6.2. The residuals do not reveal any meaningful pattern, so the regression models are likely to be valid.

The normal Q-Q plots for the synthetic data (Figures 6(d) and 6(f)) show that the respective regression models satisfy the assumption that the residuals are normally distributed, as also observed in the original data (Figure 6(b)). It is possible that the influence of the large farms or the high leverage observations in the original data set are diminished due to the random process involved in the generation of the synthetic data.

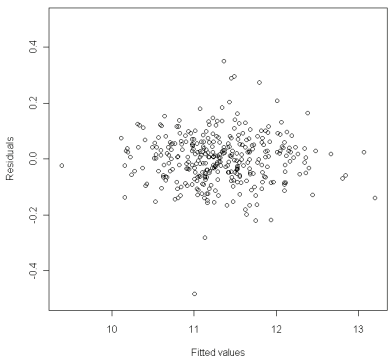
7 Choice of Stopping Criterion

The choice of stopping criterion is very important to the trade-off between the level of confidentiality protection and the degree to which the synthetic data are a good approximation of the original data. To be more precise, a lower value of k (minimum number of units per leaf in the tree) would lead to a greater number of smaller leaves, which would improve the degree to which the synthetic data are a good approximation of the original data. On the other hand, a lower value of k would lead to a lower degree of aggregation in the original data and therefore would tend to reduce the confidentiality protection afforded by the synthetic data.

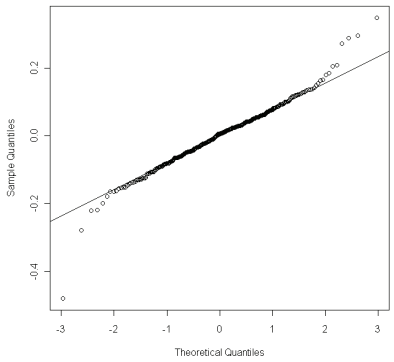
In our study we chose the value $k = 5$ after investigating this trade-off for a range of values. As we have seen, with this stopping criterion the synthetic data underestimated the correlation amongst the variables, and results for linear regression on the synthetic data set led to incorrect conclusions about the significance of variables.

In this section we investigate the effect of the parameter k on the variable correlations and linear regression results to see whether a different value might perform better. In particular, we provide correlation and linear regression results for the choices $k = 2$, $k = 5$, and $k = 7$, in Tables 6, 7, and 8. These are the analogues of Tables 3, 4, and 5. The work reported in this chapter was done as a separate study, hence new synthetic data were generated; therefore table entries in this section may differ from the corresponding table entries in earlier sections.

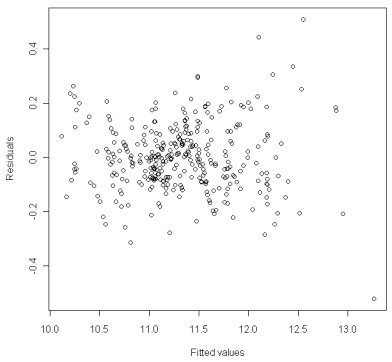
The values of the Pearson Correlation Coefficients in Table 6 generally decrease with increasing k for the same treatment of region, and are lower for region pooled than for region separate for the same value of k. Only the synthetic data set with $k = 2$ and region pooled does not underestimate the correlations amongst the variables as measured by the Pearson Correlation Coefficient.



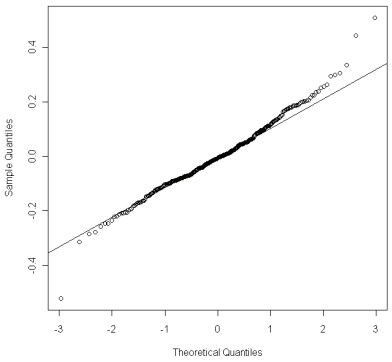
(a) Residuals by fitted values - original



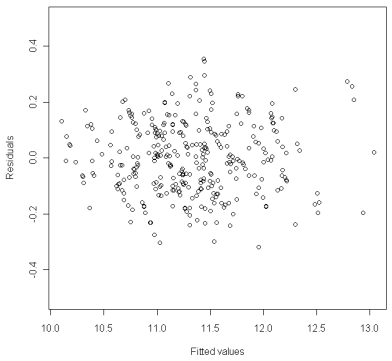
(b) Normal Q-Q Plot of Residuals - original



