Journal of Privacy and Confidentiality
Vol. 10 (2) 2020

Submitted Sept 2019
Published June 2020

SYNTHETIC BUSINESS MICRODATA : AN AUSTRALIAN EXAMPLE

CHIEN-HUNG CHIEN, A.H. WELSH, AND JOHN D MOORE

Mathematical Science Institute, The Australian National University and Methodology Division, Australian Bureau of Statistics
e-mail address: joseph.chien@abs.gov.au

College of Business and Economics, The Australian National University
e-mail address: Alan.Welsh@anu.edu.au

Statistical Data Integration Division, Australian Bureau of Statistics
e-mail address: john.moore@abs.gov.au

Abstract. Enhancing microdata access is one of the strategic priorities for the Australian Bureau of Statistics (ABS) in its transformation program. However, balancing the trade-off between enhancing data access and protecting confidentiality is a delicate act. The ABS could use synthetic data to make its business microdata more accessible for researchers to inform decision making while maintaining confidentiality. This study explores the synthetic data approach for the release and analysis of business data. Australian businesses in some industries are characterised by oligopoly or duopoly. This means the existing microdata protection techniques such as information reduction or perturbation may not be as effective as for household microdata. The research focuses on addressing the following questions: Can a synthetic data approach enhance microdata access for the longitudinal business data? What is the utility and protection trade-off using the synthetic data approach? The study compares confidentialised input and output approaches for protecting confidentiality and analysing Australian microdata from business survey or administrative data sources.

Key words and phrases: synthetic business microdata, perturbation.

Disclaimer the results of these studies are based, in part, on tax data supplied by the Australian Taxation Office (ATO) to the ABS under the Taxation Administration Act 1953, which requires that such data is only used for the purpose of administering the Census and Statistics Act 1905. Legislative requirements to ensure privacy and secrecy of this data have been adhered to. In accordance with the Census and Statistics Act 1905, results have been confidentialised to ensure that they are not likely to enable identification of a particular person or organisation. Any discussion of data limitations or weaknesses is in the context of using the data for statistical purposes, and is not related to the ability of the data to support the ATO's core operational requirements. Views expressed in this paper are those of the authors and do not necessarily represent those of the ABS. Where quoted or used, they should be attributed clearly to the authors.

JPrivacy Confidentiality
www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.733

© C.H. Chien, A. H. Welsh, and J. D. Moore
Creative Commons (CC BY-NC-ND 4.0)

Introduction

Statistical agencies are constantly facing decisions on how to best balance the trade-off between protecting data confidentiality and providing greater access to the valuable data they collect to inform decision making. Regulation 15 of the *Statistics Determination 2018* ensures safe access to ABS data in the form of unidentified individual statistical records (microdata), for research and analysis purposes. Regulation 15 stipulates that information can be disclosed if done so 'the information is disclosed in a manner that is not likely to enable the identification of the individual' [The Australian Government, 2018]. Protections are important for producing high quality statistics. However, protections have to be balanced with appropriate levels of data access and dissemination. As economist George Stigler pointed out in 1980, data is both a private and public good [Abowd, 2017]. On the one hand, statistical agencies must protect confidentiality, but at the same time they also need to ensure that data is accessible so that it can be used to inform decisions that have significant impact on the public interest [Abowd and Schmutte, 2019].

The ABS has increasingly emphasised providing better access to microdata for research. The ABS uses the Five Safes Framework to ensure microdata can be used appropriately by taking into consideration safe people, projects, settings, data and output [ABS, 2016, Desai et al., 2016]. The ABS provides three types of microdata products - TableBuilder, Confidentialised Unit Record Files or CURFs and detailed microdata [ABS, 2017]. For business microdata access, researchers can download basic CURFs for analysis in their own environment. However, these basic CURFs contain little detail and are reported at a more aggregate level [Tam et al., 2009]. The ABS also produces more detailed CURFs for research using suppression, aggregation, and top and bottom coding methodologies, to enable analysis of microdata [O'Keefe and Shlomo, 2012]. These techniques can make microdata from business surveys or administrative data sources (or business microdata) less useful because some Australian industries are characterised by oligopoly or duopoly. This means analysing business microdata could lead to the identification of units when the data contains large business units, unlike household or person microdata which have a large number of similar respondents. The ABS needs to take stronger protection measures to minimise the likelihood of disclosure [O'Keefe and Shlomo, 2012]. As a result, useful information is suppressed or aggregated to avoid re-identification of large businesses.

The ABS could consider releasing synthetic datasets for researchers to enhance access to business microdata [Chien et al., 2018]. Synthetic datasets preserve some of the relationships between variables so that researchers can make valid inferences about the target population without accessing the underlying microdata [Loong, 2012]. The US Census Bureau uses synthetic data to make its business microdata more accessible to researchers and provides a validation service [Kinney et al., 2011, Miranda and Vilhuber, 2016].

This paper compares a confidentialised input approach i.e. synthetic data and a confidentialised output approach i.e. perturbation (see Appendix A for a description). The second section of this paper describes two synthetic data methods explored in this analysis. The perturbation method is discussed in section three. The forth section provides utility and risk results for these approaches. The final section contains conclusions. Appendix A describes different disclosure methodologies. Appendix B discusses the statistical models to create the experimental dataset and how we impute missing data. All subsequent analysis is based on the completed dataset.

Disclosure control - synthetic data

This paper explores two synthetic data generation methods for Australian business microdata - the sequential regression (SR) of Raghunathan et al. [2001] and non-parametric imputation based on classification and regression trees (CART) proposed by Reiter [2005b]. The literature on synthetic data generation methods is growing, see Kim et al. [2018] and Hu et al. [2018] on non-parametric Bayesian approaches. Statistics New Zealand has also explored some Bayesian approaches to create synthetic data [Graham, 2008].

We create fully synthetic data for three variables - $\ln y$, $\ln K$ and $\ln M$ from an imputed experimental dataset (see Appendix B for details). The firm output $\ln y$ is the logarithm of total sales adjusted for the repurchase of stocks divided by the total number of employees. Firm capital $\ln K$ is the sum of equipment depreciation, business rental expenses and capital investment deductions divided by the total number of employees. Material costs $\ln M$ are the inputs used in the production process divided by the total number of employees. These variables have higher disclosure risks because the business information is more sensitive. We combine the synthetic variables with the original variables $\ln Firm\_Age$ and time indicator variables to estimate (B.1) in Appendix B for the analyses.

The SR method uses appropriate regression models for different variable types. For example, continuous variables are generated using a normal model and binary variables using a logit model. This study only creates synthetic data for continuous variables. We create three synthetic variables with $\mathbf{y}$ denoting each of the three variables $\ln y$, $\ln K$ and $\ln M$. We use $\mathbf{X}$, $\mathbf{X}^{(K)}$ and $\mathbf{X}^{(M)}$ to denote the matrix for creating synthetic data in $\ln y$, $\ln K$ and $\ln M$, respectively. So if the synthetic data variable is $\ln y$ then $\mathbf{X}$ includes all the independent variables in (B.1) in Appendix B. In comparison, if the synthetic data variable is $\ln K$ then $\mathbf{X}^{(K)}$ includes all the independent variables and $\ln y$ but excludes $\ln K$. Similarly, if the synthetic data variable is $\ln M$ then $\mathbf{X}^{(M)}$ includes all the independent variables and $\ln y$ but excludes $\ln M$.

The SR method generates a continuous vector $\mathbf{y}^{seq}$ from the parameters directly estimated from the fitted regression as follows. First draw a new value $\theta = (\sigma^2, \beta)$ from $Pr(\theta \,|\, \mathbf{y})$. Specifically, the variance is drawn from $\sigma^2 \,|\, \mathbf{X} \sim (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})\chi^{-2}_{n-k}$, where n is the total number of observations and k is the dimension of β. The coefficients are drawn from $\beta \,|\, \sigma^2, \mathbf{X} \sim \mathcal{N}(\hat{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$. Second, the synthetic values for $\mathbf{y}^{seq}$ are drawn from the regression model $\mathbf{y}^{seq} \,|\, \beta, \sigma^2, \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 I)$, where I is the identity matrix. The imputations are generated for each variable sequentially [Drechsler, 2011].

