The perils of pre-filling: lessons from the UK's Annual Survey of Hours and Earning microdata.

Damian Whittard[a]*, Felix Ritchie[a], Van Phan[a], Alex Bryson[b], John Forth[c], Lucy Stokes[d], and Carl Singleton[e]

[a]Data Research, Access, and Governance Network, University of the West of England, Bristol, UK; [b]Social Research Institute, University College London; [c]Bayes Business School, City, University of London; [d]National Institute of Economic and Social Research; [e]University of Reading

*Corresponding author: Frenchay Campus, Coldharbour Lane, Bristol, BS16 1QY, UK. Tel ++ 0044 117 3287140 damian2.whittard@uwe.ac.uk

Abstract

The role of the National Statistical Institution (NSI) is changing, with many now making microdata available to researchers through secure research environments. This provides NSIs with an opportunity to benefit from the methodological input from researchers who challenge the data in new ways. This article uses the United Kingdom's Annual Survey of Hours and Earnings (ASHE) to illustrate the point. We study whether the use of prefilled forms in ASHE may create inaccurate values in one of the key fields, workplace location, despite there being no direct evidence of it in the data supplied to researchers. We link surveys to examine the hypothesis that employees working for multi-site employers making an ASHE survey submission are more likely to have their work location incorrectly recorded as the respondent fails to correct the work location variable that has been pre-filled. In the short-term, suggestions are made to improve the quality of ASHE microdata, while longer-term, we suggest that the burden of collecting additional data could be offset through greater use of electronic data capture. More generally, in a time when statistical budgets are under pressure, this study encourages NSIs to make greater use of the microdata research community to help inform statistical developments.

Keywords: microdata; response burden; measurement error, ASHE; spatial

1. Introduction

The National Statistical Institute's (NSI) core purpose has been to produce aggregate statistics (e.g., National Accounts) with most functions set up to support such activity. This provided the NSI with a clear mandate to collect sufficient data to produce high quality statistics, while simultaneously ensuring that it limits the burden on respondents [1]. Although still guided by the Fundamental Principles of Official Statistics, which were adopted by the European Commission and the United Nations Statistical Commission in the early 1990's [2], the role of the NSI is evolving. This century the role of the NSI has expanded to be a key provider of microdata. For example, it is suggested that linking of disparate datasets across time, space and sources is probably the foremost current issue facing NSIs [3]. Recent developments have improved access for some countries, although progress is still patchy [4].

One of the reasons why this expanding role may be problematic is that analysts are likely to use the data in quite different ways to official statisticians, and this has implications for both data preparation and quality assurance [5]. Given this changing role of NSIs and the new uses of microdata, understanding how researchers can positively support NSIs in data development is vital.

In this paper we use a case study of researcher quality assurance (QA) from the United Kingdom's Annual Survey of Hours and Earnings (ASHE), collected by the UK Office for National Statistics (ONS). ONS prefills some information in ASHE, primarily to reduce respondent burden. This has been identified as a potential source of inaccuracy, but as there is no independent source of information which collects this data directly, there is no direct evidence for it, and we are therefore unable to identify the precise magnitude of the error. Our case study, however, demonstrates how the researchers identified the potential problem, established its likely existence, quantified the potential impact and fed back to ONS, as the findings have implications for geographically-based official statistics in the UK. As such, this case study also has broader implications, suggesting that in their evolving role, NSIs should work more closely with the microdata research community to help inform statistical developments. Such alliances can help to identify quality issues, point towards solutions and generally help to improve data quality and insight for researchers both within and outside NSIs.

The next section considers the pros and cons of prefilled forms. Section 3 describes the identified problem and its cause, namely a prefilled survey form entry, while Section 4 introduces the

ASHE dataset in more detail. Section 5 presents the analysis, which is followed by discussion on its implications for NSIs. Conclusions are drawn in Section 7.

2. Why prefill? Balancing conflicting needs

In a time when users are demanding more timely and more disaggregated information, and while the budgets for many statistical agencies have been stagnant or declining, prefilling answers on a survey can offer a number of benefits [6]. Providing prefilled answers to questions which are fixed or just change irregularly can reduce the burden to the respondents, cut processing costs, while also potentially improving the quality of the data.

2.1 The benefits of prefilling

NSIs have a responsibility to limit the response burden, with the three main reasons cited as bureaucratic (business competitiveness), methodological (data quality and collection costs), and strategic (relationship between NSI and respondent) [7, 8]. Prefilling forms can address all three reasons, by saving businesses time, by reducing the chance of erroneous entry (in our case, of mis-spelled postcodes), and by being seen to help the respondent by not asking for information the NSI already has, unless that information is incorrect. The use of previously reported data (PRD) in surveys is not new, with a literature going back to the mid-20th century [9]. For example, one study presented an experiment using PRD in the US Census of Agriculture's Content Test [10]. It reported that the majority of respondents view PRD as aiding in faster and easier survey completion, as well as having an overall positive reaction to its use. Others studied PRD in self-administered (paper or web-based) questionnaires for business establishments [11]. They reported that questionnaires with pre-printed values outperform questionnaires without them.

2.2 The risks of prefilling

Prefilling, however, does come with its own problems. It can reduce data quality and increase measurement error due to underreporting of changes and conservation of errors; generate a loss of confidence and goodwill if the pre-printed data is of poor quality; and increase disclosure risk [10, 12]. Much of the literature focusses on the costs and benefits of independent interviewing (respondents provide new data each survey) to dependent interviewing (DI) (respondents are shown previous responses and only asked to amend it if this has changed). One study compared proactive and reactive DI with independent interviewing for a range of survey items [12]. It concluded that the decision whether

to adopt DI techniques should weigh net costs against effects on data quality, including effects on estimates of change, item nonresponse, attrition, reliability and accuracy [13]. Given the concerns in relation to accuracy, this would then suggest that the potential benefits of prefilling are not, in themselves, justification enough to do so.

2.3 Balancing the needs

While evidence needs of policy makers can be enhanced by users of the microdata, who are becoming more proficient and innovative of their use of standalone and linked datasets, the trade-off between burden and accuracy has become less clear. As collectors, processors and guardians of the data, the NSIs can be somewhat removed from the use of the microdata. This paper therefore argues for closer alliances between the NSI's and microdata users, in order that decisions made around response burden take a full account of downstream benefits, rather than just considering the traditional needs of the NSI.

To illustrate the point, we add to the longstanding traditions of using case study approaches to demonstrate the use of microdata [14]. In support of the assertion that geography is more important for NSIs as they start to explore new data sources and ways to integrate those to study relationships [3], we explore the properties of ASHE to provide accurate spatial, labour market data. This is because, in order to limit the response burden, the NSI issue surveys (both paper and web-based) with geographic information already pre-filled, with the likelihood that this leads to inaccuracies in the data, making spatial analysis challenging.

Although the issue of pre-filling has been widely researched, the literature largely concentrates on response times and completion versus non-completion, or the reduced likelihood of entry errors. There appears to be no literature to determine whether the prefilled values are more accurate than empty fields which have been completed. This may be because there is no way to identify the true value from the completed form, making this analysis challenging. This case study, therefore goes someway to addressing this gap in the literature.

3. Problem statement

In this study we explore the impact of prefilling a key field, the workplace location variable. We do so by analysing the ASHE, which is a vital data source for UK government and academics to understand changes to labour markets, wage differentials and earnings growth. Sampling in ASHE is

based on National Insurance number (NINo - the social security number for all adults in the UK), with the same individuals being observed every year they are employed. The data is supplied by employers from payroll records, making the data high quality as it does not suffer from recall bias [15]. The accuracy and the longitudinal nature of the data make it the most important UK dataset for understanding earnings [16] and it is central to minimum wage policy [17].

One key advantage of ASHE over other data sets is that it contains information on both the employee's home and work address at detailed (postcode) level, from which higher-level geographies are defined. These location variables have been used to shed light on important aspects of labour supply and to develop labour market policy. For example, ONS reports median full-time gross weekly earnings by region, showing that the poorest performing region recently was the North East of England whose earnings fell by 1.4% [16]. ONS studied gender differences in commuting time and pay [18] and reported that men tend to have longer commutes, while for women commuting time has a greater impact on the decision to leave one's job. Another study reported that graduates tend to move to places with high average earnings, such as London, and those that grew up in places with low average earnings are more likely to move away [19]. ASHE location markers were also used in an analysis of commuting and the effects of road improvements on individual labour market outcomes [20]. The authors found a positive impact of improved accessibility on weekly wages and total hours worked. This contrasted somewhat with one study that concluded that area effects contributed to just a very small percentage of the overall variation in wages in Britain [21].

The diversity of examples demonstrates the variety of agendas for which the location variables are used. Accordingly, questions over the accuracy of the spatial data raises concerns over the inferences that were made in these publications, as well as official statistics describing geographical breakdowns of earnings.

The concern arises from processes put in place to reduce respondent burden. Data is collected via a survey (paper and web-based). The effect of survey mode on data quality has been widely investigated, with conflicting results [22, 23, 24, 25]. However, our study does not differentiate between paper and web-based surveys, but explores the differences between survey responses (paper and web-based) and 'special arrangements'. Approximately 350 companies are allowed to make a computerised submission as a dump from payroll records via what ONS terms 'special arrangements'. In this article,

we refer to ‘special arrangements’ as an electronic submission, which is distinct from a web-based survey response.

The survey approach requires employers to provide work and home postcodes. This address is pre-filled on the ASHE survey form from HM Revenue and Customs (HMRC) records, and employers are asked to change it if the actual location is different to the pre-filled one. However, the employer address is that which is registered for tax purposes, which is not necessarily the work location, especially for multi-site organisations. There is an incentive (saving time) for employers not to update this field, and discussions with the Northern Ireland Statistics and Research Agency about their similar survey suggested that uncorrected work location was a concern for them. We note that, in theory, pre-filling home address may also lead to measurement error if the employee has moved house since tax records were updated, but we assume this is a rare event.

