STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Statistics in Society
Series A

J. R. Statist. Soc. A (2018)
181, *Part* 4, *pp.* 1211–1230

Correlates of record linkage and estimating risks of non-linkage biases in business data sets

Jamie C. Moore, Peter W. F. Smith and Gabriele B. Durrant

University of Southampton, UK

[Received November 2016. Revised October 2017]

Summary. Researchers often utilize data sets that link information from multiple sources, but non-linkage biases caused by linked and non-linked subject differences are little understood, especially in business data sets. We address these knowledge gaps by studying biases in linkable 2010 UK Small Business Survey data sets. We identify correlates of business linkage propensity, and also for the first time its components: consent to linkage and register identifier appendability. As well, we take a novel approach to evaluating non-linkage bias risks, by computing data set representativeness indicators (comparable, decomposable sample subset similarity measures). We find that the main impacts on linkage propensities and bias risks are due to consenter–non-consenter differences explicable given business survey response processes, and differences between subjects with and without identifiers caused by register undercoverage of very small businesses. We then discuss consequences for the analysis of linked business data sets, and implications of the evaluation methods we introduce for linked data set producers and users.

Keywords: Bias adjustment; Business surveys; Consent to linkage; Improving data set quality; Non-linkage bias; Record linkage; Representativeness indicators

1. Introduction

The use of data sets that include linked information from multiple sources such as surveys and administrative records is increasing in many fields (see Dunne *et al.* (2009), Jutte *et al.* (2011) and Schnell (2013) for primers in economics and the health and social sciences respectively). Linked data sets can enable study of otherwise intractable questions, for instance providing longitudinal business data so job creation and destruction can be quantified (Hijzen *et al.*, 2010; Bjelland *et al.*, 2011). They can also improve estimation, as for example when administrative data supplement survey earnings estimates (Hayes *et al.*, 2007; Meijer *et al.*, 2012). Moreover, they may reduce collection costs by increasing data usage (Sala *et al.*, 2012). However, a lack of information about possible biases in data set estimates limits applications. Not all sample subjects may be linked. If linked and non-linked subjects differ, non-linkage biases, i.e. deviations of estimates from values if linkage is complete, can occur. Such biases can be substantial (Abowd and Vilhuber, 2005; Buckley *et al.*, 2007; Kho *et al.*, 2009; Sakshaug and Kreuter, 2012; McKay, 2012), but little work exists on linked–non-linked subject differences, especially in business data sets, or on bias evaluation methods. In this paper, we address these knowledge gaps in research on non-linkage biases in 2010 UK Small Business Survey (SBS) data sets.

An understanding of linked–non-linked subject differences is essential for inference from linked data sets. It informs on (reasons for) non-linkage biases in involved attribute covari-

Address for correspondence: Jamie C. Moore, Administrative Data Research Centre for England and Department of Social Statistics and Demography, University of Southampton, Southampton, SO17 1BJ, UK.
E-mail: j.c.moore@soton.ac.uk

ates and (focal) substantive correlates. It also underlies modifications to production techniques to improve data set quality (see also below). However, research so far seldom considers that differences can arise during several components of the linkage process (see Bohensky *et al.* (2010) for exceptions). Subject consent to linkage is often legally required, and consenters and non-consenters may differ. In addition, not all subjects may be linkable, and linkable and non-linkable subjects may differ. For full understanding, both components should be studied along with their product: overall linkage. Otherwise, for instance, it may be found that impacts of consent strategy modifications to improve data sets (e.g. Sakshaug *et al.* (2013) and Sala *et al.* (2014)) are reduced by extra consenters also being less linkable. Existing work mostly considers single components in (individual subject) social and health surveys. Consenter–non-consenter differences are mainly quantified by identifying survey attribute covariate consent correlates. These include sociodemographic and health attributes, which especially imply biases given associations with substantive covariates (Dunn *et al.*, 2004; Kho *et al.*, 2009; Sala *et al.*, 2012, 2014). The same methods are used to quantify linkable–non-linkable subject differences, with similar attributes identified as linkability correlates (Bohensky *et al.*, 2010).

The lack of research on the above differences in business data sets is an issue due to the growing role of such data sets in national economic analyses (e.g. Bean (2016)). We are not aware of any studies on business consent to linkage. Work on the related problem of survey non-response suggests that different dynamics operate in surveys of businesses and individuals (Cox *et al.*, 1995; Willimack *et al.*, 2004; Janik and Kohaut, 2012; Snijkers *et al.*, 2013). The importance of informed contacts and business data sharing policies for co-operation is stressed, implying associations between item response and consent. Other likely consent correlates associated with substantive covariates include business size (employee number and turnover) and performance measures such as changes in these values (response relationships and posited causes vary). Regarding linkability, linkage is often by register identifier (Abowd *et al.*, 2004; Ritchie, 2008). Registers can exhibit undercoverage of small businesses (Evans and Welpton, 2009), causing business size–linkability correlations (see Section 2.1 for detailed predictions).

A difficulty with non-linkage bias evaluation is that only linked subject values are available for linked secondary data set covariates: they are partially observed. Sakshaug and Kreuter (2012) address this by linking only a social survey consent indicator (which is not considered a response) and computing administrative data summary estimates. McKay (2012) use fully observed social survey attribute covariates and selection models to estimate biases from similar data (in both studies all consenters are linked; see also Buckley *et al.* (2007) and Kho *et al.* (2009)). Work in economics is indirect, comparing for instance longitudinal employee estimates given different production techniques (Abowd and Vilhuber, 2005). However, such methods are not suitable when not all subjects are linked (Sakshaug and Kreuter's (2012) work), or when multiple covariates are of interest and overall data set quality is important. Previously, when comparing production techniques, linkage rates have been used as overall quality measures. With the related problem of survey non-response bias though, non-respondent–respondent differences at high response rates mean that correlations with biases are weak (e.g. Groves (2006)). Similar is likely with non-linkage, so more refined quality measures are needed.

The 2010 SBS focuses on businesses with fewer than 250 employees: a key segment of the UK economy (Department for Business, Innovation and Skills, 2013). The utility of the data set, which is available from the UK Data Service (`http://www.ukdataservice.ac.uk`; the R code for our analyses will be made available through the same service), is enhanced by linkage to other surveys and administrative data. To facilitate this, if businesses consent an attempt is made to append a register identifier, enabling linkage. We begin by describing the SBS and its linkage. Then, to investigate linked–non-linked subject differences, we identify survey attribute covariate correlates of consent, consenter with identifier (e.g. linkability, which is hereafter referred to as identifier) and (overall) linkage propensities. Our analyses provide the first information of this type from a business data set.

Next, we evaluate biases in linked data sets. Given partially observed secondary data set covariates, for the first time we utilize representativeness indicators developed to study survey non-response bias (see Schouten *et al.* (2012) and de Heij *et al.* (2015)). These do not quantify specific biases but instead are comparable sample subset similarity measures computed from inclusion propensities estimated given an auxiliary covariate set. Such covariates, which must be fully observed but critically are obtained from the sample data set, should also be correlated with covariates of interest. If so, low propensity variation (representativeness) implies low bias risk (see Schouten *et al.* (2016) for support in the survey non-response context). In addition, decompositions that quantify propensity variation associated with covariates exist. These enable under-represented subgroups to be identified and their impacts compared to ensure efficient targeting of production technique modifications to improve data set quality (see Section 2.3 for details). We evaluate non-consent, non-identifier and non-linkage bias risks, using the attribute covariates tested as propensity correlates as auxiliary covariates. We quantify and decompose data set representativeness, to assess most important impacts and to inform potential production technique modifications. We then discuss findings with regard to linked SBS and other data sets, and implications of the evaluation methods introduced for data set producers and users.

2. Methods

2.1. The 2010 Small Business Survey

The SBS is commissioned by the UK Department for Business, Innovation and Skills. Roughly biannual since 2003, it surveys and focuses on topics of interest to businesses with fewer than 250 employees, collecting attribute information as well (see also Department for Business, Innovation and Skills (2011, 2013)). The 2010 survey took place from July to September. The survey sample is stratified by country and employee number, but otherwise drawn randomly from the Dun and Bradstreet commercial business database, and includes reporting and local units. Telephone interviewing is used. In the final section, subjects are asked the following question regarding consent to record linkage:

> ‘Would it be possible for BIS to link your responses to other information that you have provided previously to the Government? By this data linkage, we can reduce the burden of our surveys on your business and can improve the evidence that we use.
>
> ‘Data will only be used to inform research on businesses in aggregate—we will never release information that identifies any individual business—and your survey responses remain strictly confidential. Do you give your consent for us to do this?’

If consent is given, deterministic matching of standardized business name, postcode and type (sole owned or limited company) data held by the contractors to the Office for National Statistics ‘Inter-departmental business register’ (IDBR) is attempted. If successful, the business IDBR identifier is appended to its record, enabling linkage to other surveys and administrative data.