The CART algorithm estimates the conditional distribution of a univariate outcome given multivariate predictors by partitioning the predictors into groups with similar outcomes. The partitions are created by recursive binary splits of the predictors in a tree structure with leaves. The values in each leaf represent the conditional distribution of outcomes that satisfy the partitioning criterion. Effectively, CART preserves the underlying relationships between variables by creating models with many interaction effects [Reiter, 2005b, Burgette and Reiter, 2010].

To create $\mathbf{y}^{cart}$, we first fit a tree relating $\mathbf{y}$ to $\mathbf{X}$. We do this separately for all three variables $\ln y$, $\ln K$ and $\ln M$. The algorithm minimises the deviation of $\mathbf{y}$ within each leaf and stops splitting when the deviation is below 0.001. We do this for three variables and label these trees $tree^{(y),(K),(M)}$. We use $\mathbf{y}_{leaf}$ to represent the predicted values of terminal leaves $leaf^{(y),(K),(M)}$ in the trees. In each leaf of the tree, we use the Bayesian bootstrap to draw new values from $\mathbf{y}_{leaf}$ to create synthetic data [Reiter, 2005b]. The Bayesian

bootstrap differs from the standard bootstrap by varying the selection probabilities in the re-sampling process [Rubin, 1981]. The main advantage of using the Bayesian bootstrap is adding uncertainty in each leaf because the number of values in each leaf tends to be small [Reiter, 2005b].

We generate 20 synthetic datasets using each method and each contains three synthetic variables. We use these datasets to fit model (B.1) in Appendix B and choose the synthetic dataset with the highest log-likelihood for the analysis.

Disclosure control - perturbation

The *confidentialised input approach* produces synthetic microdata that allows researchers to analyse the microdata. In comparison, the *confidentialised output approach*, e.g. perturbation, does not allow researchers to access the underlying microdata (see Appendix A for a description). Researchers can only explore data and perform modelling analyses within a secured remote environment. In this environment, on-the-fly routines are applied to confidentialise results for analysis. These routines protect confidentiality while maximising the utility of the microdata.

The perturbation algorithm starts by considering the estimation for model (B.1) as solving $Sc(\alpha; \mathbf{X}; \mathbf{y}) = 0$, where $Sc(\alpha; \mathbf{X}; \mathbf{y}) = \mathbf{X}^\top(\mathbf{y} - \mathbf{X}^\top\alpha)$. The algorithm then adds the noise $\mathbf{e}$ to the score function. We use α^{pert} to denote the coefficients after the score function has been perturbed. The perturbed estimating equation can be expressed as

$$Sc(\alpha^{pert}; \mathbf{X}; \mathbf{y}) = \mathbf{e}. \tag{0.1}$$

The amount of perturbation is based on a record's contribution to the coefficients in the estimating equation. The perturbation is added using $\mathbf{e} = \mathbf{X}^\top(\mathbf{y} - \mathbf{X}^\top\alpha)\mathbf{u}$, where noise $\mathbf{u}$ is generated independently from the symmetric bimodal triangular distribution with modal points at -1 and 1. The choice of the distribution is to minimise bias in the model estimation[1]. In our setting, analysts will only have access to the confidentialised outputs. The estimated coefficients after perturbation are $\hat{\alpha}^{pert} = \hat{\alpha} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{e}$ where $\hat{\alpha}$ is the estimated coefficient using the original microdata. The solution of (0.1) $\hat{\alpha}^{pert}$ is an unbiased estimate of α because the noise is small and its expected value has mean zero $E(\mathbf{e}) = 0$. The perturbation has a similar effect to removing records that have large contribution to the estimated coefficients [Chipperfield and O'Keefe, 2014, Chipperfield, 2014].

Empirical results

This paper considers the utility and protection trade-off in the synthetic data and perturbation approaches in our particular setting. Duncan and Stokes [2004] and Cox et al. [2011] argue that statistical agencies should consider the risk-utility trade off when they evaluate different approaches to enhance data accessibility while protecting confidentiality.

There are many different approaches to estimating disclosure risks for individual records using the probability of matching between microdata. These individual record risk scores are then aggregated for the entire data file, see [Bethlehem et al., 1990, Shlomo, 2010, Drechsler, 2011]. We adapt the approach of Kim et al. [2018] to measure the risk-utility

[1]The choice of the perturbation distribution is based on the ABS research. The ABS does not allow us to disclose the exact perturbation distribution.

trade-off. We use the proportion of correctly linked records criterion to measure disclosure risk. However, instead of computing distance measures, we perform probabilistic linkage between unconfidentialised and confidentialised microdata. We use $\ln y$ and the total number of employees in each firm j as linkage variables. It is not possible to get an exact linkage due to disclosure protection. We calculate the percentage of correctly linked records with the threshold of linkage to 80 percent accuracy. We also compare a scenario where the linkage variables are $\ln y$, $\ln K$, $\ln M$ and the total number of employees in each firm j for comparison. We follow Kim et al. [2018] and use a propensity score approach to measure utility in the Risk-Utility map (see Appendix D for a description). The lower the propensity scores the higher the utility because the synthetic data is closer to the unconfidentialised data. Similar to measuring disclosure risks, we also consider two scenarios — one includes just $\ln y$ and the other scenario includes all three variables.

Figure 5 in Appendix E shows the Risk-Utility map for the different disclosure control methods. This Risk-Utility map illustrates the chosen methods at fixed parameter setting and is, therefore, illustrative. To select among methods, multiple parameter settings for each method must be examined. The Risk-Utility map shows that all these methods provide similar trade-offs but generally SR provides better protection because of the lower proportion of correct linked records. The slightly better utility provided by the perturbation approach may be due to the choice of noise distribution. We are interested in comparing synthetic data with the standard ABS perturbation approach in the analysis. It is also interesting to note that perturbation has the highest utility but the lowest protection in the $\ln y$ scenario but lowest utility and slightly higher protection in the scenario with three variables. This is because the noise added is independent of the data and more variables imply more noise. However, the differences in the level of protection are smaller in the scenario with three variables compared to the scenario with one variable.

This study also compares the estimated coefficients using confidentialised input and outputs approaches with the estimated coefficients using the original microdata. We consider the utility is high if the confidentialised coefficients are similar to those estimated from the original data. Figures 6 and 7 in Appendix F compare the estimated coefficients using different approaches for all industries. These figures show the results for the main variables and intercepts. There are overlaps in the confidence intervals of the coefficients estimated using perturbation, CART and no protection. SR has the least overlap in our setting. The confidence intervals are overlapping between no protection, CART and perturbation approaches. CART performs better than SR at preserving the underlying relationships between variables because it captures many interaction effects.

As expected, labour, capital and materials components contribute positively to outputs. Researchers will draw similar conclusions if analysing confidentialised and unconfidentialised microdata. Figure 8 shows the model residuals using hex-bin plots. The plotting region is broken into a mesh of tessellating hexagons, each of which is coloured indicating how many observations lie in that hexagon. Figure 9 shows the normal quantile-quantile plots for all industries. There are no notable differences when we compare different approaches with the model results using unconfidentialised data, see Appendix G.

Conclusions

This research compares synthetic data and perturbation approaches for disseminating Australian business microdata. The preliminary results show that synthetic data can be a

possible dissemination tool to make more business microdata accessible while ensuring confidentiality.

The analysis shows that the confidentialised input approach provides more protection than the confidentialised output approach in this particular setting - one percent sample file of business microdata. The protection may be needed because in this setting the confidentialised output approach never permits the researcher to access the underlying confidential data, while the confidentialised input approach does allow this access. The amount of utility loss from synthetic data and perturbation approaches is comparable because the estimated coefficients are similar and the risk and utility map also shows similar trade offs. Synthetic data could be a possible approach for the ABS to consider to enhance access to business microdata. This preliminary research has several areas for possible extension including:

- exploring multilevel models for creating synthetic data to better capture the hierarchical structure of the dataset [Drechsler, 2015].
- looking into nonparametric Bayesian methods
- considering other non-parametric approaches for synthetic data such as random forest or differential privacy [Drechsler and Reiter, 2011].
- exploring synthetic data approaches which also maintain differential privacy standards [Abowd and Vilhuber, 2008, Wasserman and Zhou, 2010, Wang, 2019]. There is emerging research interest in using methods that maintain differential privacy to better protect statistical publication [Abowd, 2018]. It would be interesting to consider Bayesian sampling approach that can also provide differential privacy for practical problems [Wang et al., 2015].