Without knowing the “true” value of the work location variable, it is challenging to demonstrate the existence or quantify the level of any systematic measurement error. Therefore, we used a number of indirect approaches to uncover the potential error, such as number of employees apparently located at the head office, or distances travelled between work and home address. The data also enables us to identify single site and multi-site organisations, and those who complete the survey (on paper or web-based) as opposed to those who provide ONS with a direct readout from their employment records (special arrangements/electronic submission). The latter group do not have pre-filled fields, and on the assumption that they are likely to accurately record employees’ work location, these four groups (single/multi-site, survey/electronic) allow us to assess whether there is potential for systematic measurement error in ASHE.

The hypothesis is that employees working for multi-site employers who make an ASHE survey submission are more likely to have their work location incorrectly recorded. The evidence for that will be a higher proportion of staff recorded as working at the registered ‘head office’ and longer commuting distances. Our assumption of longer commuting distances is based on the idea that, on average, the head office should be further away from the employees’ home address than their actual workplace location. For example, imagine a multi-site retail company with 50 sites distributed evenly across the country, and all with an equal share of employment at local retail units. If 100% of employees are mistakenly reported as working at the head office, there will be inaccuracy in the work location variable.

On average, this inaccuracy will appear as longer commute distances, even though a small percentage of employees will live closer to the head office than their actual workplace location.

This matters; for example, the wellbeing literature reports that personal stress is associated with commuting [26]; commuting has the lowest positive affect score and one of the highest negative affect scores of all activities completed in a day [27]; and individuals are putting increasing value on savings in travel time [28]. Given these findings, it is not unrealistic to expect individuals to choose to limit their commute time and opt to work close to home. As such, if their work location is 'incorrectly' recorded – (i.e., firms PAYE registered office - proxy for the head office), more often than not, this will mean that the individual appears to have a longer commute.

4. Data construction and methodology

The following sub-section describes the characteristics of both the businesses included within the sample, and the number and proportion of employees that work for them.

Table1: Number and percentage of individuals working in enterprises, by characteristic of employer.

<Insert Table 1 Here>

Table 1 shows that employees of multi-site enterprises are more likely to be employed in large businesses and those that are based in an urban location. Although there are less people working in multi-site organisations in the public sector (compared to working in the private sector), as a proportion there is a considerably higher percentage of its employees working for organisations with multi-sites (91.5% - calculated by adding Column B + Column D). There is relatively little difference between the proportion of individuals employed at multi-site businesses located inside and outside of London.

When comparing the proportional breakdown of those employed in multi-site businesses making electronic submissions (as opposed to multi-site businesses making survey submissions), apart from SME's where just 0.4% of employees in multi-site companies are covered by electronic submissions

(calculated by dividing Column D by the total of Column B and D), the proportions of employees working for all the other seven categories are broadly similar (i.e., large/public/private/UK excluding London/London/rural and urban). This ranges from 16.2% of employees working for rural businesses to 26.9% of those working in the public sector.

4.1 The potential for measurement error

Individuals included in the ASHE sample are identified as working for a specific enterprise through HMRC's tax records. Therefore, the work address of the employee is initially recorded as whatever is registered for a company's Pay As You Earn (PAYE) scheme, which in many cases will be the enterprise's head office. ONS send out the ASHE questionnaire to employers with the employees' work addresses pre-filled; if this is incorrect, respondents are required to change it. Given the implied additional burden involved in finding and supplying the postcode for an employee's local office, it is unsurprising that some respondents may take the least-cost option and not check entries or change erroneous ones. Over half the observations in ASHE (54%) relate to individuals who work for multi-site businesses who complete survey forms, equating to over 18,000 enterprises (29% of all enterprises). For multi-site companies, this therefore means that the potential for systematic measurement error in this location variable has been built into the system – particularly as some firms are required to complete survey forms for multiple employees. The challenge with testing this proposition is that there is no independent source of information that collect data on the employees work location, and therefore we are unable to know the real size of any error.

If this measurement error in workplace location was built into the system, however, it would be partially offset by the *'special arrangements'* for collection of ASHE data, which ONS has with some of the largest employers. The exact eligibility criteria for qualifying for special arrangements is not published by ONS, but they state on their website that "ONS has a special arrangement with some very large employers" [29]. Instead of completing a form for each employee in the ASHE sample, enterprises that qualify for special arrangements provide an electronic submission extracted from their employee records. In these cases, the accuracy of workplace data should improve for two reasons: first, the workplace postcode is not pre-filled and needs to be entered; second, enterprises are reading workplace location from their own staff records, rather than PAYE location.

This provides the difference we exploit to test the hypothesises. If the workplace location data is universally valid, then we work on the assumption that there is little reason to expect differences in commute distance between single site employers, irrespective of whether they make a survey or an electronic submission, via special arrangements. Likewise, we would expect very little difference between a multi-site organisation that makes either a survey or an electronic submission.

We are not aware of any studies on commuting patterns in relation to different types of firms, however, based on the findings from the wellbeing literature, given the different structure between single site and multi-site businesses, there is an expectation that there may be some difference in the commute distance between single site- and multi-site organisations. However, if firms providing survey responses do not edit the workplace (head office) address, then we would expect differences to manifest both within multi-site responses and between multi-site and single site responses. For example, Model A in Figure 1 depicts the differences we would expect to see in commuting distances without any measurement error. Model B uses red arrows to illustrate where the (inflated) differences would show up, if measurement error were present.

<Insert Figure 1 Here>

We examine two indicators of measurement error: implied commuting distances, and the share of employees located at the 'head office' compared to the official record of the number of employees at the head office location from ONS' Inter-Departmental Business Register (IDBR). The IDBR is the sampling frame for all ONS' business surveys and accordingly substantial effort is expended on keeping it up to date and accurate. The IBDR is the only other indirect source of employee work location, and therefore there were no further opportunities for triangulation.

It is also worth noting that a potential limitation of the study is that it is possible that unobserved characteristics could contribute to differences in the proxy for commuting distance across groups. For example, if firms in the special arrangement group are more actively recruiting local workforce (living closer to their workplace), this could this reduce their commuting distance. On the other hand, if survey (multi-site) firms have higher turnover and this makes it more difficult for workers to relocate, this would result in longer commuting distances. However, these are not testable hypotheses with these datasets; and so, while acknowledging these limitations, we take the simplest explanation and interpret statistically significant differences as evidence of measurement error.

For our purposes, we treat ‘head office’ as ‘PAYE location’ as we cannot distinguish them. We also do not investigate businesses (single site or multi-site) who may have a recorded ‘head office’ but no employees, or a head office with no employees and separate from all of their establishments – we treat these as shell companies.

4.2 Data linkage

To assess the validity of the ASHE workplace variable, four data sources were linked for the years 2016 to 2018. The analysis was restricted to these three years as the ‘special arrangements’ marker used in the analysis was only available from 2016, while 2018 was the latest year made available to the researchers.

The four data sources were the ASHE annual datasets [30], Eastings and Northings data (geographic point identifiers), and two versions of the Business Structure Database (BSD) – Enterprise dataset and Local Unit dataset [31].

The Eastings and Northings data, derived from postcodes, provide geographic point identifiers, which allow for the distance to be calculated in kilometres between ASHE home and work address. This allowed us to estimate commuting distances between individuals’ homes and their local workplace as a straight line between the two points.

The BSD datasets are annual snapshots of the IDBR constructed for research purposes and both enterprise (company) and local-unit (establishment) level. The assumption underpinning the analysis is that the number of employees by location data recorded in the BSD is likely to be recorded most accurately, compared to survey data collected in ASHE.

The BSD enterprise data contains the location of the registered office. This, in combination with the address of local units, allowed for the creation of a proxy ‘head office’ marker used in the analysis.

Linking the BSD between enterprises and establishments, and between the BSD and ASHE, was challenging as (a) the BSD data only contains partial postcodes and (b) there is no direct linkage variable between the BSD local unit dataset and ASHE local unit identifier. To overcome this, geographic markers were created based on census output areas (COAs) available within both the ASHE and BSD datasets - COAs are geographical units with roughly 600 residents. While ASHE workplace location could not therefore be mapped exactly to a registered establishment, COA was felt

to be small enough for all practical purposes. We were then able to assess the extent to which employment is centralised in the head office and compare this to the distribution of employees in ASHE.

5. Analysis

5.1 Do the numbers at head office provide evidence for potential measurement error?

Table 2 compares the data from the same organisation as listed in the BSD and ASHE, with the restricted sample. We create a restricted sample by excluding businesses where no employment was recorded, either at the head office or any local units – typically known as a shell company. We also exclude observations where there was no match between the location of the head office and the location of their local units, even though the company registered some employment in their local units and/or head office. In the matched dataset this equated to approximately 6,400 enterprises (10 % of ASHE sample), employing approximate 23,000 people (13% of employees). This could be for several reasons: some kind of matching issue across ASHE and BSD; delays between datasets updating records; or potentially an unusual (e.g., shell) company structure. Given the complexities of the IDBR, it beyond the scope of this paper to identify the exact reason, but given our methodology for identifying head office, these observations are excluded from the analysis and, for short-hand purposes, furthermore, referred to as 'shell companies'. It is worth noting, however, that this is a potential limitation of the study, as the excluded firms may not be completely random.

We can see that, for example, in the restricted sample both the BSD and ASHE show 100% of employees (on average) are based at the head office for micro firms under ten employees, and that this proportion steadily falls as the number of employees increases. For completeness, we include 'unmatched' observations: these are enterprises recorded in ASHE but for which no corresponding record exists in the BSD; these are believed to be due to the different sample selection times, ten months apart.

<Insert Table 2 Here>

The restricted sample shows that the percentage of employees apparently working in the head office is greater in the ASHE than the BSD, for all but the very smallest organisations (0-9 employees). These differences between ASHE and BSD persist however the numbers are analysed. For example, analysis by number of establishments, and submission type supported the general finding that ASHE always overestimates the numbers at the 'head office' location, compared to the

BSD [32]. Given the considerable effort and array of data sources used to create and maintain ONS' business register, we consider the BSD figures to be a good approximation for the 'true' value; as such, the results suggest potential workplace location measurement error in the ASHE data.

5.2 How far do employees travel to work?

The following analysis decomposes the ASHE data to demonstrate the possible structural effects arising from the potential systematic measurement error in the workplace location variable. The hypothesis that firms completing surveys are less likely to edit erroneous prefilled workplace (head office) addresses implies artificially long commute distances. We do not have exact commute time, but we calculate, as a proxy, the straight-line distance from home to work address.