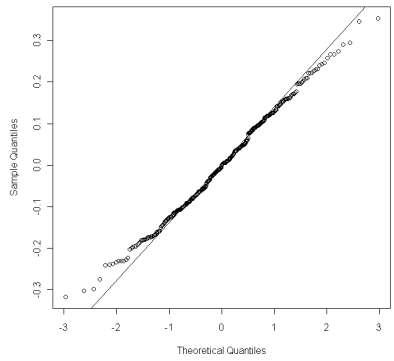
(c) Residuals by fitted values - synthetic, region separate



(d) Normal QQ Plot of Residuals - synthetic, region separate



(e) Residuals by fitted values - synthetic, region pooled



(f) Normal QQ Plot of Residuals - synthetic, region pooled

Figure 6: Diagnostics for linear regression of log(receipts) dependent on region, area, log(harvest), and log(costs) in the Sugar Farms data, for the original data, synthetic data with region separate, and synthetic data with region pooled.

		Original	Synthetic data (k=2)		Synthetic data (k=5)		Synthetic data (k=7)	
			Region pooled	Region separate	Region pooled	Region separate	Region pooled	Region separate
receipts vs area	Pearson p-value	0.888 $< 2.2e^{-16}$	0.887 $< 2.2e^{-16}$	0.862 $< 2.2e^{-16}$	0.849 $< 2.2e^{-16}$	0.848 $< 2.2e^{-16}$	0.857 $< 2.2e^{-16}$	0.733 $< 2.2e^{-16}$
	Chi-sq p-value	347 < 0.0001	357 < 0.0001	351 < 0.0001	378 < 0.0001	398 < 0.0001	375 < 0.0001	365 < 0.0001
	Cramer V	0.453	0.46	0.456	0.473	0.486	0.471	0.465
costs vs area	Pearson p-value	0.887 $< 2.2e^{-16}$	0.885 $< 2.2e^{-16}$	0.854 $< 2.2e^{-16}$	0.861 $< 2.2e^{-16}$	0.846 $< 2.2e^{-16}$	0.86 $< 2.2e^{-16}$	0.787 $< 2.2e^{-16}$
	Chi-sq p-value	361 < 0.0001	375 < 0.0001	354 < 0.0001	396 < 0.0001	399 < 0.0001	384 < 0.0001	381 < 0.0001
	Cramer V	0.462	0.471	0.457	0.484	0.486	0.477	0.475

Table 6: Pearson Correlation Coefficients, chi-square statistics, and Cramer's V statistic values for receipts and costs with area and for the original Sugar Farms data and the synthetic Sugar Farms data with $k = 2, 5$, and 7 and region separate or pooled.

For the chi-square test, all choices provide the same decision in terms of rejecting the null hypothesis of independence at a given significance level.

Table 7 shows that the regression analysis on the synthetic data for each value of k and region pooled underestimates the significance of the variable costs, though the underestimation is greater for $k = 5$ and $k = 7$. In addition, the regression analysis on the synthetic data with $k = 5$ and $k = 7$ and region pooled underestimates the significance of the variable region(2) and region(4). With $k = 7$ and region pooled the significance of the variable region(3) is also underestimated.

The residual standard errors are greater for all the synthetic data approaches than the other approaches (SDC and remote analysis, see Table 5) and original data, so the predicted values will be further from the synthetic observed values than in the other approaches.

The R squared and adjusted R squared are smaller for all the synthetic data approaches than for the other approaches and original data, except that the synthetic data with $k = 2$ has greater values than those for the SDC approach. Therefore, the proportion of total variability that is accounted for by the models is smaller for the synthetic data approaches, but the values still indicate high explanatory power for the models.

The F statistics for the synthetic data are smaller for all the synthetic data ap-