Non-linked and linked subjects may differ with respect to a range of survey attribute covariates with likely substantive correlates. In addition to potential size (employee number and turnover)—consent propensity relationships (see Section 1), very small businesses are under-represented in the IDBR (Evans and Welpton, 2009). This is because it only includes those (at sampling) with a turnover of over £70 000 *per annum* and registered to pay value-added tax or those paying salaries over £124.50 per week via the pay as you earn (PAYE) income tax mechanism. Given this, entities with turnovers below the value-added tax threshold should have lower identifier propensities than those above. Regarding PAYE impacts, sole owners and partners may be self-employed and pay themselves a dividend instead of a salary, enabling them to avoid PAYE tax if they employ no one else, and so zero employee entities should have lower identifier propensities than those with employees. We also predict a general legal status impact. Solely owned entities and partnerships may avoid PAYE tax even with employees if only owners earn above the threshold, and so should have lower identifier propensities than incorporated companies and other statuses. Similar differences have already been identified in the 2006 and 2007 SBS data sets (Meldgaard *et al.*, 2015).

Meldgaard *et al.* (2015) also reported business sector identifier propensity differences, without positing causes. In addition, identifiers are issued to businesses (reporting units) with multiple local units. If local units are sampled, linking to the reporting unit will increase failures, so identifier propensities should be lower for multisite than for single-site businesses (we do not consider linkage errors here; see Winkler (1995) for an introduction to this topic). Given these arguments, in our analyses we include survey employee number, turnover, legal status, sector and site (number) item responses as subject attribute covariates. We also include future expected employees and future expected turnover, to test for performance effects on (consent) propensities (see Section 1), and region and exporter (see Table 1 for covariate categories). Subjects confirm information with region and sector but provide it with the other covariates. Future expected employees, exporter, turnover and future expected turnover have 'don't know' categories, and the last two also 'unwilling' categories. These enable tests for survey co-operation effects on (consent) propensities (see Section 1), but we exclude exporter 'don't know' and expected future turnover 'unwilling' categories from the data set because of counts below disclosure thresholds. The other covariates lack such categories, though they are in the questionnaire. Responses of this type may be indicated by (a small number of) missing values. Excluding subjects with missing responses as well reduces the data set from 4580 to 4438 subjects. 87.4% consent to linkage. 72.1 % of these have identifiers appended. Overall, 63% are linkable.

2.2. Correlates of consent, identifier and linkage propensities

To quantify linked–non-linked subject differences in SBS data sets, we use logistic regression to identify survey attribute covariate consent, identifier and linkage propensity correlates. We fit univariable models to inform on (sources of) attribute covariate biases: for instance, average turnover will be overestimated if linkage propensity increases with turnover. We test covariate significance with χ^2-tests. We also fit multivariable models estimating conditional differences, to inform (further) on likely (reasons for) biases in attribute covariate correlates by accounting for associations and identifying main impacts on propensities. We include all covariates as main effects, but in the final model we retain only those for which the Akaike information criterion AIC increases by more than 2 on removal (see Burnham and Anderson (2002)). With both model types, we estimate category differences as contrasts to reference categories, testing significance with Z-tests. The identifier propensity data set is smaller as it includes consenters only (3895 subjects). Department for Business, Innovation and Skills (2013) also report univariable linkage propensities for some covariates.

2.3. Representativeness indicators

Representativeness indicators measure sample subset similarity in terms of inclusion propensity variation given an auxiliary covariate set and are comparable given use of the same auxiliary covariates. Two forms exist: the R-indicator and the coefficient of variation (CV) of response propensities (Schouten *et al.*, 2009, 2011, 2012; de Heij *et al.*, 2015). We utilize the CV, which

Table 1. Category level parameter estimates and covariate level χ^2s from univariable logistic regression models of respondent consent, identifier and linkage propensities†

Covariate	*Category*	*Consent*		*Identifier*		*Linkage*	
		β	*se*	β	*se*	β	*se*
Region	South	1.784	0.074	0.999	0.064	0.514	0.054
	Midlands	0.397§	0.127	0.044	0.098	0.169‡	0.085
	North	0.202	0.129	−0.094	0.103	0.001	0.089
	South, West and Northern Ireland	0.262‡	0.121	−0.137	0.095	−0.013	0.083
	χ^2 (R^2)	11.060‡ (0.003)		3.950 (0.001)		5.390 (0.001)	
Site	Single	1.937	0.052	0.956	0.041	0.538	0.036
	Multisite	0.148	0.111	0.000	0.084	0.048	0.074
	χ^2 (R^2)	1.810 (0.003)		0.010 (0.003)		0.430 (0.003)	
Employees	0	1.708	0.103	0.006	0.080	−0.304	0.075
	1–9	0.212	0.129	1.063§§	0.103	0.920§§	0.093
	10–49	0.347§	0.131	1.294§§	0.105	1.135§§	0.094
	>49	0.500§	0.160	1.134§§	0.121	1.070§§	0.109
	χ^2 (R^2)	11.640§ (0.003)		170.040§§ (0.037)		166.88§§ (0.029)	
Legal status	Sole	1.707	0.097	0.108	0.076	−0.217	0.071
	Partnership	0.264	0.166	0.612§§	0.126	0.582§§	0.114
	Company or other	0.347§	0.113	1.164§§	0.089	1.028§§	0.081
	Don't know	−0.065	0.456	0.490	0.383	0.380	0.337
	χ^2 (R^2)	9.620‡ (0.003)		173.790§§ (0.037)		170.200§§ (0.029)	
Sector	Other services	1.910	0.093	0.667	0.070	0.305	0.063
	Production	0.249	0.165	0.637§§	0.128	0.568§§	0.111
	Construction	0.030	0.180	0.300‡	0.141	0.243‡	0.123
	Transport etc.	−0.016	0.123	0.293§	0.096	0.221§	0.084
	Business services	0.154	0.138	0.424§§	0.106	0.378§§	0.093
	Primary	−0.010	0.284	0.230	0.224	0.176	0.196
	χ^2 (R^2)	4.280 (0.001)		30.750§§ (0.007)		32.210§ (0.005)	
Turnover	< £100000	1.954	0.105	−0.038	0.074	−0.284	0.070
	£100000–0.99 million	0.374§	0.141	1.401§§	0.102	1.255§§	0.092
	£1 million–4.99 million	0.285	0.149	1.423§§	0.110	1.242§§	0.099
	⩾ £5 million	0.457‡	0.194	1.285§§	0.134	1.193§§	0.121
	Don't know	−0.921§§	0.168	0.772§§	0.161	0.277‡	0.135
	Unwilling	−0.832§§	0.157	0.859§§	0.145	0.379§	0.122
	χ^2 (R^2)	119.740§§ (0.036)		251.830§§ (0.055)		289.010§§ (0.049)	
Exporter	No export	2.054	0.099	1.228	0.079	0.780	0.067
	Export	−0.108	0.111	−0.348§§	0.089	−0.296§§	0.076
	χ^2 (R^2)	0.960 (0.001)		15.810§§ (0.003)		15.470§§ (0.003)	
Future employees	Increase	2.364	0.109	1.112	0.074	0.790	0.066
	No change	−0.489§§	0.123	−0.237§	0.087	−0.335§§	0.077
	Decrease	−0.439§	0.169	0.023	0.130	−0.124	0.112
	Don't know	−1.160§§	0.314	−0.707‡	0.298	−0.944§§	0.257
	χ^2 (R^2)	24.190§§ (0.007)		14.130§ (0.003)		29.060§ (0.005)	
Future turnover	Increase	2.189	0.077	1.019	0.055	0.667	0.049
	No change	−0.190	0.143	−0.223‡	0.106	−0.232‡	0.094
	Decrease	−0.347§§	0.104	−0.050	0.080	−0.153‡	0.070
	Unwilling	−0.870§§	0.190	−0.229	0.181	−0.495§§	0.150
	χ^2 (R^2)	50.420§§ (0.007)		5.340 (0.001)		15.310§§ (0.003)	

†Category level parameter estimates are presented as logit means for the first-named categories and differences from the first category for other categories, with the significance of differences tested by Z-tests. Model R^2s are also given.

‡$P < 0.05$.

§$P < 0.01$.

