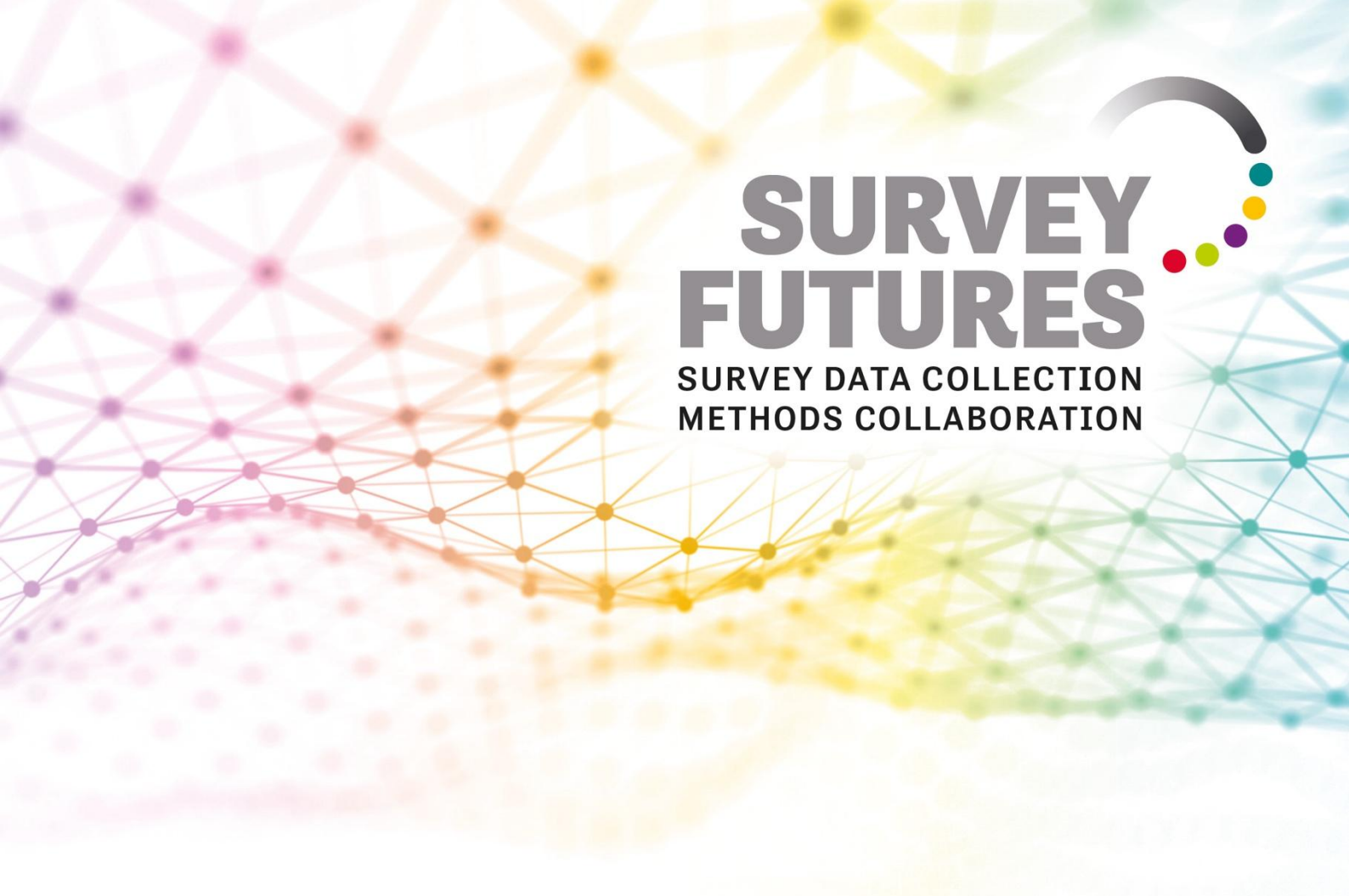




SURVEY FUTURES

**SURVEY DATA COLLECTION
METHODS COLLABORATION**

Working Paper 30:

The opportunities and challenges of sampling
frames enhanced with alternative data sources
in the UK

Paul A. Smith¹ and Anna Keyes²

¹ University of Southampton, ² National Centre for Social Research

August 2026

www.surveyfutures.net

Survey Futures is an Economic and Social Research Council (ESRC)-funded initiative (grant ES/X014150/1) aimed at bringing about a step change in survey research to ensure that high quality social survey research can continue in the UK. The initiative brings together social survey researchers, methodologists, commissioners and other stakeholders from across academia, government, private and not-for-profit sectors. Activities include an extensive programme of research, a training and capacity-building (TCB) stream, and dissemination and promotion of good practice. The research programme aims to assess the quality implications of the most important design choices relevant to future UK surveys, with a focus on inclusivity and representativeness, while the TCB stream aims to provide understanding of capacity and skills needs in the survey sector (both interviewers and research professionals), to identify promising ways to improve both, and to take steps towards making those improvements. Survey Futures is directed by Professor Peter Lynn, University of Essex, and is a collaboration of twelve organisations, benefitting from additional support from the Office for National Statistics and the ESRC National Centre for Research Methods. Further information can be found at www.surveyfutures.net.

Prior to citing this paper, please check whether a final version has been published in a journal. If so, please cite that version. In the meanwhile, the suggested form of citation for this working paper is:

Smith P A and Keyes A (2026) 'The opportunities and challenges of sampling frames enhanced with alternative data sources in the UK', *Survey Futures Working Paper* no. 30. Colchester, UK: University of Essex. Available at <https://surveyfutures.net/working-papers/>.