Acknowledgement

Authors would like to express our gratitude to the following ABS colleagues - Dr Siu-Ming Tam, Sybille McKeown, Lisette Aaron, Diane Braskic, Dr Philip Gould, Rowan Hatley, Dr Sarah Hinde, Dr Anders Holmberg, Grace Kim, David Taylor, Liza Tiy and Carter Wong for their helpful comments and support for this research. Dr Jörg Drechsler for sharing R code and the ABS Remote Execution Environment for Microdata project team for the developing the R code and Sebastian Lucie's advice and those who provided comments at the Synthetic Datasets for Statistical Disclosure Control - Research and Applications Around the World workshop at the 61st ISI World Statistics Congress 2017. We remain solely responsible for the views expressed in this paper.

References

J. M. Abowd. How will statistical agencies operate when all data are private? *Journal of Privacy and Confidentiality*, 7(3), 2017.

J. M. Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867, 2018.

J. M. Abowd and I. M. Schmutte. An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1):171–202, 2019.

J. M. Abowd and L. Vilhuber. How protective are synthetic data? In *International Conference on Privacy in Statistical Databases*, pages 239–246. Springer, 2008.

J. M. Abowd, R. H. Creecy, and F. Kramarz. Computing person and firm effects using linked longitudinal employer-employee data. Report, US Census Bureau, 2002. ftp://ftp2.census.gov/ces/tp/tp-2002-06.pdf.

ABS. Information paper transforming statistics for the future, 2016. http://www.abs.gov.au/AUSSTATS/abs@.nsf/be4aa82cd8cf7f07ca2570d60018da27/e4d483bab4e1ad93ca257f4c00170bb6!OpenDocument.

ABS. Microdata entry page, 2017. http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Statistical+Data+Integration+-+Business+Longitudinal+Analysis+Data+Environment+(BLADE).

P. Allison. Imputation by predictive mean matching promise and peril, 2015. http://statisticalhorizons.com/predictive-mean-matching.

A. N. Baraldi and C. K. Enders. An introduction to modern missing data analyses. *Journal of school psychology*, 48(1):5–37, 2010. ISSN 0022-4405.

J. G. Bethlehem, W. J. Keller, and J. Pannekoek. Disclosure control of microdata. *Journal of the American Statistical Association*, 85(409):38–45, 1990. ISSN 01621459. doi: 10.2307/2289523. URL http://www.jstor.org/stable/2289523.

R. Breunig and M.-H. Wong. A richer understanding of australia's productivity performance in the 1990s: Improved estimates based upon firm-level panel data. *Economic Record*, 84 (265):157–176, 2008.

L. F. Burgette and J. P. Reiter. Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172(9):1070–1076, 2010.

C.-H. Chien and A. Mayer. Use of a prototype linked employer-employee database to describe characteristics of productive firms. Report, Australian Bureau of Statistics, 2015. http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.055.

C.-H. Chien, A. H. Welsh, and J. D. Moore. Research paper: Synthetic microdata - a possible dissemination tool. Report, Australian Bureau of Statistics, 2018. http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.163.

C.-H. Chien, A. H. Welsh, and R. Breunig. Approaches to analysing micro-drivers of aggregate productivity. Report, Australian Bureau of Statistics, 2019. http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.055.

J. O. Chipperfield. Disclosure-protected inference with linked microdata using a remote analysis server. *Journal of Official Statistics*, 30(1):123–146, 2014. ISSN 2001-7367.

J. O. Chipperfield and C. M. O'Keefe. Disclosure-protected inference using generalised linear models. *International Statistical Review*, 82(3):371–391, 2014. ISSN 1751-5823. doi: 10.1111/insr.12054.

L. H. Cox, A. F. Karr, S. K. Kinney, J. Domingo-Ferrer, G. T. Duncan, C. M. O'Keefe, and N. Shlomo. Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act [with discussions]. *International Statistical Review / Revue Internationale de Statistique*, 79(2):160–199, 2011.

T. Desai, F. Ritchie, and R. Welpton. Five safes: designing data access for research. *Economics Working Paper Series*, 1601, 2016. URL http://eprints.uwe.ac.uk/28124/1/1601.pdf.

J. Drechsler. *Synthetic Datasets for Statistical Disclosure Control Theory and Implementation*. Lecture Notes in Statistics. Springer, New York, 2011. ISBN 978-1-4614-0325-8.

J. Drechsler. Multiple imputation of multilevel missing data—rigor versus simplicity. *Journal of Educational and Behavioral Statistics*, 40(1):69–95, 2015. doi: 10.3102/1076998614563393.

J. Drechsler and J. P. Reiter. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, 55 (12):3232–3243, 2011.

G. T. Duncan and S. L. Stokes. Disclosure risk vs. data utility: The r-u confidentiality map as applied to topcoding. *CHANCE*, 17(3):16–20, 2004. ISSN 0933-2480.

T. E. D. Enamorado, B. Fifield, and K. Imai. Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113(2):353–371, 2019. ISSN 0003-0554. doi: 10.1017/S0003055418000783.

R. E. Fay. When are inferences from multiple imputation valid? In *Proceedings of the Survey Research Methods Section*, pages 227–232. American Statistical Association, 1992.

I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969. doi: 10.1080/01621459.1969.10501049.

P. Graham. *Methods for creating synthetic data*. Statistics New Zealand, 2008. http://archive.stats.govt.nz/~/media/Statistics/about-us/statisphere/Files/official-statistics-research-series/osr-series-v3-2008-methods-creating-synthetic-data.pdf.

O. Harel and X. Zhou. Multiple imputation: review of theory, implementation and software. *Statistics in medicine*, 26(16):3057–3077, 2007. ISSN 0277-6715.

J. Honaker and G. King. What to do about missing values in time-series cross-section data. *American Journal of Political Science*, 54(2):561–581, 2010. ISSN 1540-5907.

J. Hu, J. P. Reiter, Q. Wang, et al. Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Analysis*, 13(1):183–200, 2018.

H. J. Kim, J. P. Reiter, and A. F. Karr. Simultaneous edit-imputation and disclosure limitation for business establishment data. *Journal of Applied Statistics*, 45(1):63–82, 2018.

G. King, J. Honaker, A. Joseph, and K. Scheve. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95: 49–69, March 2001.

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. *International Statistical Review*, 79(3):362–384, 2011.

F. Koller-Meinfelder. *Analysis of incomplete survey data-multiple imputation via bayesian bootstrap predictive mean matching*. Thesis, Faculty of Social and Economic Sciences, 2009.

R. J. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, 83(404):1198–1202, 1988. ISSN 0162-1459.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley and Sons, 2014. ISBN 1118625889.

B. Loong. *Topics and applications in synthetic data*. Thesis, Statistics, 2012. URL https://dash.harvard.edu/handle/1/9527319.

D. C. Mare, D. R. Hyslop, and R. Fabling. Firm productivity growth and skill. *New Zealand Economic Papers*, pages 1–25, 2016. ISSN 0077-9954. doi: 10.1080/00779954.2016.1203815.

X.-L. Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, pages 538–558, 1994. ISSN 0883-4237.

J. Miranda and L. Vilhuber. Using partially synthetic microdata to protect sensitive cells in business statistics. *Statistical Journal of the IAOS*, 32(1):69–80, 2016.

T. Nguyen and D. Hansell. firm dynamics and productivity growth in australian manufacturing and business services oct 2014. Report, ABS, 2014. http://www.abs.gov.au/.

C. M. O'Keefe and N. Shlomo. Comparison of remote analysis with statistical disclosure control for protecting the confidentiality of business data. *Trans. Data Privacy*, 5(2):403–432, 2012.

T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96, 2001. ISSN 0714-0045.

J. P. Reiter. Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):185–205, 2005a. ISSN 1467-985X. doi: 10.1111/j.1467-985X.2004.00343.x. URL http://dx.doi.org/10.1111/j.1467-985X.2004.00343.x.

J. P. Reiter. Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21(3):441, 2005b. ISSN 0282-423X.

D. B. Rubin. The bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981. ISSN 00905364. URL http://www.jstor.org/stable/2240875.