Table 3 reports the average proxy distances employees travel to work, for those working at single site and multi-site enterprises, with and without special arrangements (electronic submission). If the hypothesis is correct, there should be statistically significant differences between single- and multi-site survey respondents, and between multi-site survey and multi-site electronic submissions (i.e. special arrangements/payroll dump).

<Insert Table 3 Here>

The results are consistent with potential systematic measurement error in the workplace location variable. For example, employees working for single site organisations (making both survey and electronic submissions) and multi-site enterprises with electronic submissions have mean travel-to-work distances of 16 kilometres or less (column 2). The data for single site electronic submission is notably lower, but this is not necessarily informative: there are relatively few observations here, and an organisation which is simultaneously based on one site, but still large enough to merit special arrangements, is likely to be unusual.

In contrast, employees who work for multi-site organisations completing surveys (where systematic measurement error is hypothesised) are reported as living a mean distance of 28.2 kilometres from their reported workplace. If this were correct, then employees in these types of organisations would travel on average an additional 24 kilometres per day to commute to and from work. This difference persists across the distribution, with commuting distances longer at every percentile for these organisations compared to all other groups.

In order to interrogate the distribution of commuting distances, Figures 2 and 3 plot the Kernel density estimates of the natural log transformation of the distance travelled to work for the four main groups: single site and multi-site enterprises, with and without special arrangements. Logged values are used to improve the statistical properties to account for some large outliers.

<Insert Figure 2 Here>

Figure 2 plots the distribution for single site enterprise only. It shows that the distribution for single site companies, whether making an electronic or survey submission broadly mirror each other, albeit single site companies making survey submissions have longer tails. Potentially this may be the result of greater heterogeneity in the companies making survey submissions. For example, the left-hand tail may represent small home businesses, whereas the right-hand tail may represent services companies. It also reflects the considerably smaller number of observations in the electronic single site category.

Figure 3 plots the kernel density plot of distances travelled to work by employees working for multi-site companies making survey and electronic submissions.

<Insert Figure 3 Here>

Figure 3 clearly shows that there is a systematic difference in the distributions between multi-site companies making survey submissions and multi-site making electronic submissions. Those making survey submissions report that their employees work further away from home than those making electronic submissions at all points of the distribution. The distribution of multi-site companies making electronic submissions more closely aligns with the distributions of the single site organisations - e.g., the most frequent estimate (mode) of the natural log transformation of the distance travelled to work is approximately equal to 2 in all three cases, equating to approximately 7.5 km. This provides a clear indication that employees working for multi-site employers who make an ASHE survey submission are more likely to have their work location incorrectly recorded.

5.3 To what extent do the characteristics of the firm impact on the proportion of head office employment and reported distances travelled to work?

A priori, our expectation was that if there were evidence of systematic measurement error for multi-site companies making survey submissions, it would increase in relation to the total number of units and total number of employees of the enterprise. This is because the administrative burden of identifying

the correct site and altering the pre-filled questionnaire would increase as units and employment increased.

Related to this was the expectation that there would also be industry specific effects, particularly between public sector and private sector enterprises where the incentive to invest time to respond to national surveys may differ. For example, profit maximising firms may focus on the private cost to the firms, whereas those in the public sector may be more motivated by the social benefit of producing statistics as a public good. To further explore which factors are independently associated with the proportion of head office employment and distance travelled to work for multi-site enterprises, a number of regressions were run on the ASHE 2018 dataset, both at the level of the enterprise (Table 4) and the level of the employee (Table 5). The regression studies the characteristics of the firm (not the employee – e.g., age gender), as the hypothesis is that the firm's response is systematically biased. The results were checked by running similar regressions for 2016 and 2017 years, as well as a pooled regression for years 2016-2018. The results for all years report similar findings and are reported in the supplementary material section.

5.3.1 Enterprise level regressions – proportions at the head office

Table 4 reports the regression results for proportion of head office employment for multi-site companies only. In terms of the regressions, where there were multiple categories for each of the covariates, the largest group was always omitted. Columns 1 and 2 are both Ordinary Least Squares (OLS) linear regressions, while column 2 reports the results of a restricted model that excludes potential shell companies. Column 3 also restricts the sample but uses a Tobit (censored) model to account for the considerable number of enterprises with either zero employees (shell companies), or 100 percent of employees recorded as working at the head office.

<Insert Table 4 Here>

The main variable of interest is 'special arrangements'. Controlling for all other factors, and in all specifications, the results show that companies that have special arrangement in place (payroll dump/electronic submission) have a lower proportion of employees working at the head office, and that this difference is significant at the 1% level. As long as the option to have special arrangements is not correlated with the proportion of head office employment for some other unknown and unobserved

reason, this provides clear support for our hypothesis that there is systematic measurement error in the reported work location for multi-site companies making survey submissions.

To explore industry specific effects, we use common sector classifications as dummy variables. This is typical of labour market studies using ASHE, to mop up broad sectoral differences when the focus is on other correlations. As part of our robustness checks, we interacted special arrangements with the sector variables. This revealed that there were some significant effects (i.e., potential for bias) for special arrangements in some sectors (e.g., construction; sales; services; financial/law; creative), but less so in others. This indicates that there may be more complex interactions within sectors. This is beyond the scope of the study and therefore for simplicity, we chose to present a familiar set of sector variables to illustrate the potential for bias. As such, models 2 and 3 show that enterprises from all sectors compared to the public sector are more likely to report higher proportions of employment at their head office, with the sector controls all jointly significant at the 1% level. If we accept the premise that the response from the public sector (and health sector) is likely to be of the highest quality [33], then again this provides further support that there may be systematic mismeasurement, and this is most common in a number of sectors from the private sector.

Confirming the results reported in Table 2, lower proportions work in the head office as the size of the company increases (number of employees), although the effect is small. Excluding companies with zero HQ employment, both the OLS and censored regression results are significant at the 1% level. In terms of local unit group size, the greater number of local units in an enterprise group (2-5 was the excluded group), as expected, the lower the proportion of head office employment; this result was again significant at the 1% level.

The regional variable shows that if an organisation has its head office in London (the omitted group) it has a lower proportion of employees registered as working in the head office. This may be explained by the increased cost of workspace and labour in London, making it more efficient for organisations to outsource work from the capital to its other work locations.

The results on registered status report that those working in sole proprietors, partnerships, local authorities and non-profit bodies all record higher proportions working at the head office than registered companies, while central government bodies have less, all other things being equal.

5.3.2 Individual level regressions – travel-to-work distance

To understand what factors influence the recorded distance that employees live away from work, Table 5 reports a regression based on each individual ASHE return for the full sample (columns 1 and 2) and for a restricted sample using observations from just employees working at multi-site companies (columns 3 and 4).

For all four models reported, errors are clustered to allow for multiple employees working at the same organisation. The dependent variables in columns 1 and 3 are in actual terms (kilometres) and in logged terms for columns 2 and 4.

<Insert Table 5 Here>

The results across all four model specifications appear robust, in as much as the coefficients, for the most part, all have the same sign. The logged model appears a better fit given the heightened significance of several control variables. As the logged model better handles the skewed distribution and the small but significant number of large outliers, this is our preferred specification.

The results from the regressions provide additional support to the hypothesis that employees working for employers who make an ASHE survey submission are more likely to have their work location incorrectly recorded. The variable of interest, 'special arrangements', is significant across all specifications. Controlling for all other factors, individuals who work for enterprises with 'special arrangements' are recorded as working approximately 10 km closer to work, than individuals who have had their information provided using a survey submission. This finding is significant at the 1% level and holds both for the full sample and for just those employees working at multi-site enterprises. We conjecture that the reason for the discrepancy is that, in some instances, the ONS pre-filled postcode location in the survey questionnaire is the proxy head office (registered tax address) and not the workplace postcode for that employee.

The logged model (columns 2 and 4) supports this main assertion. Using the Halvorsen-Palmquist correction to interpret dummy variables in semilogarithmic equations [34], the results indicate that employees who have had their information provided by 'special arrangements' live approximately 31% closer to their work than those who do not, albeit with lower significance than in the non-log model. This result indicates the presence of measurement error for multi-site companies making survey

submissions and therefore has important implications for any analysis undertaken using ASHE's workplace variable.

The full models (columns 1 and 2) report that those who worked for a single site organisation lived approximately 6km, or 28%, closer to work than those who worked for multi-site employers. Intuitively one may expect employees of multi-site companies to live closer to their employer, as the sites are more distributed. Given the assumption that the work location variable is correct for single site employers, after controlling for special arrangements, the contrary result further suggest the presence of measurement error and that caution should be shown when using the workplace variable for multi-site organisations.

The logged models (columns 2 and 4), indicate that the sector is an important component in determining the distance travelled to work. For example, the results show that employees of seven of the 10 sectors for the multi-site only models are recorded as working further away than public sector employees; six of which were significant at the 1% level. Indeed, construction workers were likely to work more than double the distance from their work than public sector workers, with employees from the finance/law, utilities and primary sectors working over 50% further away. There can be a spatial dimension to service provision that gets mixed up in industry sector – for example schools and GP surgeries need to be near residential areas, whereas business services and large-scale manufacturing plants may not, while construction firms are likely to employ workers in a wide variety of 'temporary' locations. While acknowledging that there are various reasons travelling differences may be greater for individuals working in some private sector industries, with an expectation that the responses from public sector firms are likely to be of the highest quality [33], the differences indicate further support of the assertion that there is likely to be systematic measurement error in the workplace location.

Employees working for enterprises outside the main urban cities generally worked further away from home than those that worked in the main urban cities; this was significant for five of the nine urban/rural groupings (column 4). This is in line with a study which analysed the national travel surveys for the UK and the Netherlands and concluded that urban structures contributed to long-distance commuting and business travel [35]. Conceptually the result seems sound, but there is limited implication for workplace analysis from this observation.

In terms of size of employer, the results indicate that employees worked closer to their employer as the size of the employer increased. The regressions suggest that as the employer grew by 100 employees, employees work approximately 1 km closer to their employer. Conceptually, this could also be explained by the fact that larger, and potentially more productive employers, are better able to embed themselves in the local economy and make themselves a more attractive proposition to the local labour force. Therefore, similarly to the urban/rural groupings, the size of employer is a factor in workplace analysis.