Contents

	Abstract	2
1	Introduction	2
2	Some history	3
2.1	Electoral register	3
2.2	Postcode address file	4
3	The modern landscape	5
3.1	Postcode address file	5
3.2	AddressBase	5
3.3	Reference data management framework	6
3.4	Administrative data sources	6
3.5	Commercial sources	7
3.6	Aggregate auxiliary information	7
4	Opportunities	7
4.1	Reducing costs	7
4.2	Names	7
4.2.1	Named addressing	8
4.2.2	Characteristics from names	8
4.3	Characteristics	8
4.4	Coverage	9
5	Challenges	9
5.1	Access	9
5.2	Ethics	9
5.3	Linkage	10
5.4	Costs	10
6	Sampling facility	11
6.1	Legal basis	11
6.2	Funding	12
6.3	Ethics of release	12
6.4	Coordination	13
6.5	Process	13
7	Conclusions	14
	References	14

Abstract

The UK survey community has used the Postcode Address File (PAF) as a frame for general population sampling for a long time. Changes in the data landscape are making more information potentially available, paradoxically at a time when tighter ethical and data protection controls are making it harder to access and use. We review the history of sampling frames in UK general population surveys, describe the opportunities and challenges presented by the use of alternative data sources as additional information for sampling, and set out how a sampling facility based within government could offer a practical option for efficiently drawing samples with improved accuracy.

Key words: sampling frame, general population survey, administrative data, sample selection

1. Introduction

General population surveys conducted in the United Kingdom (UK) all contend with the challenge of achieving a representative sample when no single national register of the population exists to use as a sampling frame. Unlike many other European nations, in the UK records of names and addresses are fragmented across multiple administrative bodies. Access to these records for sampling general population surveys remains limited, although there is increasing recognition of their potential.

High quality general population surveys in the UK have for a long time been based on the postcode address file (PAF), which provides an up-to-date list of addresses which can be used as a sampling frame, but contains no information on who lives within. The availability of this resource, and the lack of a population register, has meant that the main focus of design has been on address-based sampling. In other countries, notably the USA, random digit dialling was the preferred approach to general population sampling while the population relied largely on landline telephones. But this approach does not work well with mobile telephones, which are not linked to geography in the same way that landlines are. Developments in the US have therefore moved towards address-based sampling (Dillman *et al.* 2014, p64-5), mirroring the approach being used for rather longer in the UK.

As more information becomes accessible (if, in general, under tight conditions and restrictions on its use), it increases the possibilities for improvements in sampling processes, through the provision of additional variables and through improved predictive accuracy. Classifying variables can be used in design and selection to improve the coverage of population subgroups of interest – without the need for a large ‘blanket’ sample which can provide sufficient cases within those subgroups with high probability. There are now more options for a sampling frame in the UK than there have been for perhaps forty years, but there is also a need to develop efficient procedures for constructing and operating with a frame based on these sources.

Properties of sampling frames (adapted from Barnes (1986) and Keyes & Smith (2026)) which would need to be evaluated in making decisions to change to a different frame:

- how frequently the frame is updated, and therefore how current it is

- completeness - assessed by the estimated undercoverage rate, including for subgroups
- no duplication or overcoverage - both of which primarily affect costs, assessed by estimated overcoverage rate, including for subgroups
- geolocated
- supplemented by characteristics information as potential stratifiers.

The key property of stratifiers is that they should be related to the variable(s) of interest to be collected in the survey. The goal of the characteristic information is prediction more than absolute accuracy; in other words, the information does not have to be accurate for each record, it just has to be a good predictor of the value(s) to be collected on the survey to be useful in design. But good coverage properties are critical for efficient surveys. If there is undercoverage, then there will be bias in estimates as a result of differences in the characteristics of the excluded parts of the population.

We investigate the possibilities presented by this greater availability of options, and some of the challenges which will need to be overcome to make the approaches useful in practice. We start in section 2 with a brief overview of the development of sampling frames for general population surveys, then consider the information sources which could be useful for sampling in section 3. Section 4 sets out some of the opportunities presented by administrative and alternative data sources, and section 5 presents some of the challenges of working with these sources, and how they might be tackled specifically in the case of constructing a sampling frame. Section 6 advocates the construction of a sampling facility from within government that could be used for social surveys run by or on behalf of public bodies/charities. Section 7 provides some conclusions.

2. Some history

In the beginnings of the Social Survey the national register from the second world war was in operation and formed a natural frame for sampling in the UK. The rating (local taxation of buildings) lists were also used, and the electoral register was used in a small way, mainly for pilot surveys. These three frames are described in some detail by Gray & Corlett (1950). The electoral register became more widely used once the national register was discontinued in 1952. Its coverage at the time is reported by Moser (1955), and Emmett (1964) summarises its properties and use after it had become embedded as the main sampling frame for social surveys in the UK. The properties of the different frames are somewhat different in Northern Ireland, and these are summarised by Bond & Lievesley (1993).

2.1 Electoral register

The electoral register continued in regular use as a sampling frame for population surveys until the mid 1980s (Barnes 1986) and had some advantages – it allowed sampling of individuals within addresses for surveys that covered only one individual per address, and gave a size measure (the number of registered adults at an address) that could be used in weighting. Name-based contacts were possible in surveys, and were generally regarded as improving contact and cooperation (Lynn & Taylor 1995). The electoral register was compiled within small districts, and therefore was easily used for cluster samples which make efficient use of field interviewer resources. Census information was linked on to provide area-level stratifying variables.

The electoral register also had some deficiencies. It was always a lagged representation of the population, since it relied on the electoral registration process which always takes place some months before the register is published, and was updated only periodically (though it can now be updated continuously). In particular, movers and new builds were not well-covered. There were also problems of coverage – there was nonresponse to electoral registration processes, which became worse when the Poll Tax was introduced (McLean & Smith 1995), and there is noncoverage of addresses where all adults are ineligible to vote.

There has been substantial reform to electoral registration since the 1980s (Johnston 2026), including the introduction of opting out of the published register from 2002, and it also suffers from substantial undercoverage even in the full version, so that it no longer has sufficient coverage to be useful by itself for general population sampling.

2.2 Postcode Address File

The PAF was introduced for sampling purposes in 1983 for the Labour Force Survey (Butcher 1986). One of the initial advantages was that PAF was available digitally, so sampling could be done by computer rather than clerically.