§§$P < 0.001$.

is the standard deviation of inclusion propensities divided by the average propensity. Its use is advised when (as here) compared data set inclusion rates differ, since R-indicators can overestimate representativeness at low rates when possible propensity variation is limited (see also Moore *et al.* (2018)). In addition, the overall CV, which measures overall data set representativeness, quantifies the maximum absolute standardized bias of survey estimate means when non-response correlates maximally to the auxiliary covariates (see Schouten *et al.* (2009, 2011) and de Heij *et al.* (2015)). Similar will hold regarding biases in linked data sets. The overall CV for sample size n and auxiliary covariate set x is

$$\widehat{\mathrm{CV}}(p_x) = \frac{\sqrt{\left[\{1/(n-1)\} \sum_{i=1}^{n} (\hat{p}_i - \hat{\bar{p}})^2\right]}}{\hat{\bar{p}}}, \tag{1}$$

where $\hat{p}_i$ is the inclusion propensity of subject i, $\hat{\bar{p}}$ the average propensity and $\sqrt{[\{1/(n-1)\} \times \Sigma_{i=1}^{n} (\hat{p}_i - \hat{\bar{p}})^2]}$ the propensity standard deviation. The less propensities differ the smaller the CV and the greater data set representativeness. The CV approximate standard errors also exist, based on linearization of a variance estimator for the propensity standard deviation derived by decomposing its distribution into that due to sampling design and that due to model parameter estimates (Shlomo *et al.*, 2012; de Heij *et al.*, 2015; Moore *et al.*, 2018). Converted into 95% confidence intervals (CIs) (CV ± 1.96 se), these enable inference regarding representativeness or not and comparative representativeness.

Partial CVs decompose overall CVs to assess inclusion propensity variation that is associated with attribute covariates (categories). Two forms exist. Unconditional CVs, CV_u, measure the deviation from representativeness (a random sample) with respect to a covariate. CV_u for covariate Z with K categories is

$$\widehat{\mathrm{CV}_\mathrm{u}}(Z, p_x) = \frac{\sqrt{\left[\{1/(n-1)\} \sum_{k=1}^{K} n_k (\hat{\bar{p}}_k - \hat{\bar{p}})^2\right]}}{\hat{\bar{p}}}, \tag{2}$$

where n_k is category k size, and $\hat{\bar{p}}_k$ average propensity in k. The larger the indicator is, the greater the variation associated with Z. Category CVs partition covariate CVs. CV_u for k is

$$\widehat{\mathrm{CV}_\mathrm{u}}\,(Z_k, p_x) = \frac{\sqrt{(\hat{n}_k/n)}(\hat{\bar{p}}_k - \hat{\bar{p}})}{\hat{\bar{p}}}. \tag{3}$$

Indicators may be positive or negative, implying respectively over- or under-representation. The further they are from 0, the greater the contribution of k to non-representativeness.

Conditional CVs, CV_c, measure the deviation from conditional representativeness (a random sample given stratifying covariates) with respect to a covariate, by comparing propensities with those estimated given set x excluding it. CV_c for covariate Z is

$$\widehat{\mathrm{CV}_\mathrm{c}}(Z, p_x) = \frac{\sqrt{\left[\{1/(n-1)\} \sum_{l=1}^{L} \sum_{i \in U_l} (p_i - \hat{\bar{p}}_l)^2\right]}}{\hat{\bar{p}}}, \tag{4}$$

where $\hat{\bar{p}}_l$ is the average inclusion propensity of the lth of L cells resulting from cross-classification of x excluding Z and propensity modelling given this covariate subset. The larger the indicator is, the greater the variation that is solely associated with Z. CV_c for category k is

$$\widehat{\mathrm{CV}}_{\mathrm{c}}(Z_k, p_x) = \frac{\sqrt{\left[\{1/(n-1)\}\sum_{l=1}^{L}\sum_{i\in U_l} h_i(p_i - \hat{\hat{p}}_l)^2\right]}}{\hat{\hat{p}}}, \tag{5}$$

where h_i details whether subject i is in k. Indicators have similar ranges and interpretations to those of the covariate counterparts. Partial CV approximate standard errors also exist. Converted into 95% CIs, these enable statistical inference concerning (conditional) representativeness with respect to covariate or category. CV_{u}s and CV_{c}s hence have parallels with univariable and multivariable propensity models respectively, but category results are in the naturally interpretable form of (under- or over-) representativeness or not rather than contrasts to a reference Schouten *et al.* (2012) also showed that indicators can be interpreted in terms of missing data mechanisms (see Little and Rubin (2002)). Overall CVs quantify the deviation of subjects from missing completely at random given the auxiliary covariates, CV_{u}s the deviation from missing completely at random with respect to a covariate given the auxiliary covariates and CV_{c}s the similar deviation from missing at random given the other auxiliary covariates, i.e. the extent to which subjects are not missing at random. Given this, when modifying production techniques to improve data sets CV_{u}s can be used to identify under-represented subgroups to target. CV_{c}s can ensure efficient targeting: a significant CV_{u} only implies an impact that is also associated with other covariates. With survey non-response, targeting categories with significant CV_{c}s and (if correlations between them exist) a subset of those with significant CV_{u}s only is advised (see also Schouten and Shlomo (2017)).

We use CVs to evaluate SBS data set non-consent, identifier and linkage bias risks. We include the nine attribute covariates as main effects in logistic regression models estimating inclusion propensities. We compute overall CVs, partial CVs for covariates (categories) and also indicator 95% CIs, using a correction to the R code of de Heij *et al.* (2015) that resolves an issue with sometimes inflated standard error estimates (see Moore *et al.* (2018)).

3. Results

3.1. Correlates of consent, identifier and linkage propensities

3.1.1. Consent propensity

Six covariates exhibit univariate correlations with consent propensity (Table 1). For these covariates, we detail category variation in terms of differences from reference categories. With region, propensities are (significantly) lower for the South than for the Midlands and Scotland, Wales and Northern Ireland categories. With employee number, propensities are lower for zero than for more than nine employee categories. With legal status, propensities are lower for sole owned than company categories. With turnover, propensities are lower for the don't know and unwilling categories than the greater than £100 000 category, which is lower than for the from £100 000 to £0.99 millon and greater than £5 million categories. With future expected employees, propensities are higher for the increase category than others. With future expected turnover, propensities are higher for the increase than decrease and unwilling categories.

Four covariates are retained in the multivariable model (Table 2). The region, future expected employees and future expected turnover category differences are as described above, as are turnover differences, except that the categories less than £100 000 and greater than £5 millon are similar. Total explanatory power is low (model $R^2 = 4.4\%$). Concerning correlation causes and links to univariable relationships, associations with being unable or unwilling to report the last three items are expected as co-operation depends on knowledgeable contacts and data sharing policies (see Section 1). Businesses expecting future employees to

Table 2. Multivariable model selection to identify business attribute correlates of consent, identifier and linkage propensities, and model parameter estimates†

Covariate	*Category*	*Consent*		*Identifier*		*Linkage*	
		β	*se*	*β*	*se*	*β*	*se*
	Intercept	2.215	0.169	−0.887	0.128	−0.972	0.116
Region (South)	Midlands	0.394§	0.130	*0.010*	*0.104*	*0.144*	*0.090*
	North	0.213	0.132	*−0.051*	*0.109*	*0.045*	*0.094*
	South, West and Northern Ireland	0.326§	0.124	*−0.104*	*0.101*	*0.043*	*0.087*
	AIC	3179.00		*4273.90*		*5481.30*	
Site (single)	Multisite	*0.069*	*0.120*	−0.357§§	0.098	−0.265§	0.085
	AIC	*3174.80*		4280.40		5485.50	
Employees (0)	1–9	*0.139*	*0.146*	0.574§§	0.118	0.492§	0.106
	10–49	*0.237*	*0.166*	0.608§§	0.144	0.562§§	0.126
	> 49	*0.377*	*0.208*	0.564§	0.183	0.588§§	0.160
	AIC	*3175.70*		4288.50		5496.50	
Legal status (sole owner)	Partnership	*0.141*	*0.176*	0.187	0.138	0.146	0.124
	Company	*0.111*	*0.130*	0.681§§	0.111	0.536§§	0.098
	Don't know	*0.233*	*0.472*	0.292	0.399	0.301	0.351
	AIC	*3178.10*		4308.10		5507.90	
Sector (other services)	Production	*−0.024*	*0.173*	0.482§§	0.139	0.374§	0.120
	Construction	*−0.178*	*0.186*	0.334‡	0.155	0.181	0.133
	Transport etc.	*−0.087*	*0.127*	0.403§§	0.107	0.305§	0.094
	Business services	*−0.034*	*0.143*	0.454§§	0.114	0.372§§	0.100
	Primary	*−0.045*	*0.292*	0.905§§	0.244	0.741§§	0.212
	AIC	*3182.00*		4287.60		5492.10	
Turnover (< £100 000)	£100 000–0.99 million	0.321‡	0.143	0.966§§	0.120	0.872§§	0.108
	£1 million–4.99 million	0.214	0.152	0.875§§	0.152	0.731§§	0.133
	⩾ £5 million	0.337	0.197	0.740§§	0.195	0.653§§	0.170
	Don't know	−0.921§§	0.171	0.489§	0.176	−0.007	0.148
	Unwilling	−0.855§§	0.158	0.629§§	0.158	0.123	0.133
	AIC	3266.50		4326.90		5568.90	
Exporter (no export)	Export	*0.114*	*0.120*	*0.014*	*0.101*	*0.050*	*0.086*
	AIC	*3174.20*		*4271.40*		*5479.60*	
Future employees (increase)	No change	−0.308‡	0.130	*0.049*	*0.093*	*−0.075*	*0.082*
	Decrease	−0.359‡	0.176	*0.036*	*0.136*	*−0.115*	*0.117*
	Don't know	−0.950§	0.327	*−0.456*	*0.318*	*−0.719*	*0.272*
	AIC	3178.10		*4272.70*		*5476.60*	
Future turnover (increase)	No change	−0.150	0.150	*−0.096*	*0.112*	*−0.132*	*0.099*
	Decrease	−0.275‡	0.109	*0.108*	*0.086*	*−0.029*	*0.074*
	Unwilling	−0.412‡	0.201	*0.176*	*0.195*	*−0.132*	*0.161*
	AIC	3175.30		*4271.20*		*5481.70*	
	Final model AIC (R^2)	3173.10 (0.047)		4269.40 (0.08)		5477.90 (0.066)	
	Null model AIC	3299.90		4602.20		5830.44	