D. B. Rubin. Statistical disclosure limitation. *Journal of official Statistics*, 9(2):461–468, 1993.

N. Schenker and J. M. G. Taylor. Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis*, 22(4):425–446, 1996. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/0167-9473(95)00057-7. URL http://www.sciencedirect.com/science/article/pii/0167947395000577.

M. Schomaker and C. Heumann. Model selection and model averaging after multiple imputation. *Computational Statistics and Data Analysis*, 71:758–770, 2014. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/j.csda.2013.02.017. URL //www.sciencedirect.com/science/article/pii/S016794731300073X.

N. Shlomo. Releasing microdata disclosure risk estimation, data masking and assessing utility. *Journal of Privacy and Confidentiality*, 2(1):7, 2010.

N. Shlomo. Probabilistic record linkage for disclosure risk assessment. In *International Conference on Privacy in Statistical Databases*, Privacy in Statistical Databases, pages 269–282. Springer International Publishing, 2014. ISBN 978-3-319-11257-2.

S.-M. Tam, K. Farley-Larmour, and M. Gare. Supporting research and protecting confidentiality. abs microdata access: Current strategies and future directions. *Statistical Journal*

of the IAOS, 26(3, 4):65–74, 2009. ISSN 1874-7655.

M. T. Tan, G.-L. Tian, and K. W. Ng. *Bayesian missing data problems: EM, data augmentation and noniterative computation*. Chapman and Hall/CRC, 2009. ISBN 1420077503.

The Australian Government. 'Statistics Determination - Reg 15', 2018. https://www.legislation.gov.au/Details/F2018L01114.

G. Vink, L. E. Frank, J. Pannekoek, and S. Van Buuren. Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68(1):61–90, 2014. ISSN 0039-0402.

Y.-X. Wang. Per-instance differential privacy. *Journal of Privacy and Confidentiality*, 9(1), 2019.

Y.-X. Wang, S. Fienberg, and A. Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pages 2493–2502, 2015.

L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

I. R. White, P. Royston, and A. M. Wood. Multiple imputation using chained equations issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011. ISSN 1097-0258. doi: 10.1002/sim.4067. URL http://dx.doi.org/10.1002/sim.4067.

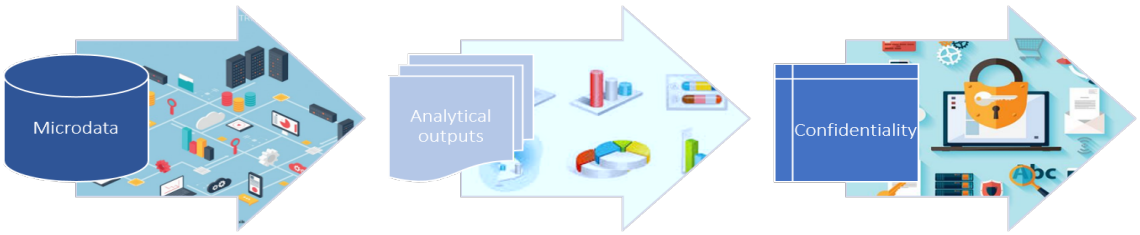
Appendix A. Disclosure control methodologies

O'Keefe and Shlomo [2012] categorise statistical disclosure control methodologies into two main approaches - *confidentialised input* and *confidentialised output*. Examples of confidentialised input methods include aggregation, geographical suppression, rounding, swapping and adding noise (see Figure 1). However, it is often difficult to quantify the amount of information loss or level of protection achieved using confidentialised input approaches. Rubin [1993] proposed a method to generate synthetic data by repeatedly sampling from a statistical model estimated from actual microdata. The synthetic datasets can be used for inference while protecting confidentiality.

Figure 1. Confidentialise input approach

Confidentialised output approaches allow data access in a remote analysis system. The system takes a query and returns the results to the analyst. The analyst does not have direct access to the microdata. The remote system imposes restrictions on the queries and applies routines to deliver confidentialised results (see Figure 2).

Figure 2. Confidentialise output approach



Appendix B. Statistical model and missing data analysis

We are interested in preserving the statistical relationships between the variables in the firm production function. The statistical model is specified as:

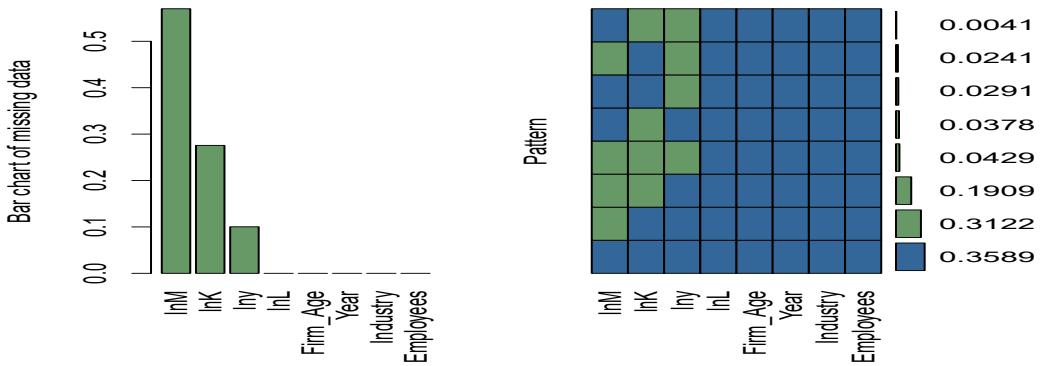
$$\ln y_{jkt} = \alpha_1 \ln L_{jkt} + \alpha_2 \ln K_{jkt} + \alpha_3 \ln M_{jkt} + \alpha_3 \ln Firm\_Age_{jkt} + \tau_{kt} + \epsilon_{jkt}, \qquad \text{(B.1)}$$

where $\ln y_{jkt}$ is the logarithm of total sales adjusted for the repurchase of stocks divided by the total number of employees for firm j in industry k at time t. The logarithm of estimated firm average labour components $\ln L_{jkt}$ for firm j in industry k at time t is derived using

the method proposed by Abowd et al. [2002]. Details can be found in Chien et al. [2019]. The logarithm of capital cost $\ln K_{jkt}$ is the logarithm of the sum of equipment depreciation, business rental expenses and capital investment deductions divided by the total number of employees for firm j in industry k at time t. The logarithm of material costs $\ln M_{jkt}$ is the logarithm of the inputs used in the production process divided by the total number of employees for firm j in industry k and time t. The logarithm of firm age is $\ln Firm\_Age_{jkt}$ for firm j in industry k at time t. We also include time fixed effects τ_{kt} for industry k at time t [Breunig and Wong, 2008, Nguyen and Hansell, 2014, Mare et al., 2016]. This gives 15 unknown regression parameters in (B.1). This study used a one percent stratified sample of business microdata from an expanded prototype dataset ($N > 45000$ firms). Chien and Mayer [2015], Chien et al. [2019] provide more details of the prototype dataset. We simplify notation in (B.1) by **removing the subscripts**. We also use different fonts i.e., $\boldsymbol{\mathcal{X}}$, to represent observed and imputed $N \times 15$ matrices containing all the independent variables in (B.1). Similarly, we use $\boldsymbol{y}$ to represent the observed vector containing dependent variable in (B.1).

The prototype sample contains missing values, particularly for material inputs. Figure 3 shows the missing data pattern; the three variables with missing values include ($\ln M$, $\ln K$ and $\ln y$) in descending order.

FIGURE 3. Missing data pattern



Note. The green tile indicates missing data. The blue tile indicates non missing data. Consider ABS and Patents subfigure at the top left, the left panel is a bar chart showing the propostion of missing data for each variable. The right panel shows the 8 missing data patterns in the data and the proportion of each pattern.

The missing values in the 1% sample are imputed assuming the data are missing at random (MAR). The consequence of this assumption is that missing values can be imputed using models fitted to the observed data [Little and Rubin, 2014]. We adapt a similar notation to Reiter [2005a]. The experimental dataset consists of $[\boldsymbol{y}, \boldsymbol{\mathcal{X}}]$, where $\boldsymbol{y}$ is $N \times 1$ vector which includes the dependent variable, and $\boldsymbol{\mathcal{X}}$ is $N \times 15$ matrix which includes all the independent variables in (B.1). We have imputed the missing variables $\ln y$, $\ln K$ and $\ln M$.