The change from the electoral register to the PAF for general population sampling was supported by a substantial programme of evaluation (Lievesley *et al.* 1978 in Wilson & Elliot 1987). The coverage and quality of the PAF were evaluated using samples of records from the 1981 census, and an assessment of eligibility rates was also made (Wilson & Elliot 1987). The PAF rapidly became the frame of choice for social surveys, but evaluations continued into the 1990s; Foster (1994) compared the PAF with the 1991 census and with similar work on the electoral register (Foster 1993). Lynn & Taylor (1995) made an assessment of the differences between PAF and the electoral register for individual-based sampling, based on a split sample experiment. In particular they concluded that the efficiency from more equal probability selection of individuals with the electoral register was offset by other properties of the PAF, and therefore that the PAF could be used for both individual and household surveys. Lynn (1997) evaluated changes to British Crime Survey estimates from changing the sampling frame from the electoral register to the PAF.

The PAF small user file contains many out of scope addresses – small businesses, vacant, communal establishments – 10-11% overcoverage overall according to Butcher (1986) and Wilson & Elliot (1987), but varying around this by area. Keyes & Smith (2026, Table 1) give an overview of deadwood rates across a selection of recent surveys, ranging from 5.6-10.6%. In addition, some eligible addresses may appear on the large user file; this was particularly of concern for the self-employed and doctors. But in an evaluation based on 1991 census responses, only 0.05% of eligible addresses were found to be on the large user file (Foster 1994).

Most of the evaluations of the PAF were based on information from the population censuses. This provided some characteristics information which could also be used to assess whether the distributions were better than from alternative sources. Foster (1994) compared the PAF with the electoral register, and found that the PAF was better for distributions of tenure and age variables.

3. The modern landscape

We have reached a juncture where several possible sampling frames are becoming available, though sometimes involving considerable effort. The main frames for general population sampling are discussed below.

3.1 Postcode Address File

The PAF has maintained its position as the premier frame for sampling high quality general population surveys. Most of the quality evaluations of the PAF took place in the 1980s and 1990s (see section 2.2), and the lack of recent evaluations reflects a general feeling that its properties are well known, and relatively stable. It has a number of advantages, including being relatively inexpensive, regularly updated by postal workers so that it remains up to date, particularly for new builds which can otherwise be slow to reach administrative data systems. Some advantages are more hidden, but can be seen by comparison with other frame resources – there is no opt-out of the PAF, for example, as there is with NHS data.

The PAF does create some difficulties at the boundaries between large and small user postcodes. Only small user postcodes are used in general population sampling, but even here there is overcoverage of business and communal establishment addresses. Some residential addresses are potentially covered by large user postcodes, and will therefore never be included in samples using only small user postcodes. The extensive use of web surveys with a postal approach also creates some issues because of the number of dead wood addresses, which cannot be distinguished from nonrespondents. It therefore becomes more challenging to make a suitable weighting adjustment, with some assumption or modelling of the dead wood propensity being needed.

3.2 AddressBase

A comprehensive address list was produced for the 2011 censuses, based on matching together the PAF, National Land and Property Gazetteer and Ordnance Survey address point information (ONS 2010). Sensitivities at the time about ownership of the different sources meant that this was a one-off exercise specifically for the census with no subsequent use. But it was obvious that this was a useful resource that should be maintained, so in the lead up to the 2021 censuses a more permanent solution was instituted with the creation of AddressBase Premium (ABP). ABP is constructed from Local Land and Property Gazetteers (LLPGs), PAF, Valuation Office Agency data and the Ordnance Survey Master Map (UKSA 2021, Annex A). It therefore incorporates some additional classification information, which in principle makes it better at weeding out nonresidential addresses.

Some quality evaluation of the address frame for the 2021 census of England & Wales was undertaken (ONS 2023a), although the COVID lockdowns meant that this was focused on desk checks. The census address frame was constructed using some additional inputs, so the results of this exercise are only indicators for the quality of ABP, but useful indicators nonetheless. It found undercoverage of 0.36%, overcoverage of 2.1% (more than half of which were addresses under construction), and 0.42% of addresses were duplicated. In addition, 0.16% of addresses were misclassified as communal establishment/residential.

ABP is not yet in wide use for general population surveys, though it has been introduced on the Transformed Labour Force Survey (TLFS). It is more expensive than the corresponding PAF

product, so although it has definite advantages over the PAF, some formal evaluation is needed to support cost-benefit decisions on whether to move more generally to ABP.

3.3 Reference Data Management Framework

The Reference Data Management Framework (RDMF; ONS 2026) is an infrastructure developed in the ONS to support the linkage of different data sources, particularly administrative data, using a set of reference numbers specific to the statistical system. It incorporates indices for people, businesses and locations, and a classification index containing appropriate classification metadata. The location index is most closely related to existing sampling frames, and indeed incorporates both AddressBase Premium and PAF, supplemented with other sources.

However, the RDMF is not by itself a fully developed sampling infrastructure (Perera *et al.* 2026), so some further evaluation of its properties and development of the infrastructure around it as the basis for sample selection would be needed. In particular, RDMF still suffers from undercoverage

The demographic index within the RDMF covers people, and has the potential to be used for person-based sampling. There are many challenges to be resolved before this could be realised, including the currency of person information, the public acceptability, and the under- and overcoverage of individuals (ONS 2023c). These coverage issues were not fully resolved in methodological development towards an administrative data based census (ONS 2023b), and so are likely to be challenging for a sampling frame too. A summary of research on the public acceptability of the use of administrative data in statistical research is given in Perera *et al.* (2026, Annex B).

3.4 Administrative data resources

The use of health records as a general population sample frame was trialled in the 2023 Lothian Health Survey. Analysis of the survey sample estimated the percentage of ineligible records using comparison with Census and Office for National Statistics (ONS) mid-year estimate data. The results suggested that ineligibility levels in the frame varied by Local Authority and age group and could be high. Examination of the address data also showed that it was of lower quality than PAF, containing errors that made some survey invitations undeliverable.

The Early Life Cohort Feasibility Study secured access to birth records linked to maternity records, which were used as a basis for a cohort sample, but this was a challenging procedure. Linkage to maternity records provided additional variables, notably ethnicity, which were used in the sampling. A summary of the sampling methodology is given by IPSOS (2025, chapter 5), but this does not fully reflect the large effort required to apply for access to health records and work with the data providers. In England the National data opt-out (a service that allows patients to opt out of their confidential patient information being used for research and planning) operated. It is possible to make a case that the opt-out should not apply, but the conditions on which this is assessed are stringent.