†AICs are those after removal from the final model for retained covariates (in normal text), and those after addition to it for others (in italics). Parameter estimates are differences from model intercepts, in logits. Covariate reference categories are in parentheses. Estimates for covariates not retained in the final models are for when they are added. Category differences from reference categories are tested by Z-tests. Model R^2s are also given.

‡$P < 0.05$.

§$P < 0.01$.

§§$P < 0.001$.

increase consenting more is consistent with a predicted performance impact, perhaps reflecting pride in achievements (see Janik and Kohaut (2012)), as is the similar future expected turnover result: the significance of both covariates, implying separate impacts, may be due to different ways of assessing achievement. The region result is similar to response rate variation in social surveys (Moore *et al.*, 2018). Lower less than £100 000 category propensities than for (most) higher turnover entities, and the univariable employee number and legal status associations not arising (implying correlations), suggest a business size impact that may occur because very small businesses are surveyed less often (e.g. Willimack *et al.* (2004)) and lack policies on consent requests.

3.1.2. Identifier propensity

Six covariates exhibit univariable correlations with identifier propensity (Table 1). With employee number, zero employee propensities are (significantly) lower than for other categories. With legal status, propensities are lower for sole owned than partnership and company categories. With turnover, propensities are lower for less than £100 000 than for the don't know, unwilling and (to a greater extent) greater than £100 000 categories. With sector, other services propensities are lower than for other categories. With exporter, non-exporter propensities are higher. With future expected employees, propensities are higher for the increase than no change and don't know categories.

Five covariates are retained in the multivariable model (Table 2). Employee number, turnover and sector category differences are as above, as are legal status differences, except that the sole owned and partnerships categories are similar. Propensities are also lower for multisite businesses. The total explanatory power is low (model $R^2 = 8.0\%$). Concerning causes and univariable relationship links, employee number, legal status and reported turnover differences are consistent with very small business IDBR undercoverage: the significance of multiple attributes implies imperfectly associated impacts. Turnover don't know and unwilling category impacts are unexpected and require further investigation. The site impact is predicted as another unit may have been issued the identifier, increasing linkage failures, but no univariable relationship exists, implying a conditional impact only. That future expected employees and exporter are not retained despite univariable relationships suggests other correlations with these impacts: the first result may be due to businesses replying 'don't know' answering turnover similarly, and those replying no change also being zero employee entities.

3.1.3. Linkage propensity

Seven covariates exhibit univariable correlations with linkage propensity (Table 1). Sector and exporter impacts reflect identifier propensity differences, and employee number and legal status impacts same signed differences in both components. With turnover, propensities are lower for the less than £100 000 than greater than £100 000 categories for similar reasons, and are lower for the less than £100 000 than don't know and unwilling categories, but less so than for identifier propensities because of opposing consent propensity differences. With future expected employees, the propensities are higher for the increase than no change and don't know categories because of same signed differences in both components. As increase and decrease do not also differ in this way though consent propensities do suggests an impact reduced by (non-significant) identifier propensity variation. With future expected turnover, increase category propensities are highest, because of consent (*versus* decrease and unwilling) and identifier (*versus* no change) propensity differences (we did not report the latter result before because of covariate non-significance). No region association exists despite consent propensity differences (though the South *versus* Midlands contrast is significant), also suggesting impacts reduced by (non-significant) identifier propensity variation (but see Section 3.2).

Five covariates are retained in the multivariable model (Table 2). Employee number impacts are as above. Legal status impacts are similar, though sole owned and partnerships are similar, as are sector impacts, except that the other services and construction categories are similar. The turnover less than £100000 and greater than £100000 categories also differ as above, but the less than £100000, don't know and unwilling categories are similar. Site differences reflect (conditional) identifier propensities. The total explanatory power is low (model $R^2 = 6.6\%$). Correlations with other covariates in involved components (see the previous sections) partly explain the non-significance of covariates with univariable relationships. Region, future expected employees and future expected turnover are not retained despite multivariable model consent propensity differences though, suggesting impacts also reduced by (non-significant) identifier propensity variation.

3.2. Small Business Survey data set representativeness

3.2.1. Risks of non-consent bias

The overall CV (0.076; 95% CI 0.060–0.092, where zero implies representativeness and no bias risk) suggests that consenters are non-representative of respondents. Partial CVs (Table 3; Figs 1 and 2) extend propensity model results and are mostly similarly explained. We focus on significant covariates and their categories. Six covariate unconditional CVs (CV_us) are significant. With region, the South is under-represented and the Midlands over-represented. With employee number, zero employees is under-represented and the greater than 49 employee category over-represented. With legal status, solely owned entities are under-represented and companies over-represented. With turnover, the categories don't know and unwilling are under-represented and the greater than £100000 categories over-represented. With future expected employees, the categories no change and don't know are under-represented and the increase category over-represented. With future expected turnover, the categories decrease and unwilling are under-represented and the increase category over-represented.

Covariate conditional CVs (CV_cs) are smaller than absolute CV_u-values, and significant for region and turnover. For these, category CV_c significance is as for CV_us, except that the category less than £100000 turnover is significant. Given causes and correlations between unconditional impacts (see also Section 3.1), from this we identify four under-represented subgroups. By CV_u-size, they are those unable or unwilling to report turnover, zero employee, solely owned entities (very small businesses, significant CV_us only imply a correlated impact) and those in the South. Future expected employees and future expected turnover CV_c non-significance (unlike in the propensity model, an unexplained discrepancy as CVs also identify new effects: see below) probably reflects correlations between the categories don't know and unwilling and turnover equivalents. Other category effects are as with CV_us and propensity models (that these, which suggest performance affects, are not also found in the non-linkage analysis below is probably due to (conditional) category non-identifier variation). The less than £100000 turnover effect arises if the other inequalities are corrected.

3.2.2. Risks of non-identifier bias

The overall CV (0.197; 95% CI 0.177–0.217) suggests that consenters with identifiers are non-representative of consenters, and more so than consenters are of respondents (CV 95% CIs do not overlap): a result that is also implied by mostly larger partial CVs (Table 3; Figs 1 and 2). Seven covariate CV_us are significant. With employee number, the zero employees category

Table 3 Consenter, consenter with identifier and linkable data set overall and partial unconditional and conditional by covariate CVs and 95% CIs†

	Consent		*Identifier*		*Linkable*	
	CV	*95% CI*	*CV*	*95% CI*	*CV*	*95% CI*
Overall	0.076	0.060–0.092	0.197	0.177–0.217	0.228	0.206–0.250
Unconditional						
Region	0.019	0.007–0.031	*0.020*	*−0.004–0.044*	0.026	0.001–0.051
Site	*0.007*	*−0.005–0.019*	*0.001*	*−23.684–23.686*	*0.007*	*−0.026–0.040*
Employees	0.019	0.007–0.031	0.135	0.113–0.157	0.149	0.125–0.173
Legal status	0.018	0.006–0.030	0.135	0.113–0.157	0.151	0.127–0.175
Sector	*0.011*	*−0.003–0.025*	0.056	0.036–0.076	0.064	0.040–0.088
Turnover	0.066	0.052–0.080	0.163	0.141–0.185	0.195	0.173–0.217
Exporter	*0.005*	*−0.009–0.019*	0.039	0.021–0.057	0.044	0.022–0.066
Future employees	0.027	0.015–0.039	0.037	0.017–0.057	0.061	0.037–0.085
Future turnover	0.029	0.017–0.041	0.023	0.001–0.045	0.045	0.021–0.069
Conditional						
Region	0.014	0.002–0.026	*0.008*	*−0.016–0.032*	*0.014*	*−0.011–0.039*
Site	*0.001*	*−0.011–0.013*	*0.019*	*−23.666–23.704*	*0.020*	*−0.013–0.053*
Employees	*0.005*	*−0.007–0.017*	0.032	0.010–0.054	0.035	0.011–0.059
Legal status	*0.002*	*−0.010–0.014*	0.045	0.023–0.067	0.045	0.021–0.069
Sector	*0.004*	*−0.010–0.018*	0.038	0.018–0.058	0.039	0.017–0.061
Turnover	0.045	0.031–0.059	0.059	0.037–0.081	0.082	0.060–0.104
Exporter	*0.003*	*−0.011–0.017*	*0.001*	*−0.019–0.021*	*0.005*	*−0.017–0.027*
Future employees	*0.011*	*−0.001–0.023*	*0.008*	*−0.014–0.030*	*0.015*	*−0.009–0.039*
Future turnover	*0.011*	*−0.001–0.023*	*0.014*	*−0.008–0.036*	*0.009*	*−0.015–0.033*