We use two Bayesian imputation approaches - Predictive Mean Matching and Expectation Maximisation and Bootstrap to impute the missing data.

The observed dataset consists of two $N \times 16$ matrices, $\boldsymbol{\mathcal{D}} = [\boldsymbol{y}, \boldsymbol{\mathcal{X}}]$, where $\boldsymbol{\mathcal{X}}$ includes all the independent variables in (B.1), and the response indicator matrix $\boldsymbol{\mathcal{R}}$ which we use to partition $\boldsymbol{\mathcal{D}}$ into the observed $\boldsymbol{\mathcal{D}}_{obs}$ and the missing $\boldsymbol{\mathcal{D}}_{mis}$. We use $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{X}}^{(K)}$ and $\boldsymbol{\mathcal{X}}^{(M)}$ to denote the matrix for imputing missing data in $\ln y$, $\ln K$ and $\ln M$, respectively. So if the missing data variable is $\ln y$ then $\boldsymbol{\mathcal{X}}$ includes all the independent variables in (B.1). In comparison, if the missing data variable is $\ln K$ then $\boldsymbol{\mathcal{X}}^{(K)}$ includes all the independent variables and $\ln y$ but excludes $\ln K$. If the missing data variable is $\ln M$ then $\boldsymbol{\mathcal{X}}^{(M)}$ includes all the independent variables and $\ln y$ but excludes $\ln M$. We impute the missing values in $\ln y$, $\ln K$ and $\ln M$ separately, using two Bayesian imputation approaches - Predictive Mean Matching (PMM) and Expectation Maximisation and Bootstrap (EMB).

PMM selects from a set of possible donors from the complete cases whose predictive means are closest to that of the missing case [Little, 1988]. The value of the selected $\boldsymbol{y}_{obs}$ are then imputed for $\boldsymbol{y}_{mis}$. This method is similar to a hot-deck imputation because it randomly choose one $\boldsymbol{y}_{imp}$ from nearest neighbour complete cases. The box 1 describes the concept of the algorithm [Vink et al., 2014].

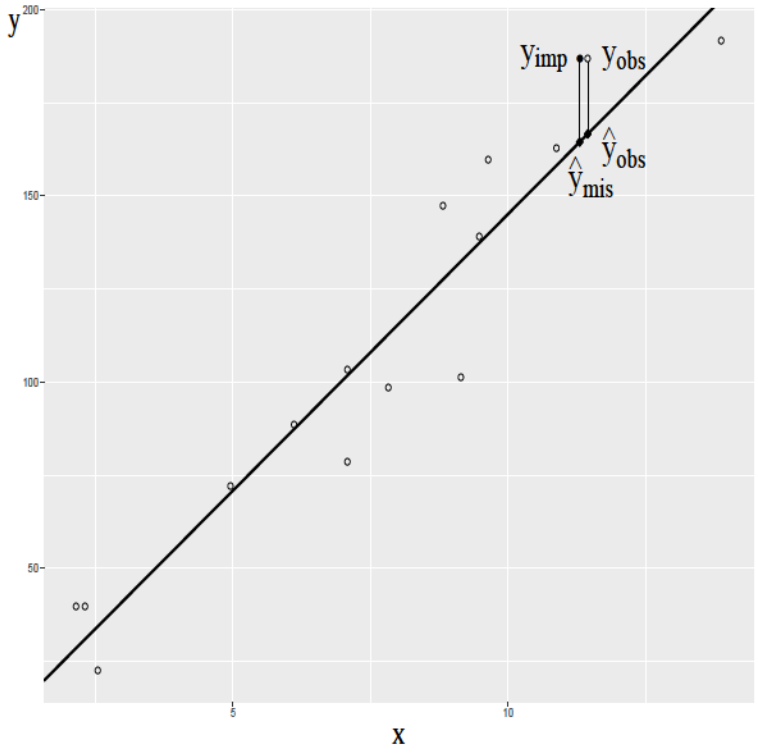
Algorithm 1: PMM algorithm

Data: use $\boldsymbol{\mathcal{D}}_{obs}$ to estimate $\hat{\beta}$ and $\hat{\epsilon}$
draw variance $\tilde{\sigma}^2$ from $\hat{\epsilon}^\top \hat{\epsilon}/A$ where A is χ^2 with $N-k$ with k is the number of parameters.
draw $\tilde{\beta}$ from a multivariate normal distribution centered at $\hat{\beta}$ with covariance matrix $\tilde{\sigma}^2(\boldsymbol{\mathcal{X}}_{obs}^\top \boldsymbol{\mathcal{X}}_{obs})^{-1}$.
calculate $\hat{\boldsymbol{y}}_{obs} = \boldsymbol{\mathcal{X}}_{obs}\hat{\beta}$ and $\hat{\boldsymbol{y}}_{mis} = \boldsymbol{\mathcal{X}}_{mis}\tilde{\beta}$

for *each* $\boldsymbol{y}_{mis}$ **do**
  find distance $\Delta_i = |\hat{\boldsymbol{y}}_{obs,i} - \hat{\boldsymbol{y}}_{mis,k}|$ where $i \neq k$.
  randomly sample one donor from Δ_i with $i = 1, \cdots, 5$ smallest elements and take the corresponding $\hat{\boldsymbol{y}}_{obs}$ to imput $\boldsymbol{y}_{mis}$.
end

Figure 4 shows how PMM imputes the missing values $\boldsymbol{y}_{imp}$ by randomly selecting one out of five plausible donors $\boldsymbol{y}_{obs}$ with smallest distance Δ. The $\boldsymbol{y}_{imp}$ has the smallest Δ in this example. PMM has the advantage of imputing real values observed from the data [Schenker and Taylor, 1996, White et al., 2011, Allison, 2015]. PMM also gives more robust estimates in the presence of misspecification in the imputation model [Koller-Meinfelder, 2009].

FIGURE 4. Imputation of a nearest neighbor with PMM



Note. ◦ indicates observed values y_{obs}, • indicates imputed value y_{imp} and ◆ indicates fitted values $\hat{y}_{obs}$ and $\hat{y}_{mis}$.
Source: adapted from [Koller-Meinfelder, 2009, p.32]

King et al. [2001] propose EMB which combines Expectation Maximisation (EM) algorithm with bootstrap sampling. Unlike PMM, EMB uses predicted values of a linear regression fitted to the observed data to impute missing values. EMB assumes variables in $\boldsymbol{\mathcal{D}}$ are multivariate normal and data are missing at random [King et al., 2001]. The imputation formula is

$$\widetilde{\mathcal{D}}^{(j)}_{mis,i} = \boldsymbol{\mathcal{D}}^{(-j)}_{obs,i}\tilde{\beta} + \tilde{\epsilon}_i, \tag{B.2}$$

where $\tilde{}$ indicates a random draw from the appropriate posterior. The symbol $\widetilde{\mathcal{D}}^{(j)}_{mis,i}$ denotes a imputed value for row i and column j and $\widetilde{\boldsymbol{\mathcal{D}}}^{(j)}_{obs,i}$ denotes the vector of values observed of all columns in row i except column j. The coefficients $\hat{\beta}$ can be calculated from the complete data parameters $\vartheta = (\mu, \Sigma)$, where μ is the mean vector and Σ is the variance-covariance matrix. The randomness of $\widetilde{\mathcal{D}}^{(j)}_{mis,i}$ is created by both estimation uncertainty due to unknown ϑ and uncertainty in $\tilde{\epsilon}_i$ because Σ is not a matrix of zero [Honaker and King,

2010]. The box 2 simplifies the notation by removing the superscripts and subscripts for $\boldsymbol{\mathcal{D}}_{mis}$ and $\boldsymbol{\mathcal{D}}_{obs}$ to describe the concept of the algorithm [Tan et al., 2009].

Algorithm 2: EMB algorithm

Data: generate m bootstrap sample of size n with replacement from the posterior $Pr(\vartheta) \int Pr(\boldsymbol{\mathcal{D}} \mid \vartheta) d\boldsymbol{\mathcal{D}}_{mis}$ described in Equation C.4b.

keep draws of $\tilde{\vartheta}$ with probabilities proportional to the importance ratio - the ratio of the posterior to the asymptotic normal approximation evalued at $\tilde{\vartheta}$. King et al. [2001] defines the importance ratio (IR) without prior as

$$IR = \frac{\ell(\tilde{\vartheta} \mid \boldsymbol{\mathcal{D}}_{obs})}{\mathcal{N}(\tilde{\vartheta} \mid \tilde{\vartheta}, V(\tilde{\vartheta}))}.$$

Result: in each sample m, fill in $\boldsymbol{\mathcal{D}}_{mis}$ by running an EM algorithm described below.