Other administrative data sources could also be used as the basis of a sampling frame. Simon (2014) evaluated the potential for DWP data to be used as the frame for the Family Resources Survey, and found that it had generally good coverage properties by comparison with the PAF, although it did include some out of date postcodes which may have represented people no

longer part of the general population. However, it also included some people at large user postcode addresses who would not have been identified through the small user PAF but who could be of policy interest.

3.5 Commercial sources

Some commercial frames have been constructed using the edited Electoral Register as a base and adding further information from a variety of (commercial) sources. These have been little used for general population surveys, though Tipping & Nicolaas (2006) undertook an evaluation of the edited electoral register and the consumer register produced by CACI as a basis for a social capital survey. They found that both frames had substantial undercoverage, differential by characteristics, and estimated a small level of overcoverage (1-2% on the electoral register and 3-5% on the consumer register, depending on the assumptions made).

3.6 Aggregate auxiliary information

Access to microdata is generally hedged with many restrictions, particularly where they are being used as the basis of sampling, which has a real impact on people (and we return to this under “ethics” in sections 5 and 6). The same sensitivity is not attached to aggregate data, however. There is a long history of matching census outputs onto sampling frames to provide stratification variables (see eg Wilson & Elliot 1987). But the same principle applies equally with aggregate administrative data, and the design of the Wealth & Assets Survey (Smith *et al.* 2011) matched aggregate administrative data to the PAF as a basis for oversampling in areas with a greater proportion of households with self-assessment income tax. Some use of administrative data can be made in this way with fewer challenges over ethics and access.

4. Opportunities

4.1 Reducing costs

The main objective of using administrative data to enhance a frame would be to reduce fieldwork costs by making the samples more accurate – by reducing deadwood addresses, and by a more efficient allocation of cases to subpopulations of particular interest.

4.2 Names

One of the developments that survey organisations are most looking for is the availability of names for individuals within sampled addresses. This was one of the key losses in the transition from the electoral register to the PAF. Names have a number of potential advantages, and we discuss these in the next sections.

However, the provision of names for fieldwork purposes, even if they are available in a dataset, presents legal and ethical challenges. A name is a personal characteristic in a way that an address is not, so a legal gateway would be needed to allow this. It is not likely that names would be available in any general way, not least because of the need to comply with data protection legislation. It is likely that some research on public acceptability would be needed in support of such a process.

4.2.1 Named addressing

Assume for a moment that names could be available. Where a name is correctly associated with an address it is likely to help participation, but where a name is incorrect through population churn or administrative error it seems likely to reduce participation; the balance of these two effects is unknown – we did not find any experimental comparisons of named and unnamed approaches in the literature. Dillman *et al.* (2014, p86) recommend not addressing to a named individual if someone else may need to complete the questionnaire.

A named individual can also be used to control within-household sampling, a process that was used with the electoral register before the transition to the PAF for most social surveys. The wider use of interviewers mediated the use of names at that time – if the named person was resident, they were interviewed, otherwise a new selection from the residents was made in the field. This is clearly more challenging when there are no field interviewers (<https://surveyfutures.net/events/2026/02/04/survey-futures-workshop-within-household-selection-methods-for-surveys-without-field-interviewers-in-the-uk/>).

Tipping & Nicolaas (2006) sought a frame with names for a social capital survey, for both of these purposes, but none of the examined options was suitable.

4.2.2 Characteristics from names

It has been known for a long time that the choice of names carries information on ethnicity (Nicoll *et al.* 1986), and fashions in names also help to predict people's gender and age (Gibin *et al.* 2025). If names are available (and for this purpose they remain within the frame and do not to be released for use in the field) this therefore provides a further predicted variable which is available for sample design. The predictions do not have to be perfect – any improvement over randomness, conditional on other parts of the design, has the potential to lower costs through achieving a better targeted sample. An example of sampling using this approach, based on names from the electoral register, is given by Hughes *et al.* (1995).

Mateos (2007) gives a helpful overview of ethnic classification from names. The recent availability of large training sets and machine learning algorithms means that current predictors are potentially more accurate than those used in older studies (eg Jain *et al.* 2022), so the provision of names could have a bigger impact on sampling accuracy and/or costs than has previously been possible. Hafner *et al.* (2024) show that there are biases (including from UK census-trained classifiers) in one or more of sex, age and ethnicity, which would be problematic if they were used for automatic classification, but contribute only to variance if used as the basis of sampling, with a respondent-provided value used for further estimation.

4.3 Characteristics

The other potentially large advantage of administrative data as a basis for sampling is the availability of characteristics information at a detailed level (rather than the area-based characteristic information discussed in section 3.3). In some very specific instances this may be sufficiently accurate to replace data collection (for example the number of rooms variable was replaced in the 2021 England & Wales census with administrative data). In general, however, the primary purpose is as a predictor of the actual characteristics of the sample unit, and the better the prediction the more scope there is to reduce the variance of estimates from the sample.

For such a use the characteristics information would not be used beyond sampling, so only a list of sample addresses would be needed, and there would be limited possibility to identify any characteristics from such a list. Using characteristics in this way is therefore more consistent with an ethical approach.

4.4 Coverage

A lot of the development of the use of administrative sources has been undertaken by the ONS as part of its evaluation of methods which might be suitable to undertake an administrative data based population census (ONS 2023b). Part of the assessment involved the construction of a population spine (the demographic index) to which different administrative datasets could be linked, and then a rule- or model-based approach to whether a record should be included – wholly, partially or not at all – in a population size estimate. Rule-based approaches may work over a fixed term, but there is a risk that they become out of date in a way which is not detectable, so a model-based assessment is preferable, but requires some real (training) data from which the model can be fitted.

In a census context this can be provided by a coverage survey, and here the intersection of administrative data and sampling is raised. The coverage survey also needs to be designed, and therefore requires a frame. The actual design parameters may be different from those in most general population surveys, since the objective is to provide good training data for training/updating the models. But the general strategy of having good sampling frame information to work from in designing the coverage survey remains.