†Non-significant covariates, based on 95% CIs spanning 0, are in italics.

is under-represented, and categories with employees over-represented. With legal status, the categories sole owned, partnership and don't know are under-represented, and companies over-represented. With sector, the category other services is under-represented, and production and business services over-represented. With turnover, the categories don't know, unwilling (for these categories, non-consent $\mathrm{CV_u}$s are larger) and less than £100 000 are under-represented, and the greater than £100 000 categories over-represented. With exporter, exporters are under-represented, and non-exporters over-represented. With future expected employees, the categories no change and don't know are under-represented, and the increase and decrease categories over-represented. With future expected turnover (which is not significant in the propensity analysis), the categories no change and unwilling are under-represented, and the increase category over-represented. The large site 95% CI reflects minimal between-category variation (see Moore *et al.* (2018)).

Covariate $\mathrm{CV_c}$s are smaller than absolute $\mathrm{CV_u}$-values, except with site, but unlike in the propensity model the statistic is not significant, possibly because (conservative) $\mathrm{CV_u}$ standard errors are used (de Heij *et al.*, 2015). $\mathrm{CV_c}$s are significant for employee number, legal status, sector and turnover. Category $\mathrm{CV_c}$ significance is similar to that for $\mathrm{CV_u}$s, though the categories greater than 49 employees, legal status don't know and greater than £4.99 millon turnover are no longer significant and transport and primary production sectors newly so. Given causes and correlations between (unconditional) impacts (see also Section 3.1), from this we identify seven under-represented subgroups. By category $\mathrm{CV_u}$-size, these are zero employee, sole owned or in partnership, or with less than £100 000 turnover (very small businesses), the other services sector

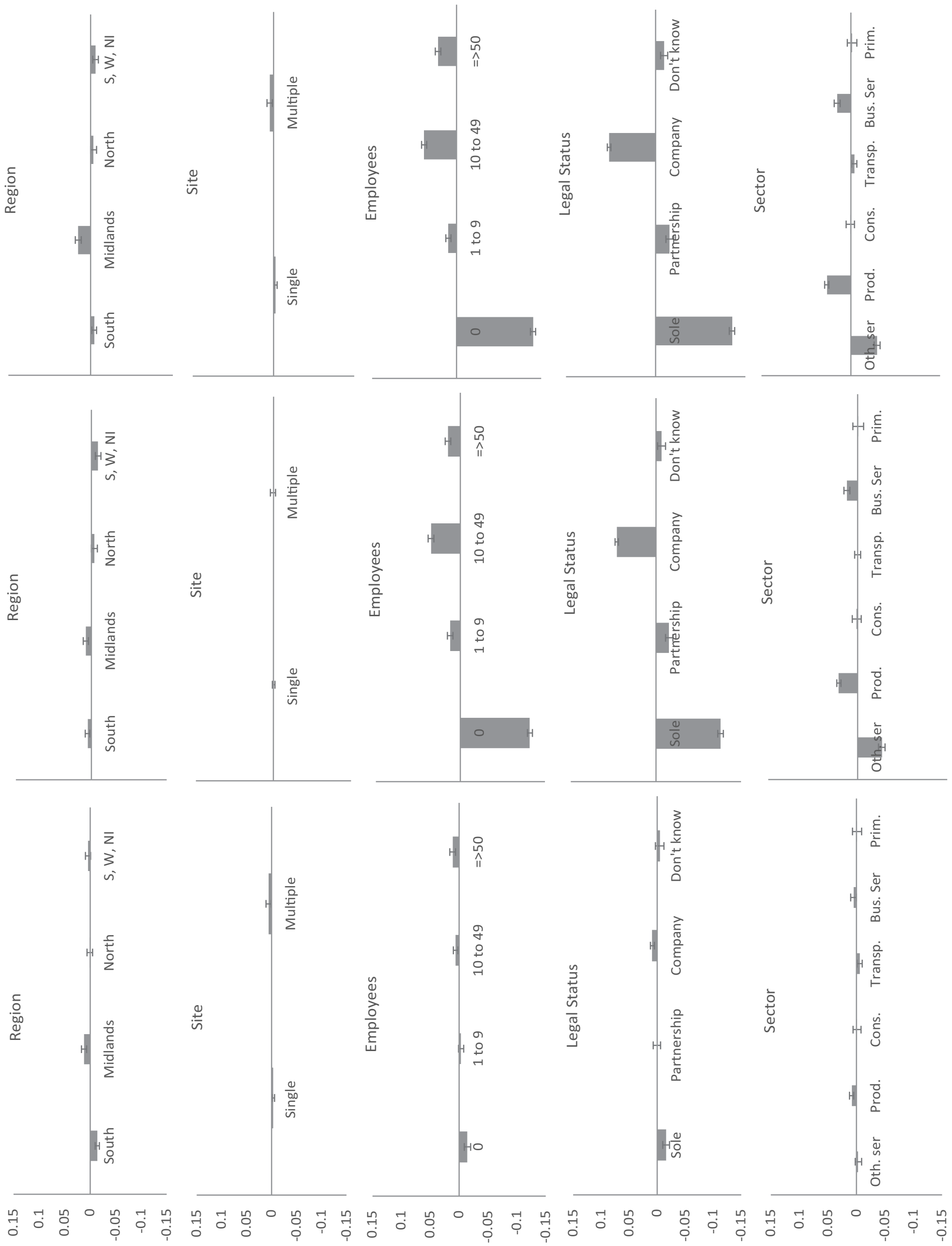

Fig. 1 *(continued)*

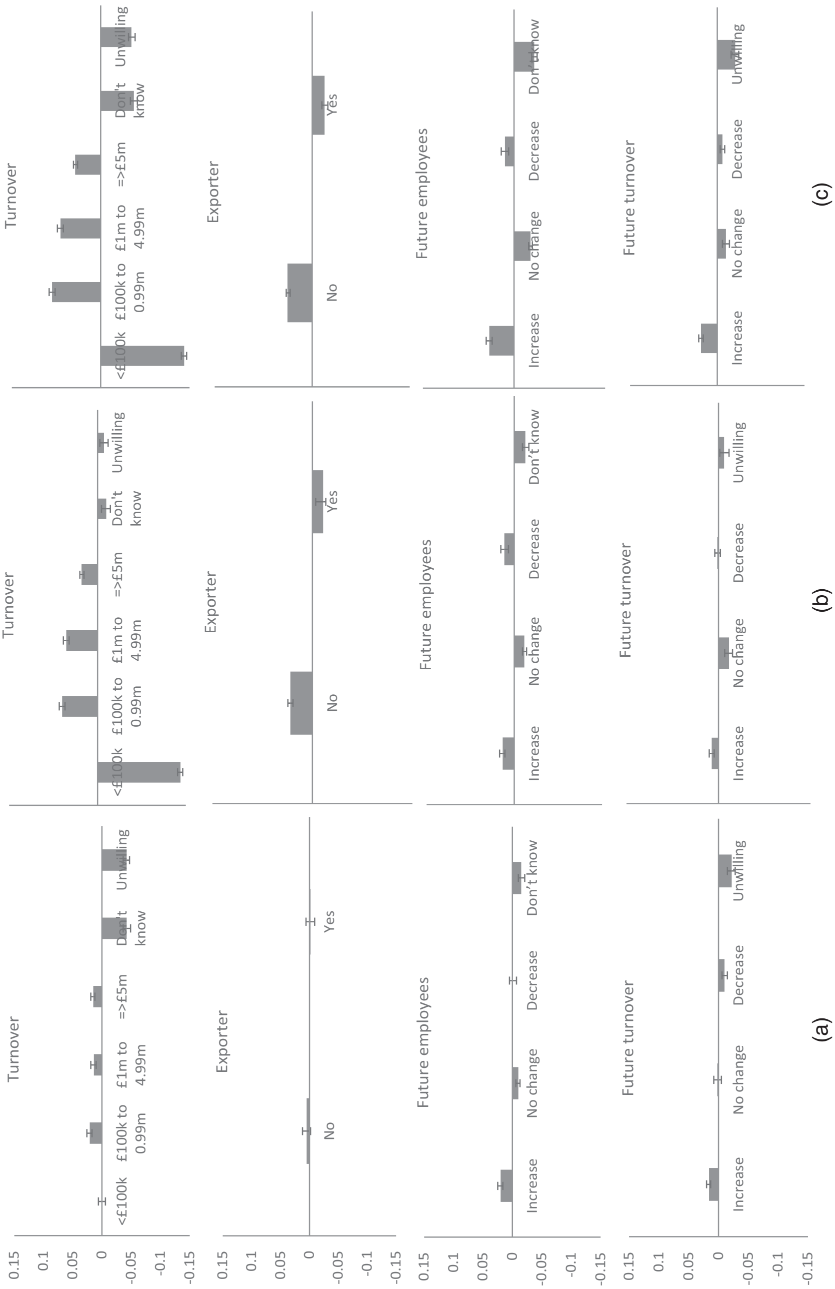

Fig. 1. Partial unconditional covariate category CVs and 95% CIs for (a) the consenter, (b) the consenter with identifier and (c) linkable data sets

