SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

Survey Practice Guide 1: Data Integration

Options for Integrating Survey and Non-Survey Data

Thomas O'Toole[1], Alexandru Cernat[1], Nikos Tzavidis[2], Natalie Shlomo[1], Joe Sakshaug[3]

[1]University of Manchester, [2]University of Southampton, [3]Institute for Employment Research (IAB) & LMU-Munich

May 2025

Contents

Executive Summary

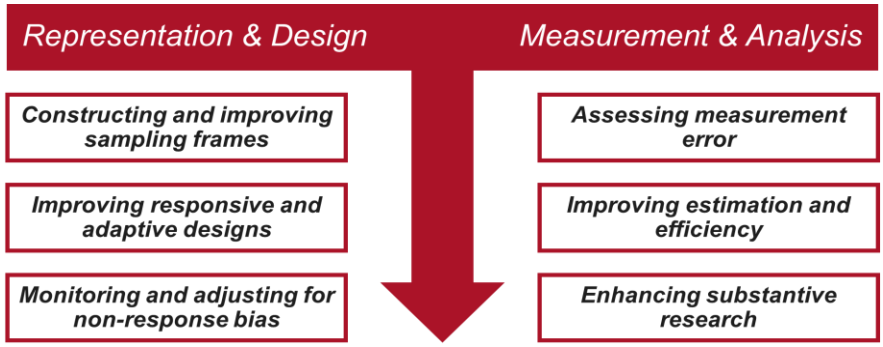
This practitioner guide offers a comprehensive overview of integrating survey and non-survey data, targeting researchers, research commissioners, and survey practitioners. Its primary aim is to assist readers in determining if, when, and how data linkage can effectively address their research questions or operational needs. The guide also includes two illustrative case studies from the Centre for Longitudinal Studies and Office for National Statistics on the integration of survey data with administrative and geospatial data.

In this guide we define data integration as the process of bringing together information from multiple data sources in a coherent and consistent manner. This process makes it possible to examine relationships between factors which might not be available in any one data source alone.

How can Non-Survey Data Enhance Survey Data?

The integration of administrative records, geospatial characteristics and digital trace data with survey data can have use-cases across stages of design, measurement and analysis (figure 1).

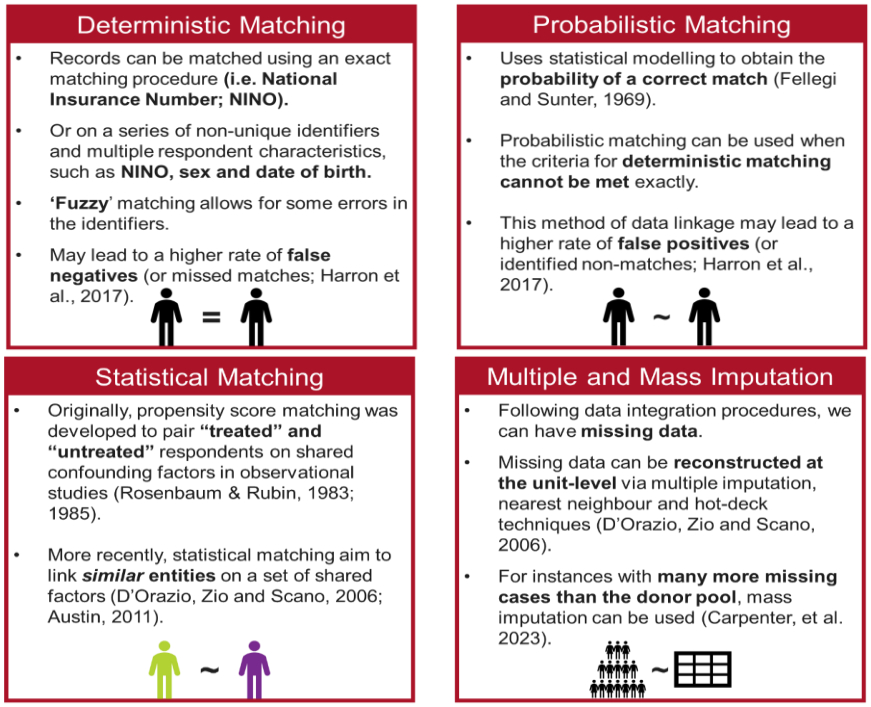
Figure 1. Data integration use-cases



How are Data Sources Integrated?

Survey and non-survey microdata can be integrated using a variety of techniques, including deterministic and probabilistic matching of the same individual across data sources. Additional procedures include statistical matching (which refers to the matching of similar entities across data sources), and multiple and mass imputation (which can be used to reconstruct missing data). See figure 2 for more information.