5. Challenges

5.1 Access

Historically, administrative data collected by government for various purposes have been safeguarded by legislation, so that they could not be shared. There were nevertheless examples of using administrative data for statistical purposes (for example the claimant count unemployment series), and of sharing of specific information for sampling purposes (for example VAT information as part of the business register which forms the sampling frame for official surveys of businesses). More recently the Statistics & Registration Service Act 2007 and the Digital Economy Act 2017 have enabled access to administrative and other data sources, though they do not require it. This has led to more sharing and linkage of sources within government, and also, under strict access restrictions, made these data available to researchers.

Access to administrative data for sampling is more challenging. Data owners would need to approve that this is an appropriate use, and that an appropriate legal gateway exists. It is likely that some public acceptability testing would be needed as evidence to support (or not) the use of administrative data for sampling in this way.

5.2 Ethics

There is not usually an opt-out for administrative data – in order to engage with a particular service it is necessary to provide data. In a few cases it may be possible to say that these data should not be further used, for example through the national data opt-out for health records in England. In general, however, there is no opportunity for data subjects to make an informed

decision on additional uses of administrative data, such as for sampling. Where existing data are used to create (aggregate) statistics without further input from the data subjects, there is a need for public interest and equality impact tests to ensure that the outputs are appropriate and do not unequally affect different subgroups.

It is more difficult to define an ethical use of administrative data as an input to a sample selection. There is still no opportunity for data subjects to make an informed decision on whether data can be used in a sampling procedure. And the whole purpose is to select some sample people/households which will be approached – and differentially according to characteristics, as that is the basis on which efficient samples can be drawn. A public interest test therefore needs to balance the inconvenience to sampled people/households against the benefit of better (lower variance) and/or more efficient (lower cost) survey outputs. And at least some demonstrable equality is needed, so any sample should retain a large element of randomness – that is, no subgroup in the designed sample should have a large probability of inclusion in the sample, or be at risk of being exhausted over surveys in several years. Most general population surveys deal with sufficiently large populations that these are not onerous restrictions, but there will be a need for some case law defining a minimum size for a design variable.

The ethical situation is perhaps assisted in that participation in social surveys in the UK is voluntary and based on informed consent, but there is still a need for sampled people to expend cognitive efforts to decide whether or not to participate.

At the moment the National Statistics Data Ethics Advisory Committee has not approved any use of data in the field (so even the practice from Electoral Register sampling days of having a named person selected for within-household sampling (section 4.2.1) might not be allowed today). See NSDEC (2024, item 2) for a useful summary of the issues to be considered in using information from administrative data in “dependent sampling”.

5.3 Linkage

There is also an ethical side to linkage – where the purpose of linkage is to produce aggregate statistics, the data may be processed under exceptions in legislation which allow for the use of data for statistical purposes. But for sampling this reasoning does not clearly hold. There is in general no way to gather informed consent with administrative data, so again a public acceptability evaluation would be needed.

Linkage operates with both methodologies and with information. The methodological toolkit is continually developing, but a certain quality of information is needed to be able to link records. In evaluation for an administrative data census the ONS used several deterministic match keys and probabilistic record linkage. Some administrative records could not be linked to the demographic index.

5.4 Costs

There are many processes in making administrative data available to enhance a sampling frame. The first step is obtaining and maintaining access, then datasets need to be gathered regularly and stored, assessed, cleaned and linked to a sampling frame. All of these activities come with costs, and we return to this issue in section 6.2 below.

6. Sampling facility

The greatest opportunity for an improvement in sampling seems to be the provision of administrative (and potentially other alternative) data sources linked with an address frame. Given the ethical challenges of using any of this information in the data collection process (see 5.1), such a facility would need to be restricted to survey sampling. It would also need to be restricted to social surveys run by or on behalf of public bodies or similar organisations, and would need to be subject to a public interest test (see below). While such a facility would undoubtedly be of interest and use more widely, such uses would not be consistent with currently available gateways for use of administrative data sources.

There are also challenges in obtaining access to administrative data for sampling purposes. In some instances it has been possible to gain access to specific data (eg Early Life Cohort Feasibility Study (IPSOS 2025)), but this can be very time-consuming. The prospect of gathering a selection of relevant administrative data sources, and with permission for linkage, seems unrealisable outside government. Within the government statistical system, however, it has gradually become more practical to gain access to administrative data and to use them for specific statistical purposes. Therefore it seems clear that any sampling facility could only be run from within government, and given its central position this could best be managed from the UK Statistics Authority.

6.1 Legal basis

The Statistics & Registration Service Act (2007) offers a basis under which such a service could be provided by the UK Statistics Authority (UKSA; referred to in the Act as “the Board”). In particular it offers:

“22 Statistical services

(1) The Board may provide statistical services to any person in any place within or outside the United Kingdom.

(2) The services which may be provided under this section include in particular—

(a) providing information, advice and technical assistance in relation to statistics;”

and

“23 Statistical research

The Board may promote and assist statistical research, in particular by providing access (where it may lawfully do so) to data held by it.”

However, the use of information under clause 22 (but not 23) is restricted in clause 38, where the UKSA would require explicit consent from the data provider(s) for the use of the data in a sampling facility.

This seems *prima facie* to enable a sampling facility to be provided by the UKSA, although there is much detail hidden in what may lawfully be disclosed. The availability of such a sampling facility may turn on the distinction between a “statistical service” and “statistical research”; an approach similar to that of the Secure Research Service which requires a

demonstration that the proposed research is in the public interest, and has appointed an independent board to make this determination, is likely to be required.

6.2 Funding

In order for such a service to be available to the survey community, it would need to be funded, as it is not a core objective of the UKSA to support the wider statistical research community. Indeed, the cost of sampling is already one factor included in costing and planning a survey when a survey organisation bids to run one (Cornick 2026). We distinguish the set-up costs from the running costs. Set-up costs for a sampling facility would involve working through the legal and ethical requirements, and in particular might involve some negotiation and agreement with the custodians of administrative data which could be added to a sampling frame. It would also involve some development of infrastructure, although components which could be used have already been developed in support of administrative data research in the ONS already (see for example ONS (2026)).