	Original	Synthetic data (k=2)		Synthetic data (k=5)		Synthetic data (k=7)	
		Region pooled	Region separate	Region pooled	Region separate	Region pooled	Region separate
Intercept	2.71	2.94	2.95	2.3	3.07	2.98	3.063
p-value	$< 2e^{-16}$	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$1.7e^{-16}$	$1.08\ e^{-21}$	$< 2.2e^{-16}$	$1.07e^{-11}$
significance	***	***	***	***	***	***	***
Factor(region)2	0.181	0.162	0.2	0.0827	0.213	0.0758	0.203
p-value	$< 2e^{-16}$	$3.72e^{-13}$	$< 2.2e^{-16}$	$2.08e^{-03}$	$6.94\ e^{-14}$	$3.98e^{-03}$	$5.63e^{-08}$
significance	***	***	***	*	***	*	***
Factor(region)3	0.239	0.172	0.241	0.0836	0.228	0.0701	0.242
p-value	$< 2e^{-16}$	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$4.4e^{-05}$	$2.68e^{-23}$	$8.26e^{-04}$	$9.18e^{-16}$
significance	***	***	***	***	***	**	***
Factor(region)4	0.118	0.102	0.123	0.057	0.112	0.0465	0.144
p-value	$< 2e^{-16}$	$1.14e^{-08}$	$5.32e^{-12}$	$5.25e^{-03}$	$1.54e^{-07}$	$2.86e^{-02}$	$6.28e^{-07}$
significance	***	***	***	*	***	.	***
area	0.0000792	0.000353	0.000605	0.00038	0.000518	0.000303	0.00102
p-value	0.773	$3.4e^{-01}$	$1.19e^{-01}$	$3.92e^{-01}$	$3.02e^{-01}$	$4.92e^{-01}$	$7.42e^{-02}$
significance							
harvest	0.866	0.881	0.839	0.907	0.797	0.902	0.683
p-value	$< 2e^{-16}$	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$1.05e^{-36}$	$3.62e^{-48}$	$< 2.2e^{-16}$	$< 2e^{-16}$
significance	***	***	***	***	***	***	***
costs	0.131	0.0979	0.126	0.077	0.146	0.0842	0.228
p-value	$4.05e^{-8}$	$1.74e^{-03}$	$1.18e^{-04}$	$2.77e^{-02}$	$2.97e^{-05}$	$2.9e^{-02}$	$1.15e^{-06}$
signficance	***	*	***	.	***	.	***

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

Table 7: Coefficient estimates and significance levels for linear regression of log(receipts) on region, area, log(harvest), and log(costs) for the original data and synthetic data with $k = 2, 5$, and 7 and region separate or pooled.

	Original	Synthetic data (k=2)		Synthetic data (k=5)		Synthetic data (k=7)	
		Region pooled	Region separate	Region pooled	Region separate	Region pooled	Region separate
Residual standard error	0.0902	0.114	0.116	0.136	0.135	0.136	0.18
Multiple R squared	0.974	0.958	0.957	0.94	0.939	0.94	0.897
Adjusted R squared	0.974	0.957	0.956	0.938	0.938	0.939	0.895
F-statistic	2070	1280	1260	871	865.7048	901	495
p-value	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$< 2.2e^{-16}$	$2.21e^{-199}$	$5.84e^{-199}$	$< 2.2e^{-16}$	$< 2.2e^{-16}$
significance	***	***	***	***	***	***	***

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

Table 8: Goodness-of-fit statistics for linear regression of log(receipts) on region, area, log(harvest), and log(costs) approaches: SDC, remote analysis, original data, and synthetic data with $k = 2, 5$, and 7 and region separate or pooled.

proaches than for the other approaches and original data, except that the synthetic data with $k = 2$ has greater values than those for the SDC approach. Therefore the level of significance of the model could be smaller, but the corresponding p-values indicate that the models are still significantly valid.

The synthetic data set with $k = 2$ and region pooled is the only synthetic data set that does not underestimate the correlations amongst the variables. In addition, the linear regression results for this case are generally comparable with the results on the original data, except that the significance of the variable costs was underestimated. On the other hand, the value $k = 2$ means that synthetic values are drawn from clusters of two records with relatively homogeneous outcomes. In many applications this is likely to be judged as insufficient to protect confidentiality.

8 Discussion and conclusions

In this paper we explored the use of synthetic data in the context of data confidentiality, and in particular for protecting the confidentiality of business data undergoing statistical analysis. We discussed a detailed example enabling a comparison of the outputs of exploratory data analysis under two variants of a synthetic data approach, and evaluation with respect to analysis of the original data. The chosen approach was identified by [5] as being best suited among four non-parametric synthesisers as a general-purpose, low-cost approach to generating partially synthetic datasets with good utility and acceptable risks. The example contributes to developing an understanding of the potential applicability of the synthetic data approach in addressing the balance between microdata access and confidentiality protection for business data, in comparison with other approaches.