The number of local units was also an important factor in determining the distance travelled to work. The results suggest that the more local units a company has, the further away an employee lives from work. For example, compared with the base category (two to five local units), those working for enterprises with 6-10 local units worked on average a further 5 km away, whereas those with 100 plus local units worked over 15km further away. This is counter intuitive to what we would be expected – i.e., individuals limiting their commuting distance by choosing to work at the closest local unit – and is potentially further evidence of systematic measurement error. However, a direct interpretation could be applied in as much as the reality could be that this preference to limit the commute could be offset by a couple of factors. First, the number of jobs in a particular local unit may be inversely related to the number of outlets. Second, working for an organisation with numerous sites may provide employees with increased internal labour mobility opportunities. Although helpful for career progression, this may come at a cost of requiring the employee to switch locations without the cost of moving house.

There is a strong regional effect with employees working for enterprises based outside London all travelling lessor distances to work than those who worked for organisations based in the capital (ranged from one to seven kilometres). Given the cost of housing and the highly sophisticated transport infrastructure and labour markets of the capital, it is unsurprising that individuals are prepared to travel further distances to work in London.

The results in relation to registered status is inconclusive with sole traders, partnerships and local authorities all working closer than registered companies do. Central government employees worked further away, and all remaining structures were insignificant. As such, there are no obvious implications for researchers undertaking workplace analysis.

6. Discussion

The detailed analysis of a single spatial variable, goes well beyond the core work of NSIs. However, given the rising importance of microdata analysis in informing policy, it highlights the important contribution the research community can make in helping NSIs better understand their data and the implications it has for its use. This is particularly true in a time when NSIs budgets are under pressure and the demand for data products is growing.

6.1 Analytical implications

Our results indicate a strong likelihood of systematic error in the ASHE workplace location for multi-site companies providing responses to pre-filled surveys. This in turn highlights two important questions:

1. Can the results of this type of analysis be used to address the issues identified in this dataset?
2. What are the implications for other similar microdatasets?

As it is not possible to observe the 'true' value of the work location, it is challenging to construct an accurate probability function to directly address this challenge[1]. Moreover, the precise impact of the error depends on the functional form of interest. However, there are some options open to the researcher:

- Inclusion of the 'special arrangements' variable as a specification test. The analyses above suggests that, even in the presence of other controlling factors, this variable has an impact on travel-to-work and head office concentration measures. This is likely to be particularly valuable if the number of local units is an explanatory variable.
- Run regressions to generate predictions, either for direct use or as instruments. The inclusion of 'special arrangements' in the selection equation, of no value to the equation of interest, may be a sufficient identifying condition.
- Create sub-samples based on the 'best' data on which to undertake analysis, perhaps by excluding outlier values. This is not easy to do, as a clerical analysis by the authors of these

[1] We are exploring with ONS whether it records whether the prefilled or the update box was filled by the respondent. ONS does not distinguish between the two in the files released to researchers, but it may be that some unreleased QA information is held by the survey team.

outlier values (e.g., home address over 300 miles from work address) showed that these are plausible.

For the last option, an extreme approach would include only observations from single site organisations or multi-site organisations with special arrangements. This would exclude over half the number of observations (97,000) from approximately one third of all enterprises (18,314) and as such be subject to losing considerable valuable information, as well as potentially biasing other results by creating an unrepresentative sub-sample. Nevertheless, it is worth considering this as it helps to illustrate the impact this measurement error can have. Table 6 shows the percentage reduction in the mean gross earnings of the high certainty sample (i.e., all observations excluding multi-site companies making survey submissions) with the full sample.

<INSERT TABLE 6 HERE>

Table 6 shows the percentage decline in mean wages having removed the 'problematic' multi-site survey responses. All regions record a fall in the average mean wage rate; this is because the removed firms are generally large and productive. However, given that the reduction in the ratio of mean gross wage is least pronounced for London (3 percentage points) compared with all other regions (e.g. 16 percentage points for the North East), this supports the assertion that there may be systematic measurement error, leading to an underestimation of regional pay gaps. This figure of course captures the composition effect of firms within region. For example, as London has a higher proportion of very large firms qualifying for special arrangements; these firms remain in the high certainty sample and therefore the fall in mean annual gross wages should be limited somewhat compared to other regions. Nevertheless, the table illustrates that the gross impact of the measurement error can be large.

The results from the regression showed that multi-site exclusions can be more subtle when a combination of other factors are also be taken into account; the three most important are sector (in particular the construction, finance/law and utilities sector), the number of local units, and potentially regional location (e.g. London). By incorporating these factors, it would be possible to improve the quality of the sample but limit exclusions and loss of data (e.g. only exclude multi-site companies without special arrangements, in three potentially problematic sectors, with over 100 local units and headquartered in London). Given the potential to limit the systematic measurement error, by creating

alternative sub-samples, researchers will be better able to test the robustness of their results, which ultimately should lead to better evidence to inform policy.

6.2 Lessons for NSIs

Our case study illustrates the tension between reducing the respondent burden and maintaining quality. The inference being that NSIs should carefully consider all the costs and benefits of pre-filling workplace location information in business surveys, before deciding to do so.

At a national level, this case study suggests ONS should undertake a full cost and benefit analysis to assess making both short-term changes to the survey form and long-term changes to the data collection model.

In the short-term ONS should assess the cost and benefits of ceasing to include pre-filled postcodes in their ASHE survey forms. Although this study suggests that by doing so, this should reduce measurement error and improve the quality of the workplace location data, it also creates the potential for a different type of measurement error. For example, some studies [11, 13] discuss the fact that with pre-filled data, the respondent does not need to recall the information and has still the opportunity to review it. Therefore, as one of the options in their cost and benefit analysis, ONS may wish to consider a hybrid approach. For example, ONS could request that the employer fills out the workplace location question the first time that they respond to the survey for each employee, while in any subsequent years, the survey could then be sent out pre-filled using the employer supplied data.

In the longer-term ONS should consider increasing the coverage of organisations using 'special arrangements' to submit their data. This appears to be particularly important for multi-site and private sector companies. The additional burden of asking and processing fuller returns from companies can be partially offset by the potential it offers to improve data quality. We are aware that moving to a blanket change overnight may be challenging, given the current systems in place, so ONS may wish to use a staged approached.

ONS could potentially focus on the following groupings in a sequential order.

- Particular sectors most severally affected – i.e., construction, finance/law and utilities sector
- Enterprise with over a specified number of local units
- Businesses head quartered in London

- All private sector enterprise
- All enterprises

More broadly, the case study has illustrated the value of allowing researchers with a different perspective from that of official statistics to undertake and feed into the quality assurance process of its datasets.

7. Conclusion

The role of the NSI is changing, with many countries now making their microdata available to researchers via secure research environments. As the use of data collected by the NSIs changes, so perhaps it is time for NSIs to review historic decisions made about data collection and processing, particularly in relation to limiting the response burden. When assessing the case in relation to response burden, the NSIs should factor in the benefits that more detailed microdata capture can offer to improving the evidence base.

It is clear that the provision and access to microdata as a public good has led to significant benefits. This has, however, brought to the fore issues with the data which previously were less well understood. In a time of increasing pressure on NSI budgets, not only does access to microdata offer the potential for researchers to explore thematic issues, but it can also provide the NSI with a 'free' pool of labour that can help to quality assure the data itself.

In this article we use the UK's ASHE dataset as a case study to demonstrate how researchers' different data quality needs can identify and illustrate concerns with the core official statistics of the NSI. We demonstrated how novel analysis could identify potential measurement error in a dataset, when the "true" value of a variable is unknowable. However, there is also significant value in the very different ways that researchers approach the problem. NSIs, when made aware of an issue, will consider "how big is it, and does it affect my results?" Researchers on the other hand are more likely to consider the same problem in the terms "what is causing it, and can I account for it?" These two approaches are both valid, and jointly can shed considerable statistical light on the problem of interest.

The analysis demonstrated that by combining datasets and segmenting the main dataset of interest, it was possible to identify mismeasurement previously assumed to be 'unknowable'. This also enabled the identification of different approaches to improve the dataset; in the short-term, this focussed

on the inclusion of newly identified control variables and the creation of sub-samples, while longer-term suggestions would change the data collection process itself. All of these findings may be of particular interest to microdata researchers who face similar challenges, while NSIs may wish to consider how best to utilise the microdata research community to support their work.

The specific findings here illustrate the power of letting researchers loose on the data. Indirect methods showed that large firms completing surveys forms on an employee-by-employee basis were likely to have (a) more employees based at the head office, and (b) employees travelling on average some 12km further to work than other workers. Both findings support the hypothesis that the pre-filling of the work address leads to measurement error, and the regression analysis confirmed this is additional to any effect arising from the characteristics of the firms themselves.

These findings also provide useful feedback to the NSI, in helping it to identify potential areas to look for remedies. In ASHE, the problem arises because the datasets used for analysis and official statistics do not distinguish between prefilled responses and corrected responses. However, it is likely that some part of the ingest process does record this information; if this information has been retained, it would instantly allow quality measures to be put on the geographical data, on the basis that a corrected entry can be assumed to represent the true value. ONS and the research team are now working to explore this possibility.

8. Acknowledgements

This work was funded by the Economic and Social Research Council [grant number: ES/T013877/1].

This work was produced using statistical data from ONS with the analysis being carried out in the Secure Research Service (SRS), part of the ONS. Data supporting this study are available for accredited researchers to access via the SRS. Access to the data is subject to project approval following accreditation in the Research Accreditation Service (RAS).

analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

The opinions that are expressed in this paper are the views of the authors alone.

9. References

[1] Giesen D, Vella M, Brady CF, Brown P, Ravindra D, Vaasen-Otten A. Response burden management for establishment surveys at four national statistical institutes. Journal of Official Statistics. 2018 Jun 1;34(2):397-418.

[2] De Vries W. Are we measuring up...? Questions on the performance of national statistical systems. International Statistical Review/Revue Internationale de Statistique. 1999 Apr 1:63-77.