Let $\tilde{\vartheta}^{(i)}$ be the current guess of $\tilde{\vartheta}$,

Expectation step computes the Q function defined by

$$\begin{aligned} Q(\tilde{\vartheta}^{(i)} \mid \tilde{\vartheta}) &= E[\ell(\tilde{\vartheta}; \boldsymbol{\mathcal{D}}_{obs}, \boldsymbol{\mathcal{D}}_{mis}) \mid \boldsymbol{\mathcal{D}}_{obs} \tilde{\vartheta}^{(i)}] \\ &= \int \ell(\tilde{\vartheta}; \boldsymbol{\mathcal{D}}_{mis}, \boldsymbol{\mathcal{D}}_{obs}) \times f(\boldsymbol{\mathcal{D}}_{mis} \mid \boldsymbol{\mathcal{D}}_{obs}, \tilde{\vartheta}^{(i)}) d\boldsymbol{\mathcal{D}}_{mis}, \end{aligned}$$

Maximisation step maximises Q with respect to $\tilde{\vartheta}$ to obtain

$$\tilde{\vartheta}^{(i+1)} = \text{argmax}\, Q(\tilde{\vartheta}^{(i)} \mid \tilde{\vartheta}).$$

repeat

both Expectation and Maximisation steps

until *convergence occurs*;

Baraldi and Enders [2010] discussed how multiple imputation methods create many copies of datasets with different imputed values. These datasets are analysed using the same estimation step to generate multiple sets of parameters and normal standard errors. The final result is derived by using model averaging to incorporate the uncertainty associated with the model selection process into standard errors and confidence intervals [Schomaker and Heumann, 2014]. It is unclear if model averaging from multiple imputed datasets provides the best results. This study applies each method 20 times to the 1% sample and we select the best imputed dataset which maximises the likelihood for (B.1) from the 40 datasets [Fay, 1992, Meng, 1994].

Appendix C. A Bayesian framework for imputation

We assume data are missing at random. The consequence of this assumption is that missing data can be imputed from fitting model on the observed data. The complete data parameters are $\vartheta = (\mu, \Sigma)$, where μ is the mean vector and Σ is the variance-covariance matrix. The likelihood of these parameters given the observed data can be expressed as

$$Pr(\boldsymbol{\mathcal{D}}_{obs}, \boldsymbol{\mathcal{R}} \mid \vartheta) = \int Pr(\boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{R}} \mid \vartheta) d\boldsymbol{\mathcal{D}}_{mis} \tag{C.1a}$$

$$= \int Pr(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\mathcal{R}}, \vartheta) Pr(\boldsymbol{\mathcal{R}} \mid \vartheta) d\boldsymbol{\mathcal{D}}_{mis}. \tag{C.1b}$$

Using Bayes' theorem we can rewrite the first term $Pr(\boldsymbol{\mathcal{D}} \mid \boldsymbol{\mathcal{R}}, \vartheta)$ in (C.1b) as $Pr(\boldsymbol{\mathcal{D}} \mid \vartheta) Pr(\boldsymbol{\mathcal{R}} \mid \boldsymbol{\mathcal{D}}, \vartheta) / Pr(\boldsymbol{\mathcal{R}} \mid \vartheta)$. Substituting the new term into (C.1b) we have

$$Pr(\boldsymbol{\mathcal{D}}_{obs}, \boldsymbol{\mathcal{R}} \mid \vartheta) = \int Pr(\boldsymbol{\mathcal{D}} \mid \vartheta) Pr(\boldsymbol{\mathcal{R}} \mid \boldsymbol{\mathcal{D}}, \vartheta) d\boldsymbol{\mathcal{D}}_{mis}. \tag{C.2}$$

Assuming the data are missing at random, the patterns of missing data depend only on the observed data, so (C.2) is simplified to

$$\begin{aligned} Pr(\boldsymbol{\mathcal{D}}_{obs}, \boldsymbol{\mathcal{R}} \mid \vartheta) &= \int Pr(\boldsymbol{\mathcal{D}} \mid \vartheta) Pr(\boldsymbol{\mathcal{R}} \mid \boldsymbol{\mathcal{D}}_{obs}, \vartheta) d\boldsymbol{\mathcal{D}}_{mis} \\ &= \int Pr(\boldsymbol{\mathcal{D}} \mid \vartheta) d\boldsymbol{\mathcal{D}}_{mis} Pr(\boldsymbol{\mathcal{R}} \mid \boldsymbol{\mathcal{D}}_{obs}) \\ &= Pr(\boldsymbol{\mathcal{D}}_{obs} \mid \vartheta) Pr(\boldsymbol{\mathcal{R}} \mid \boldsymbol{\mathcal{D}}_{obs}). \end{aligned} \tag{C.3}$$

Maximising (C.1a) over ϑ is the same as maximising the first term in (C.3) over ϑ. The likelihood can therefore be expressed as $L(\vartheta \mid \boldsymbol{\mathcal{D}}_{obs}) \propto Pr(\boldsymbol{\mathcal{D}}_{obs} \mid \vartheta)$. Harel and Zhou [2007] describe the posterior distribution to draw imputations as

$$Pr(\boldsymbol{\mathcal{D}}_{mis} \mid \boldsymbol{\mathcal{D}}_{obs}) = \int Pr(\boldsymbol{\mathcal{D}}_{mis} \mid \boldsymbol{\mathcal{D}}_{obs}, \vartheta) Pr(\vartheta \mid \boldsymbol{\mathcal{D}}_{obs}) d\vartheta, \text{ where} \tag{C.4a}$$

$$Pr(\vartheta \mid \boldsymbol{\mathcal{D}}_{obs}) \propto Pr(\vartheta) \int Pr(\boldsymbol{\mathcal{D}} \mid \vartheta) d\boldsymbol{\mathcal{D}}_{mis} \tag{C.4b}$$

is the observed posterior distribution for ϑ and $Pr(\vartheta)$ is an uninformative Jeffreys's prior for Σ.

Appendix D. Risk-Utility map

Shlomo [2014] argued statistical agencies could consider using probabilistic linkage for risk assessment because it extends the notion of population uniqueness. We use the R `fastLink` package - Fast Probabilistic Record Linkage with Missing Data to perform probabilistic linkage between confidentialised and unconfidentialised records. We use the notation from Enamorado et al. [2019] and consider two data sets - unconfidentialised data $\boldsymbol{\mathcal{D}}^{ori}$ and confidentialised data $\boldsymbol{\mathcal{D}}^{con}$ with P common variables. An agreement vector of length P is denoted by $\gamma(i,j)$. The pth element of $\gamma_{p(i,j)}$ represents the within-pair similarity for the pth variable between the ith observation of the unconfidentialised data set and the jth observation of the confidentialised data set. There is a total of L_p levels for the pth variable to measure similarity, so the element of the agreement can be defined as:

$$\gamma_{p(i,j)} = \begin{cases} 0 & \text{unmatched} \\ 1 & \\ \vdots & \text{similar} \\ L_p - 2 & \\ L_p - 1 & \text{matched} \end{cases}$$

Fellegi and Sunter [1969] propose the most commonly used probabilistic linkage model. We use M_{ij} to indicate if ith record in $\boldsymbol{\mathcal{D}}^{ori}$ matches with jth record in $\boldsymbol{\mathcal{D}}^{con}$. We follow Enamorado et al. [2019] to specify the basic model as:

$$\gamma_{p(i,j)} \mid M_{ij} = m \overset{indep.}{\sim} \text{Multinomial}(L_p, \pi_{pm}), \tag{D.1}$$

$$M_{ij} \overset{iid}{\sim} \text{Bernoulli}(\lambda) \tag{D.2}$$

where π_{pm} is a vector of length L_p giving the probability of each agreement level for the pth variable given that the pair is matched ($m = 1$) or unmatched ($m = 0$), and λ representing the probability of a match across all pairwise comparison. The prototype dataset contains more than 45000 firms so the potential search space for pairwise comparison is large. We use the available blocking variables such as industry and year to reduce the computational problem [Fellegi and Sunter, 1969].