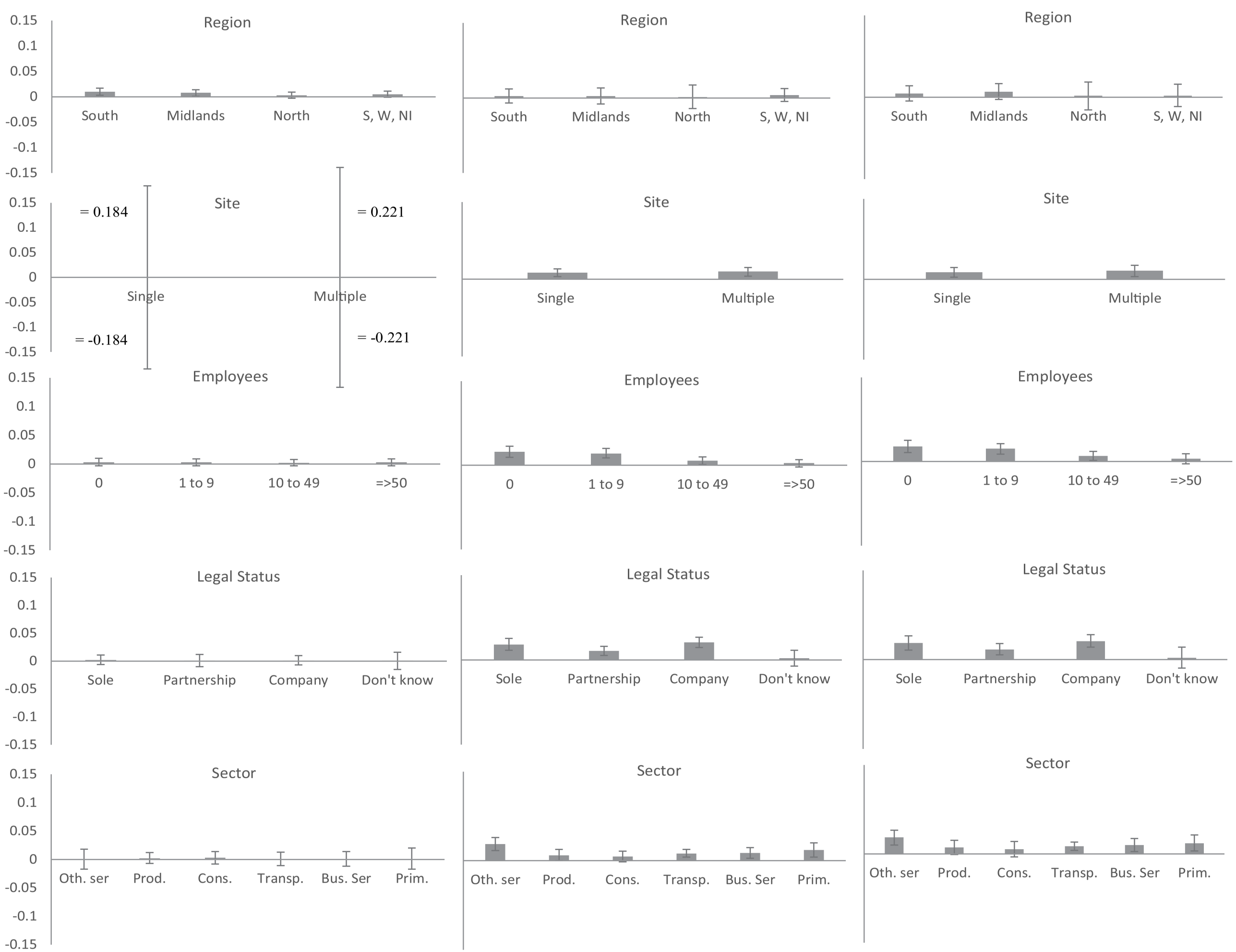

Fig. 2 (*continued*)

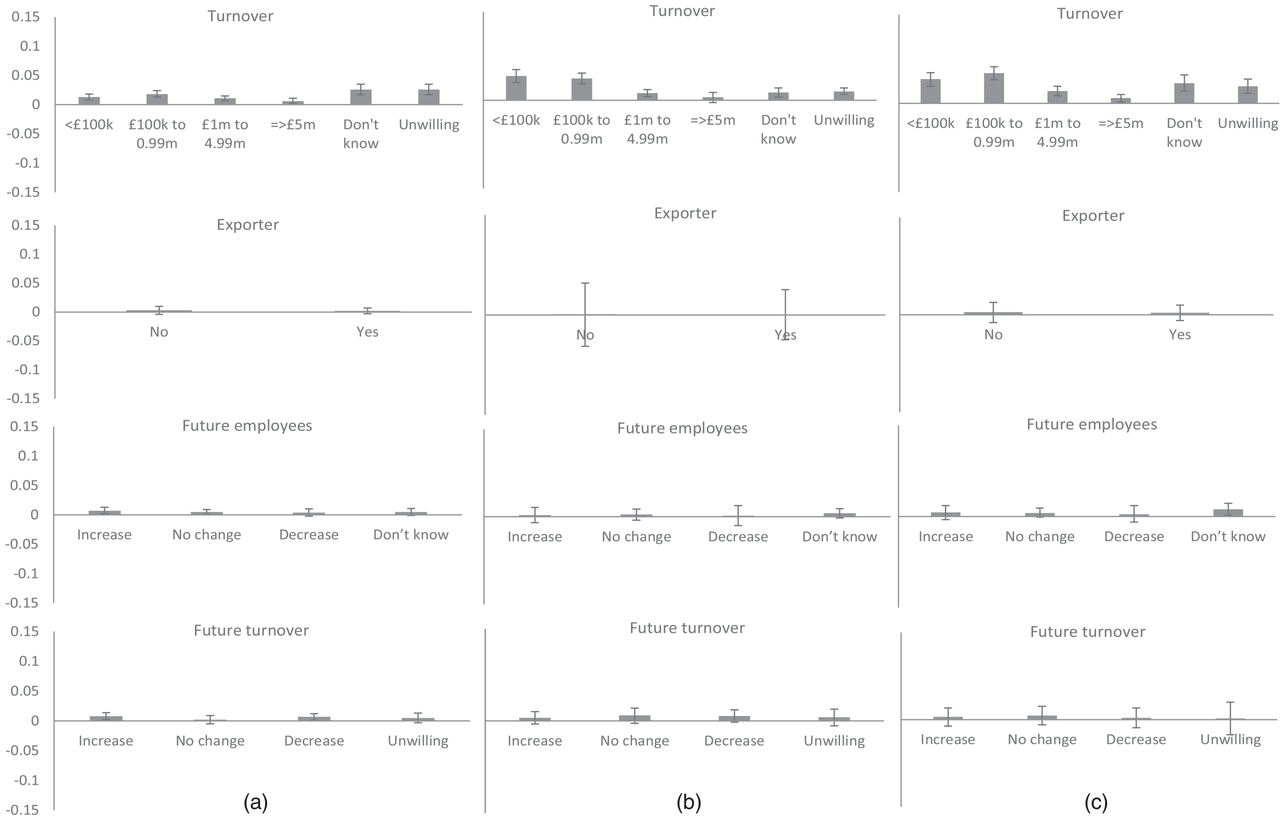

Fig. 2. Partial conditional covariate category CVs and 95% CIs for (a) the consenter, (b) the consenter with identifier and (c) linkable data sets

(though the covariate CV_c also reflects impacts arising if other inequalities are corrected), and those unable or unwilling to report turnover. Similar reasons to those given earlier for the future expected employees result may explain why the future expected turnover CV_u-association is not conditionally significant.