[3] Pina-Sánchez J, Koskinen J, Plewis I. Adjusting for measurement error in retrospectively reported work histories: An analysis using Swedish register data. Journal of Official Statistics. 2019 Mar 1;35(1):203-29.

[4] Ritchie F. UK release practices for official microdata. Statistical Journal of the IAOS. 2009 Jan 1;26(3, 4):103-11.

[5] Ritchie F. Improving data quality: the user as data detective. Conference of European Statistics Stakeholders. 2016 October.

[6] Abraham KG. Big Data and Official Statistics. Review of Income and Wealth. 2022.

[7] Bavdaž M, Giesen D, Černe SK, Löfgren T, Raymond-Blaess V. Response burden in official business surveys: Measurement and reduction practices of national statistical institutes. Journal of Official Statistics. 2015 Dec 1;31(4):559-88.

[8] Hoogendoorn, A.W. and Sikke, D., 1998. Response burden and panel attrition. *Journal of Official Statistics*, *14*(2), p.189.

[9] Hansen MH, Hurwitz WN, Pritzker L. The accuracy of census results. American Sociological Review. 1953 Aug 1;18(4):416-23.

[10] Rodhouse JB, Ott K. Respondent Perceptions of Previously Reported Data. Survey Practice. 2022 Jun 30;15(1). Rodhouse JB, Ott K. Respondent Perceptions of Previously Reported Data. Survey Practice. 2022 Jun 30;15(1).

[11] Holmberg A. Pre-printing effects in official statistics: an experimental study. Journal of Official Statistics. 2004 Jun 1;20(2):341.

[12] Jackle A. Dependent interviewing: effects on respondent burden and efficiency of data collection. Journal of Official Statistics. 2008 Jan 1;24(3):411.

[13] Lynn P, Jäckle A, Jenkins SP, Sala E. The impact of interviewing method on measurement error in panel survey measures of benefit receipt: evidence from a validation study. ISER Working Paper Series; 2004.

[14] McGuckin RH, Nguyen SV. Public use microdata: Disclosure and usefulness. Journal of Economic and Social Measurement. 1990 Jan 1;16(1):19-39.

[15] Ritchie F, Whittard D, Dawson C. Understanding official data sources. London: Low Pay Commission. 2014 Feb.

[16] Office for National Statistics. Employee earnings in the UK: 2020. 2021

[17] Low Pay Commission. Summary of Findings Report. 2022 November

[18] Office for National Statistics. Gender Differences in Commute and Pay. 2019

[19] Britton J, Waltmann B, Xu X. London calling? Higher education, geographical mobility and early-career earnings: Research report: September 2021.

[20] Sanchis-Guarner, R. and Lyytikäinen, T. Driving up wages: The effects of road improvements in Great Britain. 2012. In conference paper, www.ieb.ub.edu/aplicacio/fitxers/WS12Sanchis-Guarner.pdf.

[21] Gibbons S, Overman H, Pelkonen P. The decomposition of variance into individual and group components with an application to area disparities. InTechnical Report 2012. mimeo. London, LSE.

[22] Schork J, Riillo CA, Neumayr J. Survey mode effects on objective and subjective questions: Evidence from the labour force survey. Journal of Official Statistics. 2021 Mar 1;37(1):213-37.

[23] Felderer B, Kirchner A, Kreuter F. The effect of survey mode on data quality: Disentangling nonresponse and measurement error bias. Journal of Official Statistics. 2019 Mar 1;35(1):93-115.

[24] Galesic M. Dropouts on the web: Effects of interest and burden experienced during an online survey. Journal of official statistics. 2006 Jun 1;22(2):313.

[25] Haas GC, Eckman S, Bach R. Comparing the Response Burden between Paper and Web Modes in Establishment Surveys. Journal of Official Statistics. 2021 Dec 1;37(4):907-30.

[26] Chatterjee K, Chng S, Clark B, Davis A, De Vos J, Ettema D, Handy S, Martin A, Reardon L. Commuting and wellbeing: a critical overview of the literature with implications for policy and future research. Transport reviews. 2020 Jan 2;40(1):5-34.

[27] Kahneman D, Krueger AB, Schkade DA, Schwarz N, Stone AA. A survey method for characterizing daily life experience: The day reconstruction method. Science. 2004 Dec 3;306(5702):1776-80.

[28] Koster PR, Koster HR. Commuters' preferences for fast and reliable travel: A semi-parametric estimation approach. Transportation Research Part B: Methodological. 2015 Nov 1;81:289-301.

[29] Office for National Statistics. Annual Survey of Hours and Earnings, Low Pay and Annual Survey of Hours and Earnings Pension Results QMI. 2021.

[30] Office for National Statistics. 2020. Annual Survey of Hours and Earnings, 1997-2020: 2020 Secure Access. [data collection]. 17th Edition. UK Data Service. SN: 6689,

[31] Office for National Statistics. Business Structure Database, 1997-2018: 2019 Secure Access. [data collection]. 10th Edition. UK Data Service. SN: 6697, http://doi.org/10.5255/UKDA-SN-6697-10

[32] Whittard, D., Ritchie, F., Phan, V., Forth, J., Bryson, A., Stokes, L., Singleton., C and McKenzie, A. Exploring the workplace location problem in the Annual Survey of Hours and Earnings. 2022. UWE

[33] Ritchie F. Microdata access and privacy: What have we learned over twenty years?. Journal of Privacy and Confidentiality. 2021 Feb 3;11(1).

[34] Halvorsen R, Palmquist R. The interpretation of dummy variables in semilogarithmic equations. American economic review. 1980;70(3):474-75.

[35] Limtanakool N, Dijst M, Schwanen T. On the participation in medium-and long-distance travel: A decomposition analysis for the UK and the Netherlands. Tijdschrift voor economische en sociale geografie. 2006 Sep;97(4):389-404.

Table 1: Number and percentage of individuals working in enterprises, by characteristic of employer.

	Organisation type			
	(A) Single site, survey	(B) Multi-site, survey	(C) Single site, electronic	(D) Multi-site, electronic
Business Size				
SME (<250 employees)				
Number	49,503	17,159	147	67
Proportion	74.0%	25.7%	0.2%	0.1%
Large business				
Number	7,144	79,474	86	25,133
Proportion	6.4%	71.1%	0.1%	22.5%
Sector				
Public				
Number	3,135	26,070	173	9,616
Proportion	8.04%	66.86%	0.44%	24.66%
Private				
Number	48,476	61,434	51	15,047
Proportion	38.8%	49.1%	0.0%	12.0%
Location				
UK (excluding London)				
Number	48,005	82,761	213	21,947
Proportion	31.39%	54.12%	0.14%	14.35%
London				
Number	8,642	13,872	20	3,253
Proportion	33.5%	53.8%	0.1%	12.6%
Location				
Rural				
Number	11,244	11,295	17	2,179
Proportion	45%	46%	0%	9%
Urban				
Number	41,608	76,891	184	18,941
Proportion	30.2%	55.9%	0.1%	13.8%

Source: Authors' calculations based on ONS' ASHE

Table 2: Numbers of enterprises and proportions of their employees based in head offices, split by size of enterprise, ASHE vs BSD (2018)

size (n. of employees)		BSD (1)	ASHE (2)	BSD, restricted sample (3)	ASHE, restricted sample (4)
micro 0-9	Number of firms.	13,923	13,977	10,919	10,959
	Proportion of employees based at head office	99%	78%	100%	100%
Small 10-49	Number of firms	20,402	20,402	16,286	16,286
	Proportion of employees based at head office	94%	79%	96%	99%
Medium 50-249	Number firms	14,792	14,792	11,390	11,390
	Proportion of employees based at head office	81%	73%	87%	95%
Large 250 - 999	Number of firms.	4,976	4,976	3,703	3,703
	Proportion of employees based at head office	60%	57%	67%	76%
Xlarge 1000-10,000	Number of firms	1,883	1,883	1,497	1,497
	Proportion of employees based at head office	35%	37%	39%	46%
Xxlarge 10,000+	Number of firms	192	192	155	155
	Proportion of employees based at head office	15%	21%	17%	26%
Unmatched	Number of firms	-	976	-	-

Source: Authors' calculations based on ONS' BSD and ASHE

Table 3: Distance (in kilometres) between recorded employee work and home addresses in the ASHE (2018)

Organisation type	Observations (1)	Mean (2)	10th percentile (3)	50th percentile (4)	90th percentile (5)
Single site - survey	56,647	16.6	0.5	5.7	29.5
Multi-site - survey	96,633	28.2	1.1	7.7	54.4
Single site, electronic	233	8.5	0.5	4.9	19.5
Multi-site, electronic	25,200	16.4	0.9	5.5	30.9

Source: Authors' calculations based on ONS' ASHE

Table 4: Regressions on the proportion at head office, all multi-site enterprises and multi-site enterprises ignoring 'shell' companies (2018)

Dependent variable: proportion at head office	OLS (1)	OLS - restricted (2)	Tobit - restricted (3)
Special arrangements	-0.14***	-0.11***	-0.14***
Omitted Category: Public Sector	-	-	-
Primary	0.10***	0.07***	0.22***
Manufacturing	0.09***	0.04***	0.09***
Utilities	0.01	0.05*	0.12*
Construction	0.08***	0.10***	0.26***
Sales	0.01	0.05***	0.15***
Services	-0.05***	0.03**	0.08**
Financial/law	0.02	0.06***	0.14***
Health	-0.07***	0.07***	0.18***
Creative	-0.01	0.02	0.07*
Other	0.02	0.09***	0.26***
Omitted Category: Urban City and Town	-	-	-
Rural hamlets and Isolated Dwellings in a sparse setting	0.03	-0.01	-0.06
Rural hamlets and isolated dwellings	0.04**	0.01	0.02
Rural town and Fringe in a sparse setting	-0.02	0.02	0.05
Rural town and Fringe	-0.00	0.01	0.03
Rural village in a sparse setting	-0.04	0.01	0.02
Rural village	-0.02	-0.00	-0.02
Urban city and town in a sparse Setting	-0.12*	-0.02	-0.09
Urban minor conurbation	-0.05**	-0.04***	-0.10***
Urban major conurbation	-0.01	-0.02**	-0.05**
Number of employees	-0.00	-0.00***	-0.00***
Omitted Category: 2-5 units	-	-	-
1	-	-	-
6-10	-0.19***	-0.17***	-0.42***
11-50	-0.30***	-0.34***	-0.69***