The running costs should be met by the survey organisations using the sampling facility. So the cost of maintaining the system and drawing individual samples should be recovered through charges for specific samples. The cost-benefit of such a process might need further evaluation – would the cost be prohibitive, or would the improved quality of sampling, coupled with reduced need to maintain an in-house sampling facility, be sufficient to offset the costs?

6.3 Ethics of release

Several ethical issues on the use of data are covered in section 5.1. Here, we focus on the ethical issues around the provision and use of samples. It would clearly not be appropriate to release information about the characteristics of households or people directly to the survey organisations. The most that seems defensible is to provide a list of addresses which has been obtained through a probability sampling procedure. Probability sampling is important – both as a basis for sound inference, and as a basis for explaining how addresses are included in samples. In effect the addresses are stratified by some suitable characteristics (which may or may not be the actual characteristics of the address/people, but will in general be a strong predictor of those characteristics), and then a probability sample (“random sample”) is taken within each stratum. It would be appropriate to have an upper threshold for the selection probability within a stratum, in order to (a) avoid any suggestion that particular addresses are targeted by the procedure; (b) ameliorate the impact on any particular subpopulations; and (c) ensure that relevant subpopulations are not exhausted within the sampling system (see also 6.4 below).

The impact of (b) is that samples which are focused on specific subpopulations will not be suitable for such a general sampling facility, at least in the initial stages. If such a sampling facility is successful, then further discussion of incorporating more detailed stratification suitable for surveys of subpopulations of policy interest might be worthwhile. But in the first instance it is important to have a general population focus so that the ethical challenges are less.

Although there is considerable interest in being able to sample individuals, this also seems to raise too many ethical questions – a name derived from administrative sources might be of someone no longer residing at the sampled address. Characteristics of individuals can be used as stratifiers for addresses, but they cannot be used directly for sampling. Therefore within-

household selection processes will still be needed, and beyond the control of the sampler for the time being.

All general population surveys in the UK with the exception of the population censuses operate on the principle of informed consent: participation in the surveys is voluntary, and sample members will be provided with sufficient information to be able to make a decision on whether to respond. This is an important principle which is needed to support making samples available to survey organisations.

In addition to the actual sample of addresses, some further information is strongly desirable for use in weighting and estimation. Population totals by different (frame) characteristics would allow samples to be calibrated to at least partially account for differential response within the selected samples. As long as the population totals are for sufficiently large groups/areas, there should be minimal risk of disclosure and there are not likely to be any ethical concerns. Calibration using released data in this manner would rely on the variables being collected in the survey aligning with those on the frame. But some mismatch may occur, which could reduce the quality of calibration slightly (though noting that such calibration would already be a major improvement over what is available from the PAF or, for example, census outputs). If a sampling facility were a success, it could be extended to a weighting facility which would use the original characteristics in calibration. This would, however, require a two-way flow of information at least on response, which would be too ambitious in a first version of the facility.

6.4 Coordination

A benefit of a sampling facility is that samples could be coordinated across the whole social survey system. The ONS already operates a system where a household selected in one of its surveys cannot be included in another for 3 years (with some exceptions for 4-year rotation in APS and during COVID (see <https://www.ons.gov.uk/surveys/informationforhouseholdsandindividuals/respondentcharterforsurveysofhouseholdsandindividuals>)), a type of negative sample coordination induced in this case through rules. But it would be possible to extend this to a wider sample coordination system (Matei & Smith 2023) which could spread surveys more fairly across addresses – even in the case that some subpopulations are overcovered – a likely situation since many surveys seek adequate coverage of subpopulations of policy interest. Potentially this could be extended to exclude households already participating in one of the major longitudinal surveys, if consent to link these were forthcoming.

A facility would also allow some evaluation of the total load imposed both in general and on specific subpopulations by surveys in the UK. In order to be really effective, feedback on which households responded would be valuable, but the challenge of two-way flows of data is again too much to tackle in a first version of the facility.

6.5 Process

In moving towards such a facility, resilience is vital. A long-term commitment is needed to develop and maintain a facility. Since it may eventually replace sampling selection mechanisms within survey organisations, there needs to be confidence in its continued provision over the long term. The initial steps towards a sampling frame have already been put in place within the ONS (ONS 2026, Perera et al. 2026). An initial implementation that demonstrates that the sampling facility is workable and provides some improvement over PAF

or AddressBase by themselves is needed, to justify the resources and effort. This suggests that an important step is an evaluation of the coverage and quality of the RDMF's location index, and this would most readily be achieved through an evaluation connected with Census 2031; indeed this is a golden opportunity to evaluate sampling provision and provide the sound scientific basis for future sampling that has not been done since the evaluations of the PAF (see section 2.2).

7. Conclusions

It seems obvious that AddressBase Premium (ABP) offers advantages over PAF as an address frame for general population surveys, but it also has higher costs. An evaluation of the differences in accuracy between the two sources is therefore needed to facilitate a cost-benefit assessment. While this can be done with a split sample approach (as in Lynn & Taylor 1995), it could very effectively be included as an objective of a census nonresponse link study, such as have been run for several censuses, providing invaluable information on response and coverage of surveys around the time of the census.

A sampling facility is a medium to long term goal with the potential to improve general population survey taking in the UK. Some initial feasibility assessment is included here as a basis for further discussion of legal, ethical and practical approaches. Considerable elaboration will be needed to develop a formal proposal and make it into reality. Advocacy for such a system is important, to make the case that it can be operationally feasible and articulate its benefits against the associated costs. The setup costs would need to be financed from a suitable source – it is unrealistic to expect that these costs can be covered during the setup phase by the use of the service, so some up-front investment in infrastructure will be needed. There will also be a need for further negotiation with suppliers of administrative data sources to ensure that data can be used in such a fashion. Starting the facility with the least controversial situations (by setting constraints which restrict the available use cases) may be a way to build acceptability and confidence before considering whether it is possible to extend the system to a wider range of applications.