3.2.3. *Risks of non-linkage bias*

The overall CV (0.228; 95% CI 0.206–0.250) suggests that linkable subjects are non-representative of respondents, and more so than consenters are of respondents (the 95% CIs do not overlap). They are similarly non-representative though, as consenters with identifiers are of consenters, consistent with the component contributing most to inequalities. Partial CVs imply similar conclusions (Table 3; Figs 1 and 2). Eight covariate CV_us are significant. Business sector and exporter impacts mainly reflect non-identifier inequalities. Employee number and legal status impacts are similar but are increased by smaller non-consent inequalities. This also holds for future expected employees and future expected turnover, though component inequalities are more equally sized, and turnover, except for the categories don't know and unwilling impacts increased by larger non-consent inequalities. This description of turnover impact causes seems to differ from with the propensity model because they are framed as deviations from representativeness: opposing differences in components between mentioned categories and less than £100000 are visible in plots. For region, which is non-significant in the propensity model, the South, North and Scotland, Wales and Northern Ireland are under-represented, because of non-consent inequalities in the first case and in the other cases non-identifier inequalities not reported earlier because of covariate non-significance. The Midlands is over-represented because of inequalities in both components.

Covariate CV_cs are mostly smaller than absolute CV_u-values. As in the non-identifier analysis, the exception is the non-significant site (in this case, its 95% CI is less inflated). Similar as well, employee number, legal status, turnover and sector CV_cs are significant. Category significance is as for CV_us, except that the categories greater than 49 employees and legal status don't know are no longer significant and primary production newly so. Given causes and correlations between (unconditional) impacts (see also Section 3.1), from this we also identify the same seven under-represented subgroups as in the non-identifier analysis. By CV_u-size, very small business (zero employee, sole owned or partnership and less than £100000 turnover) impacts are again largest. The categories unable and unwilling to report turnover impacts are larger than the other services sector impact. Largest component impacts (production technique modification targets with greatest potential benefits) are very small business non-identifier impacts, followed by unable and unwilling to report turnover non-consent impacts and then the other services non-identifier impact. Very small business non-consent and don't know and unwilling to report turnover non-identifier impacts are smaller again and similarly sized. A non-significant region CV_c despite its CV_u implies correlations with these impacts.

4. Discussion

Data sets linking subject information from multiple sources are often used in economic and other research, but there is little work on estimating non-linkage biases, especially in business data sets, and how they should be evaluated. Such biases can occur when linked–non-linked subject differences, potentially arising in several linkage process components, exist. We address these questions in 2010 UK SBS data sets. We identify survey attribute covariate correlates of linkage propensity and its components: whether subjects consent (consent propensity), and whether an IDBR identifier enabling linkage is appended (identifier propensity). This is the first detailed study of business data set linked–non-linked subject differences. Then, to evaluate biases we report representativeness indicators (CVs of inclusion propensities) for data sets. Given sample information on correlates of inclusion propensities and estimates of interest, these quantify overall data set quality and also impacts of covariate associated inequalities, informing potential production technique modifications. Our use of them in the record linkage context is novel: they have been developed to evaluate survey non-response biases (Wagner, 2012; Schouten *et al.*, 2012).

An understanding of linkable–non-linkable subject differences informs on (causes of) non-linkage biases in both attribute covariates and their correlates, and on production technique modifications to improve data sets (e.g. Sakshaug *et al.* (2013) and Sala *et al.* (2014)). Our univariable propensity models suggest that differences arise for most tested covariates in one or other linkage component (see Section 3.1 for details). Some do not impact on linkage propensity because of variation in the other component: for instance, only region consent propensities differ. Most do though, implying non-linkage biases. Of particular interest out of these are very small business (zero employee, sole owned or partnership, less than £100000 turnover) lower linkage propensities. This suggests, for instance, that linked data set turnover estimates will be positively biased. Similar biases in survey estimates, caused by undercoverage of these entities, are one reason why the SBS is conducted (Department for Business, Innovation and Skills, 2013). Our research suggests that more consideration is needed if they are not to remain an issue with linked SBS data sets (see also Meldgaard *et al.* (2015)).

Our multivariable propensity models quantify conditional differences to inform further on biases in attribute covariate correlates by identifying main impacts. Not all covariates with univariable associations are retained, implying that some are correlated (see Section 3.1 for details). Again, differences in one component may be ameliorated in the other: those expecting future employee or turnover increases have higher consent propensities only (a business performance effect: see also Section 2.1). When linkage propensities differ, one or both components are involved. Identifier propensity differences cause business sector impacts. Employee number, legal status and reported turnover impacts partly reflect identifier propensity differences caused by IDBR very small business undercoverage, but also entity consent propensities are lower. Impacts of those who are unable or unwilling to report items are due to lower consent (probably reflecting a role of informed contacts and data sharing policies) and (unexpectedly) identifier propensities. Any inference from linked SBS data sets must consider how (correlated) estimates are affected by these differences. In addition, they show why all linkage components should be studied. Given information on consent, to improve data sets strategy modifications targeting very small businesses might be proposed but, unless their lower identifier propensities are recognized, benefits will be overestimated. We note though, that the explanatory power of models is low ($R^2 = 6.6\%$ for linkage propensity). This is a characteristic of similar studies (e.g. Meldgaard *et al.* (2015)) and also work on the related issue (both entail subject co-operation) of survey non-response (see Kreuter (2013)), leading researchers to argue that many factors acting in subject-specific contexts determine linkage (response) outcomes. Given this, future work on business linkage propensities should test other possible correlates as well.

Our use of CVs to evaluate non-linkage biases overcomes the issue of partially observed linked secondary data set covariates by using survey attribute covariates only. Biases are presumed small if variation in inclusion propensities estimated given these covariates (which should also be correlates of covariates of interest) is low (see Section 2.3). Overall CVs enable overall data set quality comparisons. Previously, linkage rates have been used (Sakshaug *et al.*, 2013; Sala *et al.*, 2014), but work on survey non-response suggests that linked–non-linked subject differences at high rates will mean that correlations with non-linkage biases are weak (e.g. Groves (2006)). Partial

CVs enable comparisons of attribute-covariate-associated propensity variation: unconditional and conditional versions with parallels to univariable and multivariable propensity models exist that report inequalities in the form of (under- or over-) representation or not. In the SBS data set, these suggest that the largest impacts are (under-representation) of very small businesses in the identifier component. Next are of those who are unable or unwilling to report items in the consent component, followed by that of the other services sector in the identifier component, and then the similarly sized impacts of those who are unable or unwilling to report items in the identifier component and very small businesses in the consent component (see Section 3.2). This information elaborates on that from propensity models. It enables the comparative influence of (causes of) biases to be included in inference and identifies subgroups and linkage components (production technique modification targets) impacting most on data set quality.

We mention that an alternative to modifying production techniques to improve linked data sets is to compute post-data-set production adjustments to reduce non-linkage biases. In fact, it is the only option for users of prelinked data sets, although more generally modifying production techniques may also have the benefit of increasing data set size. Often, methods utilized (for instance, inclusion propensity weights: see Roberts *et al.* (1987) for an introduction in the survey non-response context) assume that subjects are missing at random given an auxiliary covariate set. Consequently, with use of the same auxiliary covariates they should eliminate overall CV-identified non-representativeness (i.e. the biases quantified by partial unconditional CVs): such indicators measure deviations from missing completely at random given a covariate set (see Section 2.3). Impacts of adjustments of this type on other (partially observed) covariates of interest, though, will depend on their correlation with the auxiliary covariates and the extent to which the latter describe linkage propensity variation (a potentially common constraint), as detailed previously. Beyond this, we also note the fully observed auxiliary covariates available in such situations enable use of methods that (with further information or assumptions) quantify and address biases in specific covariates and/or without assuming that data are missing at random, such as selection (McKay, 2012) or pattern–mixture (Andridge and Little, 2011) models. However, these are outside the scope of this paper.

Implications of our introducing CVs to evaluate non-linkage biases are wide ranging. They are easily used by others who are interested in linked data set quality, given consideration of the points emphasized in the previous paragraph. As here, they can be utilized post production to assess biases and contributory factor influence and to inform improvements to estimates. This is of use to both data set producers (ideally, evaluations of this type would be part of data set metadata) and researchers otherwise lacking such information. In addition, they can be utilized by data set producers to identify targets for production technique modifications to improve data set quality. SAS and R code to compute indicators is available from `www.risq-project.eu.` (de Heij *et al.* (2015) and Moore *et al.* (2018); see also Section 2.3). It requires similar levels of knowledge to that of propensity models, and in some ways less as inference on (all) contrasts from the average propensity (the most natural interpretation of differences that are offered by category CVs) entails further programming (e.g. for SAS; Littell *et al.* (2002)). That the perspective that is gained on data sets by using these methods is necessary is shown by our SBS analyses: otherwise scenarios could occur, for example, where production technique modifications are adopted that capture different subgroups so that biases actually increase. In this sense, parallels exist between the study of these biases and the more mature field of survey non-response from where indicators originate (see also Sakshaug and Kreuter (2012)). After empirical work, it is apparent that a focus on inclusion (response) rates as quality measures must be replaced by a more refined approach. Given similarities between problems, we end by noting that it is likely that research in this field can also provide other insights into the issue of non-linkage bias.