51-99	-0.35***	-0.50***	-0.89***
100+	-0.38***	-0.54***	-0.92***
Omitted Category: London	-	-	-
North East	0.02	0.03**	0.02
North West	0.01	0.03***	0.04*
Yorkshire & Humberside	0.02	0.04***	0.07**
East Midlands	0.03*	0.05***	0.10***
West Midlands	0.03**	0.06***	0.12***
South West	0.03**	0.05***	0.11***
East	0.03**	0.05***	0.09***
South East	0.03**	0.06***	0.12***
Wales	0.01	0.10***	0.28***
Scotland	-	-	-
Omitted Category: Company	-	-	-
Sole proprietor	-0.01	0.09***	0.55***
Partnership	0.16***	0.05***	0.29***
Public Corporation	0.11**	-0.01	-0.08
Central Government Body	0.05**	-0.06***	-0.16***
Local Authority	0.18***	0.08***	0.12***
Non-profit Making Body	0.13***	0.04***	0.05*
Constant	0.53***	0.82***	1.16***
Observations	17,183	10,364	10,364
R-squared	0.12	0.38	
Probability>F=	0	0	
Pseudo R2			0.2401
Probability			0

Source: Authors' calculations based on ONS' BSD and ASHE

Table 5: Distance travelled to work, all employees and adjusted for enterprise clustering (2018)

	Distance - all observations		Distance- multi-site only	
Dependent variable: distance travelled to work	Kilometres (1)	Logs (2)	Kilometres (3)	Logs (4)
Special arrangements	-9.59***	-0.16**	-9.50***	-0.16**
Single site	-6.02***	-0.27***		
Omitted Category: Public Sector	-	-	-	-
Primary	4.84	0.18**	19.38*	0.55***
manufacturing	0.28	0.23***	4.23	0.33***
utilities	3.76	0.47***	5.72	0.59***
Construction	9.79***	0.67***	20.98***	1.01***
Sales	-3.67*	-0.01	-4.88	-0.02
Services	-3.71**	0.02	-5.00*	0.06
Financial/law	12.55***	0.58***	15.84***	0.67***
Health	-3.77**	-0.07*	-3.49*	-0.05
Creative	-6.13***	-0.11**	-9.13***	-0.20***
Other	1.31	0.11**	5.87	0.33***
Omitted Category: Urban City and Town	-	-	-	-
Rural hamlets and Isolated Dwellings in a sparse setting	3.10	0.51***	-2.23	0.42***
Rural hamlets and isolated dwellings	3.50**	0.51***	2.21	0.47***
Rural town and Fringe in a sparse setting	-3.59**	-0.15**	-4.16*	-0.06
Rural town and Fringe	1.86	0.10**	2.30	0.10
Rural village in a sparse setting	3.59	0.35**	4.07	0.40**
Rural village	2.34	0.45***	1.63	0.43***
Urban city and town in a sparse Setting	-0.16	-0.10	-0.87	-0.03
Urban minor conurbation	1.37	0.09	2.85	0.09
Urban major conurbation	2.84**	0.13***	3.89**	0.15***
Number of employees	-0.01***	-0.00***	-0.01***	-0.00***
Number of local units: Omitted Category: 2-5	-	-	-	-
1	-	-	-	-
6-10	5.40**	0.12***	5.32**	0.11***
11-50	10.50***	0.23***	10.48***	0.23***

51-99	15.13***	0.35***	15.62***	0.35***
100+	15.30***	0.17***	16.48***	0.19***
Omitted Category: London	-	-	-	-
North East	-4.20	-0.32***	-5.06	-0.32***
North West	-6.18***	-0.34***	-7.57***	-0.34***
Yorkshire & Humberside	-7.60***	-0.38***	-9.33***	-0.37***
East Midlands	-6.06***	-0.30***	-6.49**	-0.26***
West Midlands	-5.23***	-0.26***	-5.05*	-0.22***
South West	-4.04	-0.36***	-3.17	-0.32***
East	-1.40	-0.19***	-0.56	-0.18***
South East	0.74	-0.16***	1.55	-0.15**
Wales	-6.70**	-0.28***	-7.03*	-0.24***
Scotland	-	-	-	-
Omitted Category: Company	-	-	-	-
Sole proprietor	-10.35***	-0.60***	-15.88***	-0.73***
Partnership	-11.06***	-0.50***	-16.26***	-0.48***
Public Corporation	3.58	0.20	4.24	0.23
Central Government Body	-7.57***	0.08*	-7.77***	0.14**
Local Authority	-20.96***	-0.30***	-21.10***	-0.23**
Non-profit Making Body	-4.86***	-0.01	-5.36**	0.07
Constant	25.51***	2.03***	24.12***	1.93***
Observations	151,448	151,448	102,869	102,869
R-squared	0.04	0.08	0.04	0.08
p	0	0	0	0

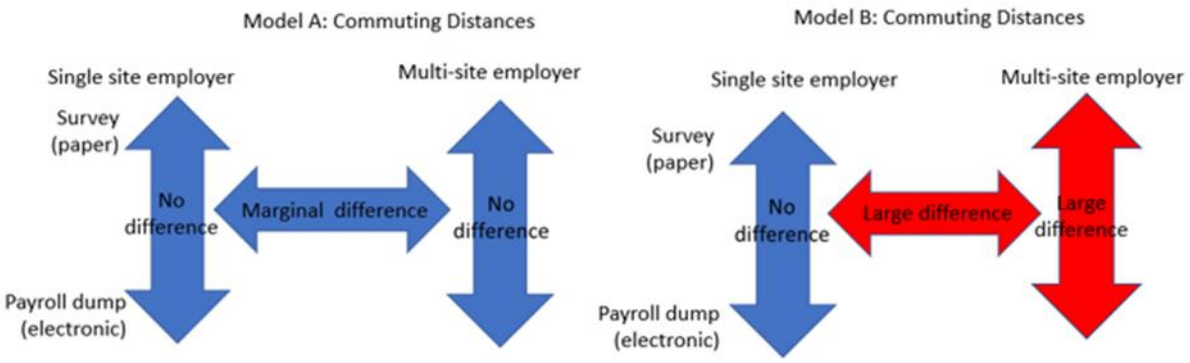
Source: Authors' calculations based on ONS' BSD and ASHE

Table 6: Comparison of the effect on estimated mean annual gross wage by excluding multi-site companies making paper submissions.

Government Office Regions: In rank order	Percentage reduction in mean gross earnings having removed multi-site companies making survey submissions
London	3%
East Midlands	7%
East	7%
Scotland	7%
Yorkshire & Humberside	10%
South West	10%
South East	10%
West Midlands	11%
North West	12%
Wales	15%
North East	16%

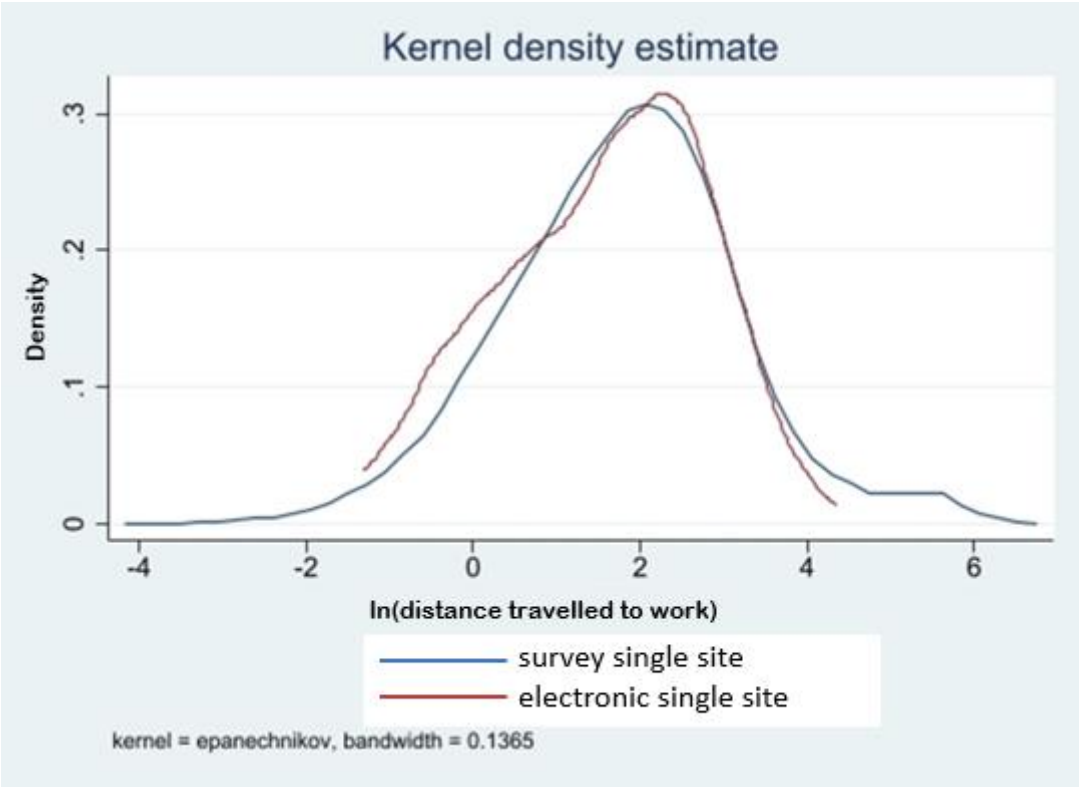
Source: Authors' calculations based on ONS' BSD and ASHE