Figure 2. Linkage and matching procedures



Sources of Integrated Data

Accessing Data

- In the United Kingdom, survey-to-non-survey integrated data is often available via the data holder's trusted research environment.
- Accredited researcher status under the Digital Economy Act is required to access potentially disclosive data.
- Any outputs must adhere to ethical and statistical disclosure requirements.

Trusted Research Environments

The most prominent secure data access services include:

- *The UK Data Service (UKDS)*
- *The Office for National Statistics (ONS)*
- *The UK Longitudinal Linkage Collaboration (UKLLC)*
- *The Secure Anonymised Information Linkage (SAIL) Databank*
- *Research Data Scotland's (RDS) Research Access Service*

Administrative Data

Background

- **Administrative data** is primarily collected for **routine, operational purposes** and is recorded when an individual interacts with (an often public) service.
- As such, **administrative data** is often tied to an **observed event or phenomena**
- **Administrative data** is often linked to survey data at the **individual level**, using unique and non-unique identifiers such as **National Insurance Number, sex and date of birth.**

Administrative Data Research UK (ADR UK) is a UK-wide partnership focussing on getting public sector data "research ready".

Challenges

- Administrative data is often not **"research ready",** which can lead to errors in inference when integrated with survey data.
- Administrative data can **lack the conceptual specificity** of social surveys.
- **Missing data** can occur because of incomplete recording but also because of a failure to interact with a service.
- **Consent to linkage**, along with missed or incorrect linkages, can introduce further bias and errors in the dataset.
- **Over-coverage** can also be a source of error as outdated information is not deleted appropriately.

The UKDS, ONS, UKLLC, SAIL and RDS house and are permitted to integrate a range of survey and administrative data sources, including:

Health Data

- NHS England, Scotland and Wales hospital episode statistics: outpatient, admitted patient care and accident & emergency, and cancer and Office for National Statistics mortality records

Education Data

- The national pupil database (pupil records in Scotland and Wales) and individualised learner records from the Department for Education.

Employment and Income Data

- Benefit receipt, tax credits from the Department for Work and Pensions and PAYE data and HM Revenue & Customs.

Geospatial Data

Background

- **Geospatial data** is collected via satellite imagery or sensors and can be processed to produce area-level statistics for a given zone, for example:
    - ***Government region (Figure 3)***
    - ***Middle/Lower Super Output Area (M/LSOA)***
    - ***Postcode***
    - ***km x km grid***
    - ***Respondent unit***
- These variables can be linked at the selected spatial scale with survey data to add contextual geospatial variables for each respondent.

WorldPop develops global, open access geospatial and demographic datasets to produce global gridded population estimates.

Challenges

There are a number of challenges inherent to working with integrated geospatial and survey data:

- There can be **temporal inconsistencies** between the survey and geospatial datasets.
- Geospatial data is often historical data, and the **reliability of estimates may change** over time as measurement technologies improve.
- **Aggregation** to the selected spatial scale **may lead to a loss of information**.
- Administrative **boundaries may introduce statistical bias** from using arbitrarily classified units to report spatial patterning.

For example, via the UK Longitudinal Linkage Collaboration, there are a number of permitted linkages of geospatial characteristics, including:

Air quality

- Nitrogen dioxide (NO2) and fine particulate matter (PM2.5) from the Department for Environment, Food & Rural Affairs.

Access to Healthy Assets and Hazards

- Retail environment, health services, physical environment and Air quality (NO2, PM10, SO2) from the Consumer Data Research UK.

Energy Performance Certificates

- Energy efficiency: average energy efficiency ratings, energy use, carbon dioxide emissions, fuel costs, average floor area sizes and numbers of certificates recorded from the Department for Levelling Up, Housing and Communities.

Figure 3. Map of government regions (GOV.UK, 2021)

Digital Trace Data

Background

- **Digital trace data** is derived from interactions with digital platforms, capturing behaviours and trends.
- **Digital trace data** is often **collected** or **donated** at the respondent level from a subsample of **consenting survey respondents**. Digital trace data can be collected via:

  Web scraping
  - A programming interface that allows data to be collected directly from applications.

  Smart apps
  - URLs, app usage, geolocation.

  Document scanning
  - Via mobile receipt-scanning apps.

  Data donation
  - Downloaded by survey respondents and donated.

Challenges

- The integration of survey and digital trace data can also present challenges for data quality, such as noise, data sparsity, and non-response bias.
  - For example, recent **restrictions on platform access**, such as Twitter's (X) API paywall, further complicate its integration with survey data.
  - **Digital trace data** must be **identifiable** and linkable to a unit of interest.
  - **Measurement error** can be difficult to assess, as similar digital trace and survey data may not capture the same **underlying construct**.