A sampling facility based on linking administrative data to PAF and/or AddressBase could also be evaluated alongside the 2031 censuses. Although nonresponse link studies have focused on nonresponse assessments for both surveys and the census since 2001, there is a strong case for building evaluation of the current options for sample frames into the processing of the 2031 censuses (in all the nations of the UK). A part of this is likely to be included automatically as the quality of the address list will need to be evaluated as a component of the quality measurement for the census outputs. But we call for inclusion of other elements of frame evaluation, with appropriate resources.

References

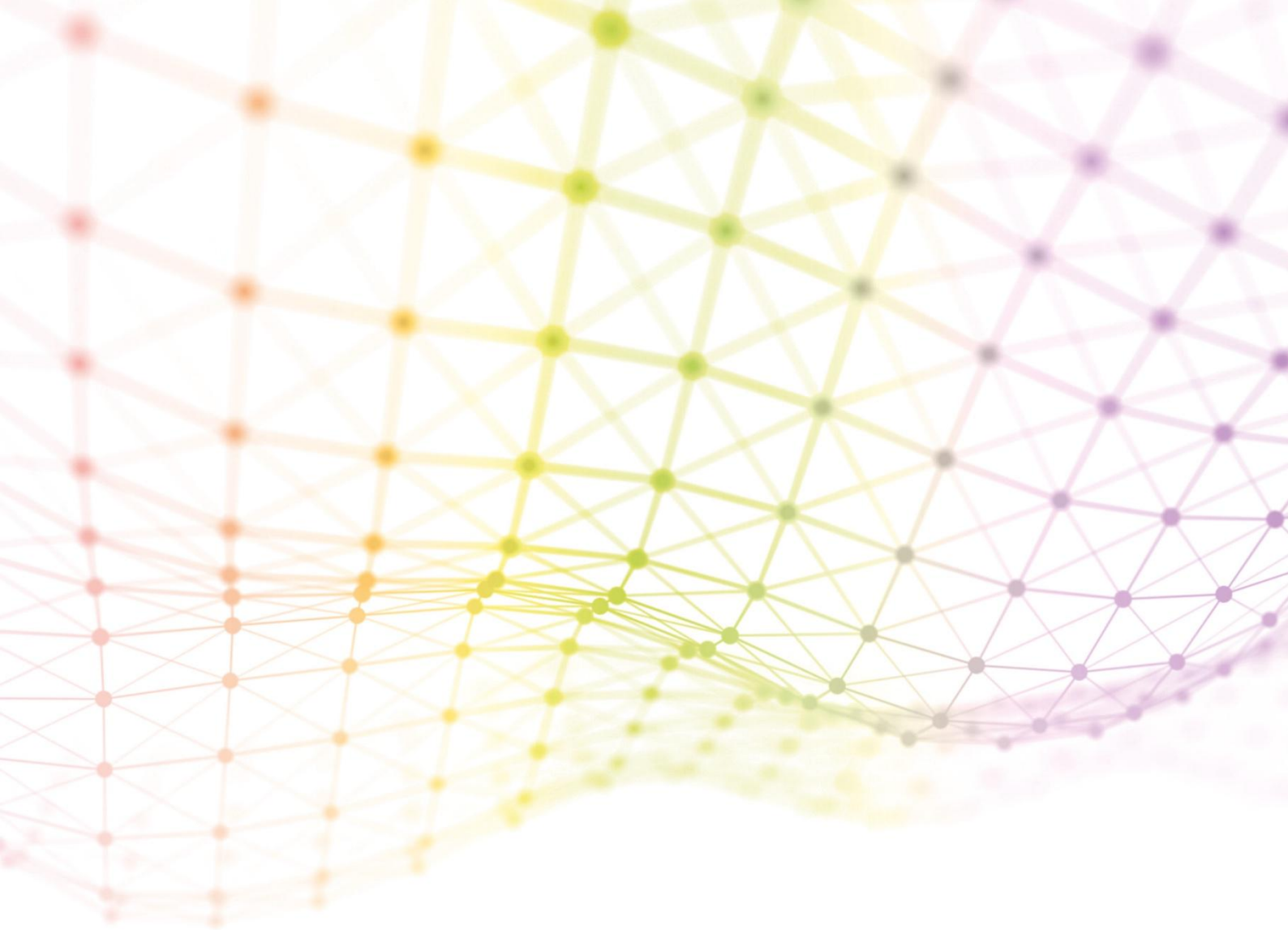
- Barnes, R. (1986) Use of registers and population censuses as sampling frames in Great Britain. *Statistical Journal of the United Nations ECE* **4** 81-92.
- Bond, D. & Lievesley, D. (1993) Address-based sampling in Northern Ireland. *Journal of the Royal Statistical Society, Series D* **42** 297-304.
- Cornick, P. (2026) Using the ONS Reference Data Management Framework: a cost-quality comparison. *Survey Futures Report* **16**. <https://surveyfutures.net/wp-content/uploads/>

[2026/07/report-16-ons-reference-data-management-framework-cost-quality-comparison.pdf](#).