Figure 1: Testable differences in firm data



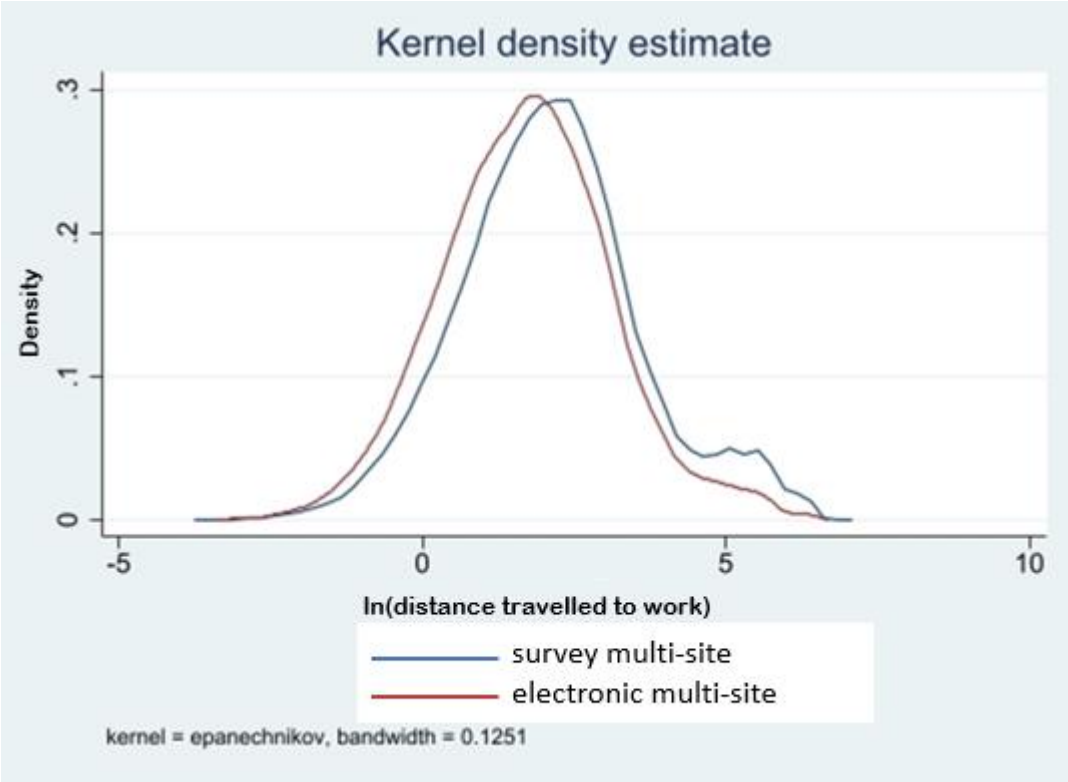
Source: Authors' model

Figure 2: Kernel density plot of natural logarithm distance travelled to work of employees of single site enterprises, by survey and electronic submission (2018)



Source: ONS' BSD and ASHE

Figure 3: Kernel density plot of natural logarithm distance travelled to work of multi-site enterprises, by survey and electronic submission (2018)



Source: ONS' BSD and ASHE

9. Supplementary Material

Regressions for 2016

Table 1a: Regressions on the proportion at head office, all multi-site enterprises and multi-site enterprises ignoring 'shell' companies (2016)

	(1)	(2)	(3)
	OLS	OLS - restricted	Tobit - restricted
Special arrangements	-0.07**	-0.09***	-0.10**
Single site	-	-	-
Unmatched	-	-	-
Primary	0.06	0.05*	0.15*
manufacturing	0.09***	0.02	0.04
utilities	-0.02	0.04	0.12*
Construction	0.04*	0.04***	0.11***
Sales	-0.00	0.04***	0.12***
Services	-0.05**	0.02*	0.08**
Financial/law	0.01	0.05***	0.14***
Health	-0.08***	0.05***	0.16***
Creative	0.01	0.04**	0.14***
Other	0.02	0.07***	0.25***
Rural hamlets and Isolated Dwellings in a sparse setting	-0.03	-0.05	-0.19
Rural hamlets and isolated dwellings	0.03*	0.01	0.04
Rural town and Fringe in a sparse setting	-0.13***	-0.04	-0.12
Rural town and Fringe	-0.00	0.01	0.03
Rural village in a sparse setting	-0.07	0.02	0.06
Rural village	0.00	-0.01	-0.03
Urban city and town in a sparse Setting	0.03	0.03	0.12
Urban minor conurbation	-0.06***	-0.02	-0.05
Urban major conurbation	-0.02*	-0.01*	-0.04*
Number of employees	-0.00	-0.00***	-0.00***

1	-	-	-
6-10	-0.21***	-0.17***	-0.46***
11-50	-0.30***	-0.34***	-0.72***
51-99	-0.36***	-0.52***	-0.94***
100+	-0.39***	-0.59***	-1.01***
North East	-0.00	0.03**	0.04
North West	0.01	0.03***	0.04*
Yorkshire & Humberside	0.01	0.04***	0.07**
East Midlands	0.04**	0.06***	0.13***
West Midlands	0.01	0.06***	0.12***
South West	0.01	0.05***	0.10***
East	0.02	0.06***	0.12***
South East	0.03**	0.07***	0.16***
Wales	-0.00	0.08***	0.22***
Scotland	-	-	-
Sole proprietor	-0.05	0.05*	0.27***
Partnership	0.12***	0.04***	0.25***
Public Corporation	0.09*	-0.02	-0.10
Central Government Body	0.10***	-0.02*	-0.12***
Local Authority	0.17***	0.08***	0.13***
Non-profit Making Body	0.14***	0.04***	0.04*
Constant	0.55***	0.84***	1.21***
Observations	16,468	9,781	9,781
R-squared	0.12	0.41	
p	0	0	0
r2_p			0.261

Table 1b Distance travelled to work, all employees and adjusted for enterprise clustering (2016)

	(1)	(2)	(3)	(4)
	Kilometres	Logs	Kilometres	Logs
Special arrangements	-9.68***	-0.13*	-10.30***	-0.13*
Single site	-6.58***	-0.27***		
Unmatched	-	-	-	-
Primary	5.80	0.23**	12.33	0.45**
manufacturing	-1.76	0.20***	-0.52	0.23***
utilities	2.37	0.45***	2.88	0.50***
Construction	10.55***	0.66***	20.51***	0.93***
Sales	-5.21*	-0.07	-7.13*	-0.10
Services	-1.94	0.06	-2.50	0.12
Financial/law	11.70***	0.54***	13.92***	0.60***
Health	-8.19***	-0.19***	-10.09***	-0.24***
Creative	-8.23***	-0.24***	-11.06***	-0.35***
Other	0.58	0.05	3.68	0.25**
Rural hamlets and Isolated Dwellings in a sparse setting	1.39	0.45***	-2.63	0.30**
Rural hamlets and isolated dwellings	0.00	0.45***	-1.47	0.40***
Rural town and Fringe in a sparse setting	0.60	-0.11	2.10	0.01
Rural town and Fringe	0.39	0.06	-0.02	0.04
Rural village in a sparse setting	7.81**	0.36***	5.50	0.38**
Rural village	1.13	0.39***	-0.22	0.35***
Urban city and town in a sparse Setting	4.98**	0.03	7.31**	0.19
Urban minor conurbation	0.85	0.09	1.16	0.08
Urban major conurbation	2.54*	0.11***	3.90*	0.13***
Number of employees	-0.01***	-0.00***	-0.01***	-0.00***
1	-	-	-	-
6-10	1.43	0.07**	1.32	0.07**
11-50	9.66***	0.25***	9.48***	0.25***
51-99	11.74***	0.33***	12.07***	0.33***
100+	16.46***	0.20***	17.17***	0.21***

North East	-1.97	-0.30***	-1.53	-0.28***
North West	-6.57***	-0.35***	-8.26***	-0.36***
Yorkshire & Humberside	-8.36***	-0.40***	-10.04***	-0.41***
East Midlands	-3.15	-0.25***	-2.32	-0.20**
West Midlands	-3.78	-0.25***	-3.54	-0.22***
South West	-0.96	-0.30***	-0.37	-0.26***
East	0.08	-0.16***	0.86	-0.15**
South East	-0.10	-0.18***	0.19	-0.18***
Wales	-8.50***	-0.29***	-9.79***	-0.28***
Scotland	-	-	-	-
Sole proprietor	-8.36***	-0.56***	-11.98***	-0.60***
Partnership	-8.97***	-0.48***	-13.70***	-0.44***
Public Corporation	5.91	0.31	5.36	0.33
Central Government Body	-3.88**	0.17***	-2.58	0.27***
Local Authority	-22.10***	-0.40***	-22.60***	-0.37***
Non-profit Making Body	-4.92***	-0.03	-5.57**	0.04
Constant	25.57***	2.04***	25.15***	1.98***
Observations	149,146	149,146	105,137	105,137
R-squared	0.05	0.08	0.05	0.09
p	0	0	0	0

Regressions for 2017

Table 2a: Regressions on the proportion at head office, all multi-site enterprises and multi-site enterprises ignoring 'shell' companies (2017)

	(1)	(2)	(3)
	OLS	OLS - restricted	Tobit - restricted
Special arrangements	-0.13***	-0.10***	-0.11**
Single site	-	-	-
Unmatched	-	-	-
Primary	0.04	0.04	0.13*
manufacturing	0.08***	0.03**	0.07**
utilities	-0.06	0.01	0.06
Construction	0.07***	0.06***	0.17***
Sales	0.00	0.04***	0.11***
Services	-0.05***	0.03**	0.09**
Financial/law	0.01	0.04***	0.10***
Health	-0.08***	0.06***	0.18***
Creative	0.01	0.03*	0.12**
Other	0.02	0.06***	0.19***
Rural hamlets and Isolated Dwellings in a sparse setting	0.03	-0.05	-0.17
Rural hamlets and isolated dwellings	0.03*	0.00	0.00
Rural town and Fringe in a sparse setting	-0.03	0.01	0.02
Rural town and Fringe	0.02	0.00	0.01
Rural village in a sparse setting	-0.08	0.01	-0.03
Rural village	-0.01	-0.00	-0.00
Urban city and town in a sparse Setting	-0.06	0.07	0.38*
Urban minor conurbation	-0.03	-0.03**	-0.09**
Urban major conurbation	-0.02*	-0.02***	-0.06***
Number of employees	-0.00	-0.00***	-0.00***
1	-	-	-
6-10	-0.21***	-0.16***	-0.43***
11-50	-0.31***	-0.35***	-0.73***
51-99	-0.37***	-0.52***	-0.94***