We use the propensity score approach of Kim et al. [2018] to measure utility. We first concatenate unconfidentialised data $\boldsymbol{\mathcal{D}}^{ori}$ and confidentialised data $\boldsymbol{\mathcal{D}}^{con}$ and add an indicator variable whose values equal one for all observations from $\boldsymbol{\mathcal{D}}^{con}$ and equal zero for all observations from $\boldsymbol{\mathcal{D}}^{ori}$. Let p_i be the probability the indicator variable equals one. Then we fit a logistic regression

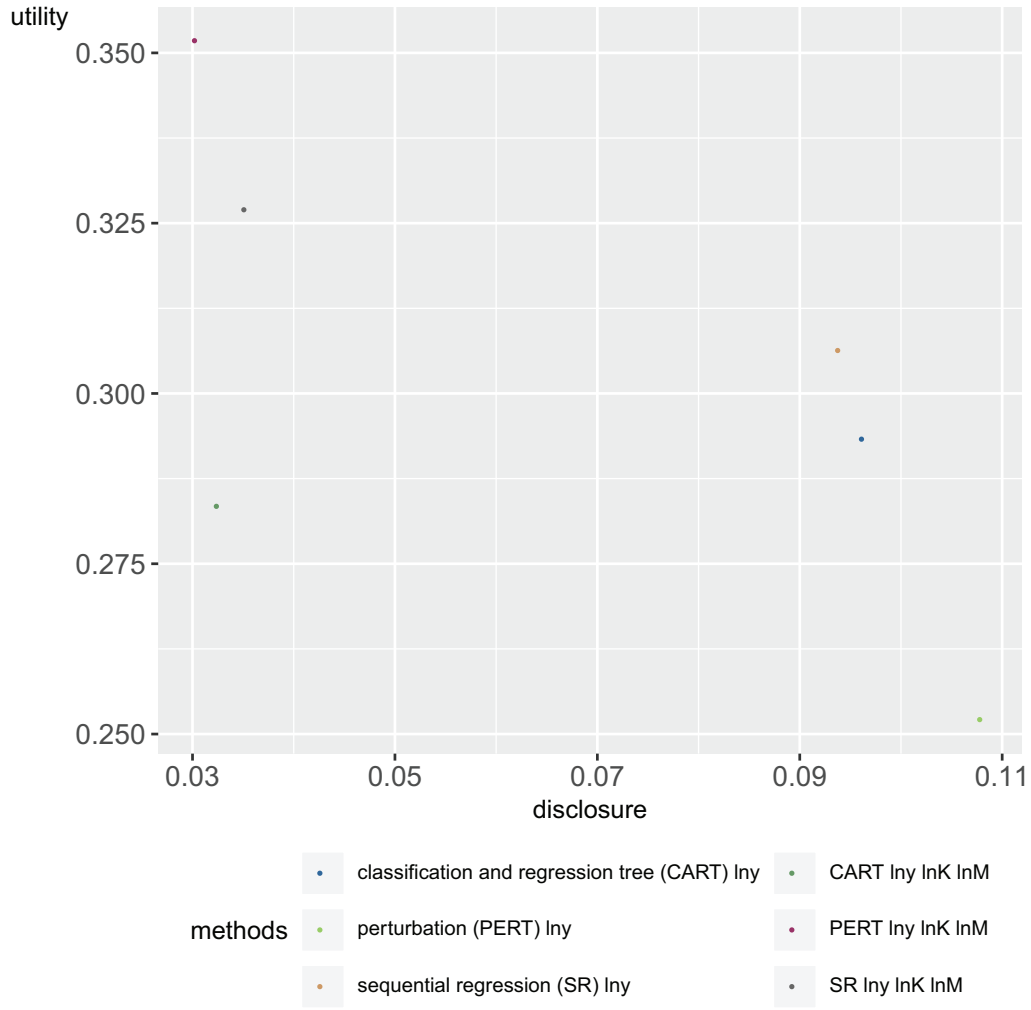
$$\log\Big[\frac{p_i}{1-p_i}\Big] = \mathbf{X}_i^\top \theta, \tag{D.3}$$

where $\mathbf{X}_i$ includes $\ln y$, $\ln K$ and $\ln M$ with main effects and all interactions. For $i = 1, \cdots, 2N$, we compute the predictive probabilities $\hat{p}_i$ and then compute Kim et al.'s [2018] measure

$$U_{prop} = \frac{1}{2N} \sum_{i=1}^{2N} (\hat{p}_i - \frac{1}{2})^2.$$

Appendix E. Risk and utility map

Figure 5. Risk and utility map



Appendix F. utility measure results - all indutries

Figure 6. coefficients plots - main variables

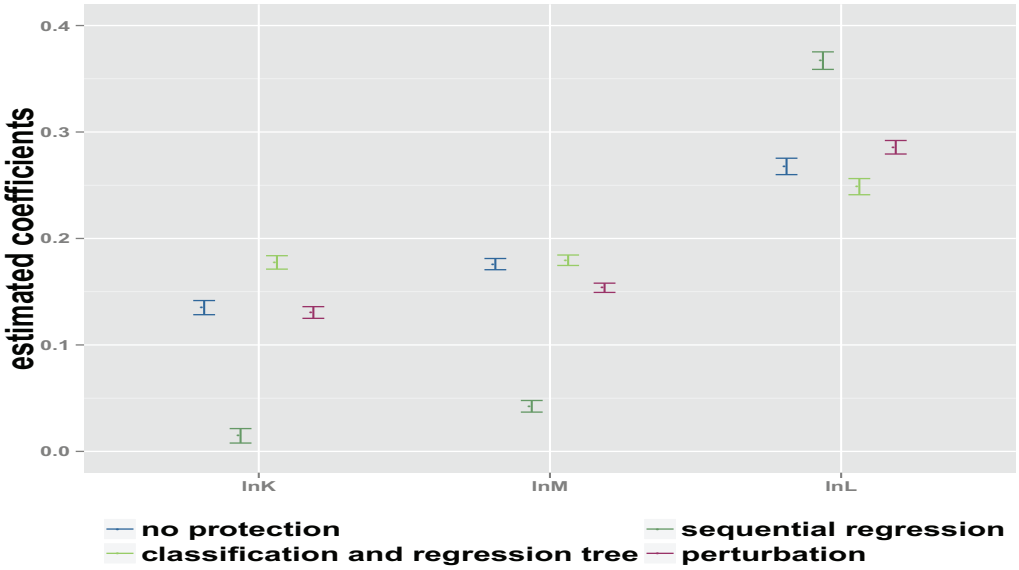
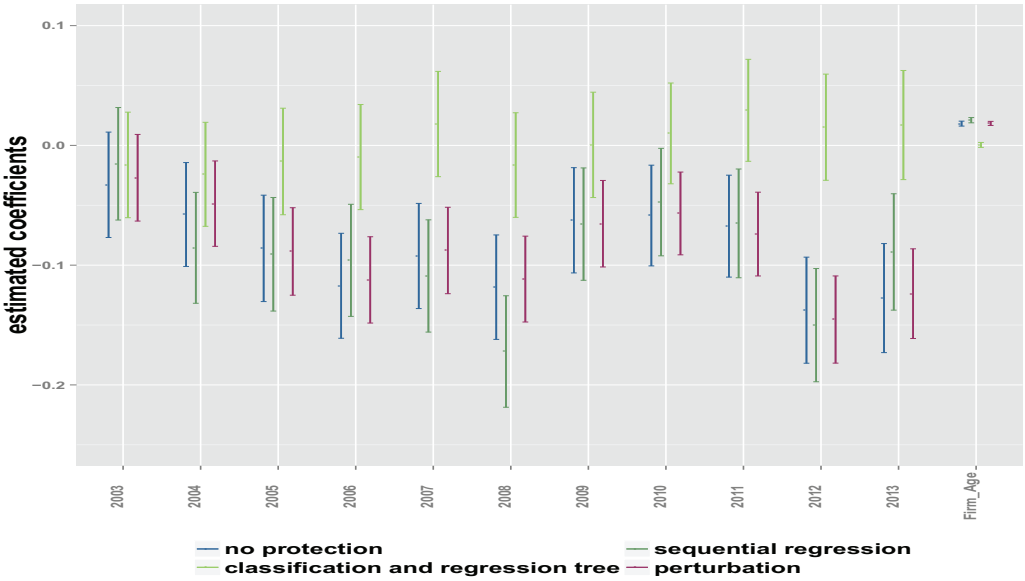