In our example of analysis on the Sugar Farms business data set, we obtained similar results for univariate exploratory data analysis on the synthetic data and on the original data. However, this was not the case for bivariate exploratory data analysis and linear regression, probably because the CART models underestimate the strong correlation structure between the variables in the data set. The extent of underestimation seems to depend mostly on the minimum number k of observations in a leaf allowed during the generation of trees for the CART model. We investigated the impact of different choices for k. The synthetic data set with $k = 2$ and region pooled was the only synthetic data set that did not underestimate the correlations amongst the variables. In addition, the linear regression results for this case were generally comparable with the results on the original data, except that the significance of the variable costs was underestimated. On the other hand, the value $k = 2$ means that synthetic values are drawn from clusters of two records with relatively homogeneous outcomes. In many applications this is likely to be judged insufficient to protect confidentiality.

[5] suggested that synthesisers based on regression trees can give rise to synthetic data that provide reliable estimates and low disclosure risks, and are easily implemented. They identified a subset of the 2002 Uganda census public use microdata file to treat as a population, and repeatedly took random samples from it. For each sample, they gener-

ated partially synthetic data sets from each of four types of nonparametric synthesizers and computed point estimates and 95% confidence intervals for 162 estimands spanning representative analyses, including regressions and a variety of population percentages.

Our example seems to indicate that such synthesisers based on regression trees cannot be assumed to be immediately applicable in the context of business data. This may be because business data can have strong correlation structures, a situation exacerbated by the presence in business data of records corresponding to large enterprises that are outliers on each of many variables. These could be expected to make large contributions to correlation measures. As we have already noted, regression trees struggle to adequately capture correlation structure in data sets, suggesting a possible reason for the issues observed.

While it is certainly true that any method designed to protect confidentiality introduces error, and may indeed give misleading conclusions, our analysis of the results for synthesisers based on CART models has provided some evidence that this error is not random but is due to the particular characteristics of business data. We conclude that more careful analysis needs to be done in applying these methods and end users certainly need to be aware of possible discrepancies.

References

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

[2] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth, Inc.

[3] Burgette, L. F. and Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172:1070–1076.

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

[5] Drechsler, J. and Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, 55:3232–3243.

[6] Duncan, G. T., Keller-McNulty, S. A., and Stokes, S. L. (2001). Disclosure risk vs data utility: The R-U confidentiality map. Technical Report LA-UR-01-6428, Los Alamos National Laboratory.

[7] Duncan, G. T., Elliot, M., and Salazar-Gonzàlez, J.-J. (2011). *Statistical Confidentiality*. New York: Springer.

[8] Gini, C. (1912). Variabilitá e mutabilitá. C. Cuppini, Bologna, pp. 156. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1955).

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

[10] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and de Wolf, P.-P. (2012). *Statistical Disclosure Control*. Wiley Series in Survey Methodology. John Wiley & Sons.

[11] Iacus, S. M. and Porro, G. (2007). Missing data imputation, matching and other applications of random recursive partitioning. *Computational Statistics & Data Analysis*, 52:773–789.

[12] Little, R. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9:407–426.

[13] O'Keefe, C. M. (2012). Confidentialising exploratory data analysis output in remote analysis. *Journal of Official Statistics*, 28:591–613.

[14] O'Keefe, C. M. and Good, N. M. (2009). Regression output from a remote analysis system. *Data & Knowledge Engineering*, 68:1175–1186.

[15] O'Keefe, C. M. and Shlomo, N. (2012). Comparison of remote analysis with statistical disclosure control for protecting the confidentiality of business data. *Transactions on Data Privacy*, 5:403–432.

[16] Raghunathan, T. E., Lepkowsi, J. M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*, 27:85–96.

[17] Raghunathan, T. E., Reiter, J. P., and Rubin, D. R. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19:1–16.

[18] Reiter, J. P. (2003). Model diagnostics for remote-access regression systems. *Statistics and Computing*, 13:371–380.

[19] Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21:441–462.

[20] Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9:130–134.

[21] Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9:462–468.

[22] Shlomo, N. and De Waal, T. (2008). Protection of micro-data subject to edit constraints against statistical disclosure. *Journal of Official Statistics*, 24:1–26.

[23] Sparks, R., Carter, C., Donnelly, J., O'Keefe, C. M., Duncan, J., Keighley, T., and McAullay, D. (2008). Remote access methods for exploratory data analysis and statistical modelling: Privacy-Preserving Analytics™. *Computer Methods and Programs in Biomedicine*, 91:208–222.

[24] Sutcliffe, P., Caruso, M., and Teasdale, H. (2004). 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. pp. 32.