100+	-0.39***	-0.56***	-0.97***
North East	-0.01	0.01	-0.02
North West	0.01	0.02*	0.03
Yorkshire & Humberside	0.00	0.02	0.03
East Midlands	0.02	0.05***	0.10***
West Midlands	0.02	0.05***	0.10***
South West	0.02	0.03***	0.07**
East	0.02	0.05***	0.11***
South East	0.03*	0.05***	0.11***
Wales	0.01	0.07***	0.21***
Scotland	-	-	-
Sole proprietor	0.00	0.07**	0.40***
Partnership	0.15***	0.05***	0.31***
Public Corporation	0.07	-0.04	-0.17**
Central Government Body	0.06***	-0.04***	-0.14***
Local Authority	0.17***	0.08***	0.13***
Non-profit Making Body	0.14***	0.05***	0.07***
Constant	0.55***	0.85***	1.23***
Observations	16,253	9,763	9,763
R-squared	0.13	0.41	
p	0	0	0
r2_p			0.257

Table 2b: Distance travelled to work, all employees and adjusted for enterprise clustering (2017)

	(1)	(2)	(3)	(4)
	Kilometres	Logs	Kilometres	Logs
Special arrangements	-12.77***	-0.23***	-13.05***	-0.23***
Single site	-5.12***	-0.24***		
Unmatched	-	-	-	-
Primary	0.84	0.15*	5.45	0.39**
manufacturing	0.48	0.24***	4.91	0.35***
utilities	3.40	0.49***	6.14	0.60***
Construction	10.35***	0.67***	21.50***	1.00***
Sales	-4.68*	-0.03	-5.61	-0.05
Services	-2.50	0.07	-2.61	0.13
Financial/law	11.61***	0.58***	15.35***	0.66***
Health	-3.76**	-0.09*	-3.82*	-0.09
Creative	-7.35***	-0.16**	-11.22***	-0.28***
Other	-0.54	0.04	3.42	0.26**
Rural hamlets and Isolated Dwellings in a sparse setting	5.59**	0.57***	0.62	0.39***
Rural hamlets and isolated dwellings	2.09	0.50***	1.77	0.47***
Rural town and Fringe in a sparse setting	-1.45	-0.14*	-2.13	-0.07
Rural town and Fringe	1.50	0.10**	1.54	0.09
Rural village in a sparse setting	7.71**	0.46***	8.53*	0.52***
Rural village	1.70	0.43***	-0.74	0.38***
Urban city and town in a sparse Setting	1.17	-0.02	1.56	0.05
Urban minor conurbation	1.55	0.11**	2.68	0.10
Urban major conurbation	4.13***	0.15***	5.33***	0.17***
Number of employees	-0.01***	-0.00***	-0.01***	-0.00***
1	-	-	-	-
6-10	5.34**	0.10**	5.37**	0.10**
11-50	10.25***	0.24***	10.12***	0.24***
51-99	17.95***	0.40***	18.33***	0.41***
100+	16.78***	0.19***	17.88***	0.22***
North East	-0.60	-0.29***	-0.51	-0.28***

North West	-6.38***	-0.35***	-7.83***	-0.35***
Yorkshire & Humberside	-7.88***	-0.39***	-10.39***	-0.39***
East Midlands	-4.61*	-0.27***	-5.22	-0.24***
West Midlands	-4.57*	-0.25***	-5.03	-0.23***
South West	-4.25	-0.38***	-5.19	-0.37***
East	0.58	-0.15***	0.87	-0.14**
South East	2.09	-0.13***	2.71	-0.10*
Wales	-6.13**	-0.26***	-6.41*	-0.22***
Scotland	-	-	-	-
Sole proprietor	-10.05***	-0.61***	-11.96***	-0.66***
Partnership	-10.45***	-0.49***	-15.41***	-0.42***
Public Corporation	7.12	0.37	6.39	0.38
Central Government Body	-7.52***	0.11**	-7.08**	0.18**
Local Authority	-20.71***	-0.29***	-20.22***	-0.23**
Non-profit Making Body	-5.24***	-0.02	-5.18*	0.07
Constant	23.67***	1.98***	22.24***	1.89***
Observations	148,496	148,496	104,057	104,057
R-squared	0.05	0.08	0.05	0.08
p	0	0	0	0

Regressions for pooled years (2016-2018)

Table 3a: Regressions on the proportion at head office, all multi-site enterprises and multi-site enterprises ignoring 'shell' companies (2016-2018)

	(1)	(2)	(3)
	OLS	OLS - restricted	Tobit - restricted
Special arrangements	-0.12***	-0.11***	-0.13***
Single site	-	-	-
Unmatched	-	-	-
Primary	0.09**	0.08***	0.22***
manufacturing	0.09***	0.04***	0.10***
utilities	-0.02	0.02	0.06
Construction	0.07***	0.05***	0.12***
Sales	-0.00	0.04***	0.10***
Services	-0.06***	0.01	0.03
Financial/law	-0.00	0.04***	0.08***
Health	-0.08***	0.03***	0.08***
Creative	0.01	0.03	0.07**
Other	0.02	0.06***	0.16***
Rural hamlets and Isolated Dwellings in a sparse setting	0.04	-0.03	-0.10
Rural hamlets and isolated dwellings	0.03**	0.01	0.03
Rural town and Fringe in a sparse setting	-0.05	-0.03	-0.06
Rural town and Fringe	0.00	0.00	0.01
Rural village in a sparse setting	-0.08*	-0.03	-0.07
Rural village	-0.01	-0.01	-0.01
Urban city and town in a sparse Setting	-0.04	0.04	0.10
Urban minor conurbation	-0.03*	-0.03*	-0.06*
Urban major conurbation	-0.01	-0.01*	-0.03*
Number of employees	-0.00	-0.00***	-0.00***
1	-	-	-
6-10	-0.20***	-0.17***	-0.34***
11-50	-0.29***	-0.34***	-0.57***
51-99	-0.35***	-0.49***	-0.73***
100+	-0.38***	-0.53***	-0.76***
North East	-0.00	0.03**	0.05*
North West	-0.00	0.01	0.01
Yorkshire & Humberside	0.00	0.03***	0.06***
East Midlands	0.03**	0.04***	0.08***
West Midlands	0.01	0.04***	0.07***
South West	0.01	0.04***	0.07***
East	0.02	0.04***	0.06**
South East	0.02**	0.05***	0.09***
Wales	-0.00	0.08***	0.17***

Scotland	-	-	-
Sole proprietor	0.01	0.08***	0.29***
Partnership	0.15***	0.08***	0.28***
Public Corporation	0.12**	-0.00	-0.02
Central Government Body	0.05***	-0.04***	-0.10***
Local Authority	0.18***	0.12***	0.18***
Non-profit Making Body	0.13***	0.05***	0.08***
2016	0.03***	-0.15***	-0.49***
2017	0.03***	-0.07***	-0.28***
Constant	0.52***	0.91***	1.44***
Observations	22,110	14,187	14,187
R-squared	0.12	0.35	
p	0	0	0
r2_p			0.226

Table 3b: Distance travelled to work, all employees and adjusted for enterprise clustering (2016-2018)

	Kilometres	Logs	Kilometres	Logs
Special arrangements	-10.83***	-0.18***	-11.11***	-0.18***
Single site	-5.89***	-0.26***		
Unmatched	-	-	-	-
Primary	3.86	0.19**	12.79*	0.47***
manufacturing	-0.28	0.22***	3.04	0.31***
utilities	3.17	0.47***	5.00	0.57***
Construction	10.26***	0.67***	21.19***	0.99***
Sales	-4.52**	-0.04	-5.80*	-0.06
Services	-2.66	0.05	-3.16	0.11
Financial/law	12.01***	0.57***	15.21***	0.65***
Health	-5.18***	-0.11**	-5.65***	-0.12**
Creative	-7.15***	-0.16***	-10.30***	-0.27***
Other	0.49	0.07	4.46	0.28***
Rural hamlets and Isolated Dwellings in a sparse setting	3.34*	0.51***	-1.42	0.37***
Rural hamlets and isolated dwellings	1.89	0.48***	0.82	0.45***
Rural town and Fringe in a sparse setting	-1.50	-0.14**	-1.37	-0.04
Rural town and Fringe	1.25	0.09**	1.26	0.08
Rural village in a sparse setting	6.37*	0.39***	6.07	0.43***
Rural village	1.73	0.43***	0.24	0.39***
Urban city and town in a sparse Setting	1.94	-0.03	2.63	0.07
Urban minor conurbation	1.27	0.09*	2.26	0.09
Urban major conurbation	3.17***	0.13***	4.39***	0.15***
Number of employees	-0.00***	-0.00***	-0.00***	-0.00***
1	-	-	-	-
6-10	4.09**	0.09***	4.03**	0.09***
11-50	10.13***	0.24***	10.03***	0.24***
51-99	14.94***	0.36***	15.33***	0.36***
100+	16.25***	0.19***	17.27***	0.21***
East	-2.23	-0.30***	-2.31	0.29***
North West	-6.34***	-0.34***	-7.84***	-0.34***
Yorkshire & Humberside	-7.93***	-0.39***	-9.91***	-0.39***
East Midlands	-4.59**	-0.27***	-4.66	-0.23***
West Midlands	-4.49**	-0.25***	-4.48	-0.22***
South West	-3.04	-0.34***	-2.86	-0.32***

East	-0.21	-0.17***	0.44	-0.16**
South East	0.92	-0.16***	1.50	-0.14**
Wales	-7.08***	-0.27***	-7.70**	-0.24***
Scotland	-	-	-	-
Sole proprietor	-9.60***	-0.59***	-13.25***	-0.67***
Partnership	-10.15***	-0.49***	-15.07***	-0.45***
Public Corporation	5.66	0.30	5.35	0.32
Central Government Body	-6.34***	0.12***	-5.83**	0.20***
Local Authority	-21.18***	-0.33***	-21.14***	-0.27***
Non-profit Making Body	-4.99***	-0.02	-5.30**	0.06
2016	-0.97*	-0.03***	-1.06	-0.03*
2017	-0.44	-0.01	-0.59	-0.02
Constant	25.30***	2.03***	24.18***	1.94***
Observations	449,090	449,090	312,063	312,063
R-squared	0.04	0.08	0.05	0.08
p	0	0	0	0