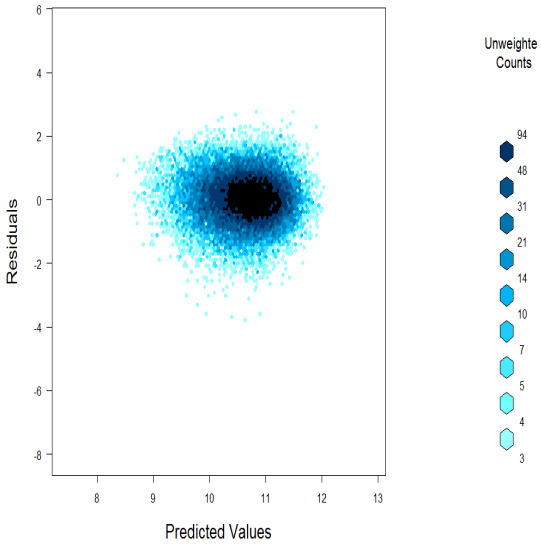
Figure 7. coefficients plots - all other variables



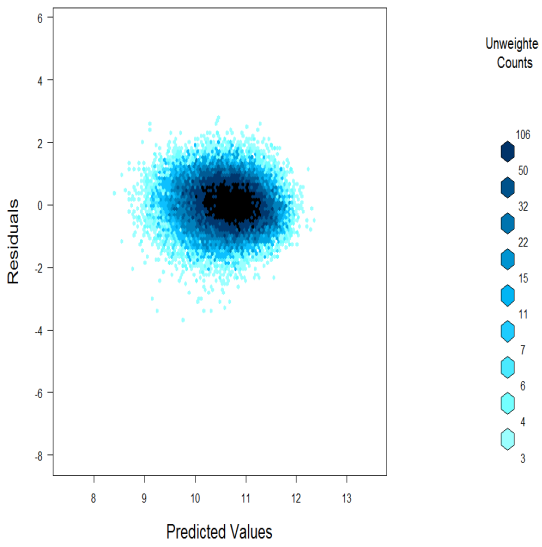
Appendix G. selected diagnostics

Figure 8. Confidentialised residual plots - ALL industries

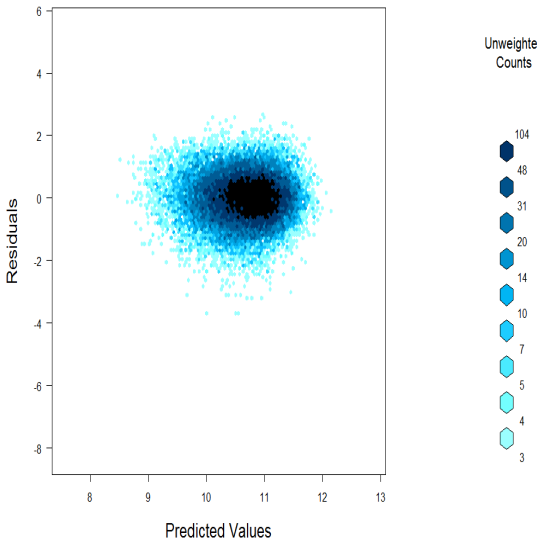
(a) original



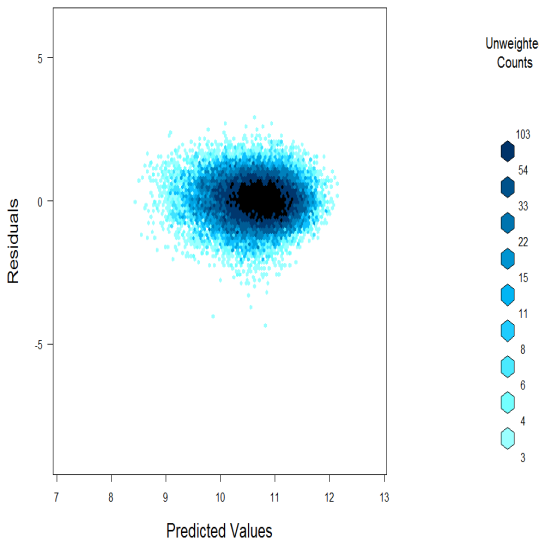
(c) synthetic data - SR



(b) perturbation



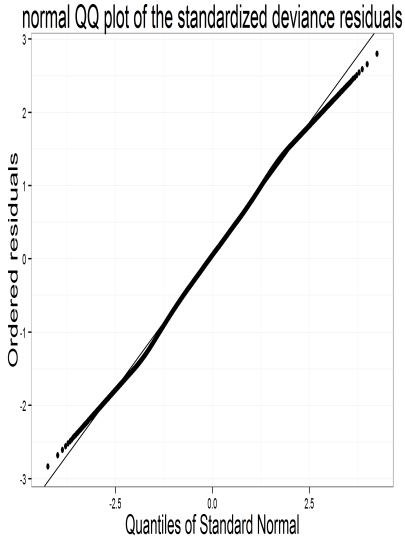
(d) synthetic data - CART



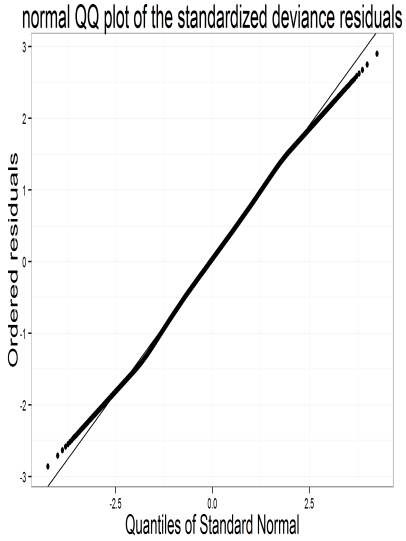
Note. Residuals come from fitting (B.1) to different approaches. The plotting region on these figures is broken into a mesh of tessellating hexagons, each of which is coloured indicating how many observations lie in that hexagon.

FIGURE 9. QQ Norm plots - ALL industries

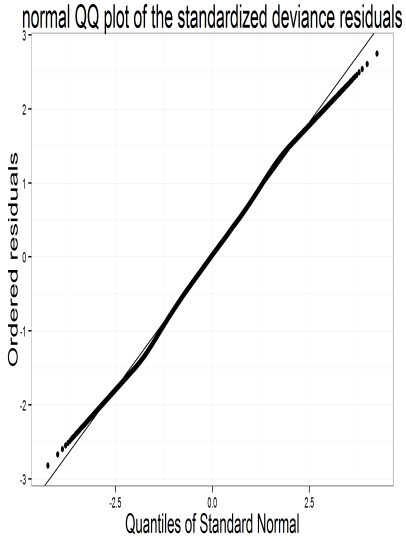
(A) original



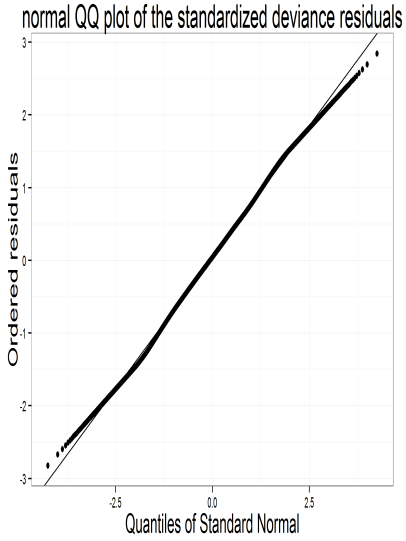
(C) synthetic data - SR



(B) perturbation



(D) synthetic data - CART



Note. Residuals come from fitting (B.1) to different approaches. A 45 degree line indicates that residuals are normally distributed.

This work is licensed under the Creative Commons License Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0). To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-nd/4.0/ or send a letter to Creative Commons, 171 Second St, Suite 300, San Francisco, CA 94105, USA, or Eisenacher Strasse 2, 10777 Berlin, Germany