Acknowledgements

We thank the Department for Business, Innovation and Skills and the UK Data Service for providing access to the SBS 2010 data set. Neither the Department for Business, Innovation and Skills nor the UK Data Service bear any responsibility for the analyses or interpretations that are reported in this presentation. This work was supported by the Economic and Social Research Council National Centre for Research Methods Workpackage1 (grant reference ES/L008351/1) and the Economic and Social Research Council Administrative Research Centre for England (grant reference ES/L007517/1).

References

Abowd, J. M., Haltiwanger, J. and Lane, J. (2004) Integrated longitudinal employer-employee data for the United States. *Am. Econ. Rev.*, **94**, 224–229.

Abowd, J. M. and Vilhuber, L. (2005) The sensitivity of economic statistics to coding errors in personal identifiers. *J. Bus. Econ. Statist.*, **23**, 133–152.

Andridge, R. R. and Little, R. J. A. (2011) Proxy-pattern mixture analysis for survey nonresponse. *J. Off. Statist.*, **27**, 153–180.

Bean, C. (2016) *Independent Review of UK Economic Statistics*. London: Cabinet Office.

Bjelland, M., Fallick, B., Haltiwanger, J. and McEntarfer, E. (2011) Employer-to-employer flows in the United States: estimates using linked employer-employee data. *J. Bus. Econ. Statist.*, **29**, 493–505.

Bohensky, M. A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D. V., Scott, I. and Brand, C. A. (2010) Data linkage: a powerful research tool with potential problems. *BMC Hlth Serv. Res.*, **10**, 346–352.

Buckley, B., Murphy, A. W., Bryne, M. and Glynn, L. (2007) Selection bias resulting from the requirement for prior consent in observational research: a community cohort of people with ischaemic heart disease. *Heart*, **333**, 196–198.

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference: a Practical Information-theoretic Approach*, 2nd edn. Berlin: Springer.

Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J. and Kott, P. S. (eds) (1995) *Business Survey Methods*. New York: Wiley.

Department for Business, Innovation and Skills (2011) *BIS Small Business Survey 2010: a User Guide for Analysts*. London: Department for Business, Innovation and Skills.

Department for Business, Innovation and Skills (2013) *Small Business Survey 2012: SME Employers*. London: Department for Business, Innovation and Skills.

Dunn, K. M., Jordan, K., Lacey, R. J., Shapley, M. and Jinks, C. (2004) Patterns of consent in epidemiologic research: evidence from over 25,000 responders. *Am. J. Epidem.*, **159**, 1087–1094.

Dunne, T., Jensen, J. B. and Roberts, M. J. (eds) (2009) *Producer Dynamics: New Evidence from Micro Data*. Chicago: University of Chicago Press.

Evans, P. and Welpton, R. (2009) Business structure database: the Inter-Departmental Business Register (IDBR) for research. *Econ. Lab. Markt Rev.*, **6**, 71–75.

Groves, R. M. (2006) Nonresponse rates and nonresponse bias in household surveys. *Publ. Opin. Q.*, **70**, 646–675.

Hayes, R., Ormerod, C. and Ritchie, F. (2007) Earnings: summary of sources and developments. *Econ. Lab. Markt Rev.*, **1**, 42–47.

de Heij, V., Schouten, B. and Shlomo, N. (2015) RISQ Manual 2.1: Tools in SAS and R for the computation of R indicators and partial R indicators. Statistics Netherlands, The Hague. (Available from `www.risq-project.eu`.)

Hijzen, A., Upward, R. and Wright, P. W. (2010) Job creation, job destruction and the role of small firms: firm-level evidence for the UK. *Oxf. Bull. Econ. Statist.*, **72**, 621–647.

Janik, F. and Kohaut, S. (2012) Why don't they answer?: Unit non-response in the IAB establishment panel. *Qual. Quant.*, **46**, 917–934.

Jutte, D. P., Roos, L. L. and Brownell, M. D. (2011) Administrative record linkage as a tool for public health research. *A. Rev. Publ. Hlth*, **32**, 91–108.

Kho, M. E., Duffett, M., Willison, D. J., Cook, D. J. and Brouwers, M. C. (2009) Written informed consent and selection bias in observational studies using medical records: systematic review. *Br. Med. J.*, **338**, 866–873.

Kreuter, F. (2013) Facing the nonresponse challenge. *Ann. Am. Acad. Polit. Socl Sci.*, **645**, 23–35.

Littell, R. C., Stroup, W. W. and Freund, R. J. (2002) *SAS for Linear Models*, 4th edn. Cary: SAS Institute.

Little, R. J. and Rubin, D. B. (2002) *Statistical Inference with Missing Data*. New York: Wiley.

McKay, S. (2012) Evaluating approaches to family resources survey data linking. *Working Paper 110*. Department for Work and Pensions, London.

Meijer, E., Rohwedder, S. and Wansbeek, T. (2012) Measurement error in earnings data: using a mixture model approach to combine survey and register data. *J. Bus. Econ. Statist.*, **30**, 191–201.

Meldgaard, J., Vaze, P., Derbyshire, J. and Davies, D. (2015) Small Business Survey: linking 2006 and 2007 waves to the IDBR. *Research Paper 253*. Department for Business, Innovation and Skills, London.

Moore, J. C., Durrant, G. B. and Smith, P. W. F. (2018) Data set representativeness during data collection in three UK social surveys: generalizability and the effects of auxiliary covariate choice. *J. R. Statist. Soc.* A, **181**, 229–248.

Ritchie, F. (2008) Secure access to confidential microdata: four years of the virtual microdata laboratory. *Econ. Lab. Markt Rev.*, **2**, 29–34.

Roberts, G., Rao, J. N. K. and Kumar, S. (1987) Logistic regression analysis of sample survey data. *Biometrika*, **74**, 1–12.

Sakshaug, J. W. and Kreuter, F. (2012) Assessing the magnitude of non-consent biases in linked survey and administrative data. *Surv. Res. Meth.*, **6**, 113–122.

Sakshaug, J. W., Tutz, V. and Kreuter, F. (2013) Placement, wording and interviewers: identifying correlates of consent to link survey and administrative data. *Surv. Res. Meth.*, **7**, 33–144.

Sala, E., Burton, J. and Knies, G. (2012) Correlates of obtaining informed consent to data linkage: respondent, interview, and interviewer characteristics. *Sociol. Meth. Res.*, **41**, 414–439.

Sala, E., Knies, G. and Burton, J. (2014) Propensity to consent to data linkage: experimental evidence on the role of three survey design features in a UK longitudinal panel. *Int. J. Socl Res. Methodol.*, **17**, 455–473.

Schnell, R. (2013) Linking surveys and administrative data. In *Improving Survey Methods: Lessons from Recent Research* (eds U. Engel, B. Jann, P. Lynn, A. Scherpenzeel and P. Sturgis), pp. 273–287. London: Routledge.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N. and Skinner, C. (2012) Evaluating, comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators. *Int. Statist. Rev.*, **80**, 382–399.

Schouten, B., Cobben, F. and Bethlehem, J. (2009) Indicators for the representativeness of survey response. *Surv. Methodol.*, **35**, 101–113.

Schouten, B., Cobben, F., Lundquist, P. and Wagner, J. (2016) Does more balanced survey response imply less non-response bias? *J. R. Statist. Soc.* A, **179**, 727–748.

Schouten, B. and Shlomo, N. (2017) Selecting adaptive survey design strata with partial R-indicators. *Int. Statist. Rev.*, **85**, 143–163.

Schouten, B., Shlomo, N. and Skinner, C. (2011) Indicators for monitoring and improving representativeness of response. *J. Off. Statist.*, **27**, 231–253.

Shlomo, N., Skinner, C. J. and Schouten, B. (2012) Estimation of an indicator of the representativeness of survey response. *J. Statist. Planng Inf.*, **142**, 201–211.

Snijkers, G., Haraldsen, G., Jones, J. and Willimack, D. (eds) (2013) *Designing and Conducting Business Surveys*. Chichester: Wiley.

Wagner, J. (2012) A comparison of alternative indicators for the risk of nonresponse bias. *Publ. Opin. Q.*, **76**, 555–575.

Willimack, D. K., Lyberg, L., Martin, J., Japec, L. and Whitridge, P. (2004) Evolution and adaptation of questionnaire development, evaluation, and testing methods for establishment surveys. In *Methods for Testing and Evaluating Survey Questionnaires* (eds S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin and E. Singer), pp. 385–407. New York: Wiley.

Winkler, W. E. (1995) Matching and record linkage. In *Business Survey Methods* (eds B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge and P. S. Kott), pp. 355–384. New York: Wiley.