- Dillman, D.A., Smyth, J.D. & Christian, L.M. (2014) *Internet, phone, mail and mixed-mode surveys*, 4th edition. Wiley: Hoboken.
- Emmett, B.P. (1964) Reflections on the state of population sampling in the United Kingdom. *Journal of the Royal Statistical Society, Series C* **13** 146-157.
- Foster, K. (1994) The coverage of the Postcode Address File as a sampling frame. *Survey Methodology Bulletin* **34** 9-18.
- Gibin, M., van Dijk, J., Ye, Z. & Longley, P. (2025) From names to numbers: modelling age and gender profiles from consumer data. *Information Geography* **1**(2) 100023.
- Gray, P.G. & Corlett, T. (1950) Sampling for the social survey (with discussion). *Journal of the Royal Statistical Society, Series A* **113** 150-206.
- Hafner, L., Peifer, T.P. & Hafner, F.S. (2024) Equal accuracy for Andrew and Abubakar—detecting and mitigating bias in name-ethnicity classification algorithms. *AI & Society* **39** 1605-1629.
- Hughes, A.O., Fenton, S. & Hine, C.E. (1995) Strategies for sampling black and ethnic minority populations. *Journal of Public Health* **17** 187-192.
- IPSOS (2025) *Early Life Cohort Feasibility Study Technical Report*. Available from <https://cls.ucl.ac.uk/wp-content/uploads/2026/02/ELC-FS-Technical-Report.pdf> (accessed 10 June 2026)
- Jain, V., Enamorado, T. & Rudin, C. (2022) The importance of being Ernest, Ekundayo, or Eswari: An interpretable machine learning approach to name-based ethnicity classification. *Harvard Data Science Review* **4**(3). <https://doi.org/10.1162/99608f92.db1aba8b>
- Johnston, N. (2026) *Electoral registers and access to them*. House of Commons Library. Available from <https://researchbriefings.files.parliament.uk/documents/SN01020/SN01020.pdf> (accessed 19 May 2026).
- Keyes, A. & Smith, P.A. (2026) Enhanced sampling frames: a literature review. *Survey Futures Working Paper 17*, University of Essex, Colchester. Available from <https://surveyfutures.net/wp-content/uploads/2026/05/working-paper-17-enhanced-sampling-frames-a-literature-review.pdf>
- Lievesley, D., Breeze, E. & Owen, D. (1978) *Postcode sampling*. London: OPCS.
- Lynn, P. (1997) Sampling frame effects on the British Crime Survey. *Journal of the Royal Statistical Society, Series A* **160** 253-269.
- Lynn, P., & Taylor, B. (1995) On the bias and variance of samples of individuals: a comparison of the Electoral Registers and Postcode Address File as sampling frames. *Journal of the Royal Statistical Society, Series D* **44** 173-194.
- Matei, A. & Smith, P.A. (2023) Sample coordination methods and systems for establishment surveys. Chapter 28 (pp637-657) in Snijders, G., Bavdaž, M., Bender, S., Jones, J., MacFeely, S., Sakshaug, J.W, Thompson, K.J. & van Delden, A. (eds) *Advances in Business Statistics, Methods and Data Collection*. Wiley: Hoboken.
- Mateos, P. (2007) A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place* **13** 243-263.

- McLean, I. & Smith, J. (1995) The poll tax, the electoral register, and the 1991 census: an update. Pp. 128-147 in Broughton, D., Farrell, D.M., Denver, D. & Rallings, C. (eds.) *British Elections and Parties Yearbook 1994*. Frank Cass: London.
- Moser, C.A. (1955) Recent developments in the sampling of human populations in Great Britain. *Journal of the American Statistical Association* **50** 1195-1214.
- Nicoll, A., Bassett, K. & Ulijaszek, S.J. (1986) What's in a name? Accuracy of using surnames and forenames in ascribing Asian ethnic identity in English populations. *Journal of Epidemiology & Community Health* **40** 364-368.
- NSDEC (2024) National Statistician's Data Ethics Advisory Committee, Minutes 23rd April 2024. Available from <https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2024/08/National-Statisticians-Data-Ethics-Advisory-Committee-minutes-23-April-2024.pdf> (accessed 15 June 2026).
- ONS (2010) Building the address register for the 2011 Census. Available from https://www.ons.gov.uk/file?uri=/census/2011census/howourcensusworks/designforthe2011census/compilingtheaddressregister/address-register-overview_tcm77-189342.pdf (accessed 15 June 2026).
- ONS (2023a) Evaluation of addressing quality: Census 2021. Available from <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/evaluationofaddressingqualitycensus2021> (accessed 2 June 2026).
- ONS (2023b) Developing Statistical Population Datasets, England and Wales: 2021. Available from <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/developingstatisticalpopulationdatasetsenglandandwales/2021> (accessed 15 June 2026).
- ONS (2023c) Understanding quality of the Statistical Population Dataset in England and Wales using the 2021 Census - Demographic Index linkage. Available from <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/understandingqualityofthestatisticalpopulationdatasetinenglandandwalesusingthe2021censusedemographicindexlinkage/2023-02-28> (accessed 10 June 2026).
- ONS (2026) Survey Futures: Reference Data Management Framework project - what is the RDMF? Available from https://www.ons.gov.uk/file?uri=/aboutus/whatwedo/programmesandprojects/referencedatamanagementframework/referencedatamanagementframeworkwhatistherdmf_v03.pdf (accessed 2 June 2026).
- Perera, Z.M., Nicolaas, G. & Waugh, N. (2026) How could the Reference Data Management Framework improve survey quality? *Survey Futures Report* no. 15. University of Essex, Colchester. Available from <https://surveyfutures.net/wp-content/uploads/2026/06/report-15-how-could-rdmf-improve-survey-quality.pdf>.
- Simon, A. (2014) Using administrative data for constructing sampling frames and replacing data collected through surveys. *International Journal of Social Research Methodology* **17** 185-196. doi: 10.1080/13645579.2012.733176.
- Smith, P., Ashworth, K. & Lound, C. (2011) Using administrative data sources to target sampling in the Wealth and Assets Survey in the UK. *Proceedings of Statistics Canada Symposium 2010* 135-140.
- Tipping, S. & Nicolaas, G. (2006) In search of a population sampling frame for UK postal surveys. *Survey Methodology Bulletin* **58** 34-36.

- UKSA (2021) 2021 Census Address Checking. Available from <https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2021/07/EAP152-Address-Checking.pdf> (accessed 2 June 2026).
- Wilson, P.R. & Elliot, D.J. (1987) An evaluation of the Postcode Address File as a sampling frame and its use within OPCS. *Journal of the Royal Statistical Society, Series A* **150** 230-240



University of Essex



University of
Southampton



Economic
and Social
Research Council

NCRM NATIONAL CENTRE FOR
RESEARCH METHODS

Office for
National Statistics

UCL

WARWICK
THE UNIVERSITY OF WARWICK

MANCHESTER
The University of Manchester

CITY
UNIVERSITY OF LONDON
UNIVERSITY OF LONDON

**National Centre
for Social Research**

LSE LONDON SCHOOL
OF ECONOMICS AND
POLITICAL SCIENCE

Uster University

Unil
UNIL | Université de Lausanne

Ipsos

verian
Formerly Kantar Public

www.surveyfutures.net