The Smart Data Donation Service is a new initiative for UK smart data donation and integration.

Types of digital trace data often integrated with survey data include:

Social media
- Platform-level data such as posts, likes, shares and follows, in addition to post-level sentiment, syntax and lexical variables.

Digital transactions
- Banking information/transactions, loyalty card data.

Health data
- Wearable trackers (e.g. accelerometry data).

GPS data
- Real-time information from geographical positioning systems.

Sensor information
- For example, air quality captured by sensors worn by individuals.

Future Directions

Optimising questionnaire design: reducing respondent burden by integrating administrative records for routinely collected variables such as benefit receipt, PAYE and tax

Updating sampling frames: using geospatial gridded sampling methods to update sampling frames more frequently than PAF/Census-based methods allow for.

Improving measurement: reducing measurement error, calculating more effective non-response weighting and better targeting responsive and adaptive deigns.

1. Introduction

Social surveys can collect rich, self-reported information on a wide range of topics from a sample of respondents. Ideally, these variables should be aligned with theoretical constructs of interest and of high quality. This methodology allows researchers to test specific hypotheses or explore nuanced aspects of human behaviour that are representative of the underlying population. However, surveys can be limited by a range of errors in measurement (such as validity, reliability and processing error) and in representation (such as coverage, sampling and non-response error), all of which can affect the accuracy and representativeness of the information gathered (Groves, 2010). Moreover, survey data can be costly to collect, particularly through large-scale probability-based surveys, which often require large setup costs, infrastructure, and staffing.

Non-survey data, such as administrative records, geospatial characteristics, and digital trace data, can offer potentially cost-effective and complementary information to the information collected by surveys. Non-survey data can include objective measures such as medical diagnoses from health records (e.g. Hospital Episode Statistics), geospatial characteristics (e.g. air pollution data) or social media interactions (e.g. from X [formerly known as Twitter]), which would be difficult to collect with high accuracy via traditional social surveys. By integrating or linking surveys with non-survey data, we can create richer data for researchers and policymakers. However, issues such as linkage error and data confidentiality pose challenges for researchers working with integrated survey and non-survey data (Harron et al., 2017).

This document is a guide that covers the concepts and rationale behind various forms of data integration, as informed by the current literature. We present an illustrative typology of different survey and non-survey data sources, for which a systematic review of integrated data literature was conducted. The guide concludes with practical examples of recent data integration conducted by the Centre for Longitudinal Studies (CLS) and the Office for National Statistics (ONS).

1.1. What is Data Integration?

Data integration refers to the process of bringing together information from multiple data sources in a coherent and consistent manner, making it possible to examine relationships between factors which might not be visible from any one data source alone. Data integration has multiple varied use cases and can supplement conventional surveys or create population-level cohorts entirely derived from administrative and non-survey data sources (Harron, 2022). Similarly, multiple sources of survey data may be integrated with non-survey data to create a larger dataset with broader content and population coverage than any one source alone. This can be done directly through record linkage approaches or indirectly using statistical matching or model-based mass-imputation techniques (D'Orazio, Di Zio and Scano, 2006; Han & Lahiri, 2018). This guide focuses on the various options available to researchers and practitioners when using integrated survey data and non-survey data.

1.2. Purposes of Data Integration

In this section we provide more information on specific uses for integrating survey and non-survey data. They can be categorised into several broad purposes.

The improvement of survey representation and design:

- Through the construction and improvement of sampling frames
- Improvement of responsive and adaptive designs
- By monitoring and adjusting for non-response bias

The enhancement of measurement and analysis:

- Through the assessment of measurement error
- Improvement of estimation and efficiency
- Enhancement of substantive research

1.2.1. Constructing and improving survey sampling frames.

The linkage of survey and non-survey data can create more comprehensive and detailed sampling frames that can better enumerate and describe the target population (Mooney & Garber, 2019). For example, the Postal Address File (PAF) is a critical resource for constructing survey sampling frames in the UK, providing comprehensive address information and enabling researchers to target specific populations more effectively. While there is very little under-coverage in the PAF, using the PAF to construct sampling frames has inherent limitations. The PAF only includes address and geographical location and lacks the necessary information to accurately stratify and target traditionally under-represented groups.

As such, the integration and augmentation of the PAF with data sources such as the census, small area statistics or neighbourhood statistics (WorldPop, 2025), is necessary to construct sufficiently detailed sampling frames. This enables the oversampling of underrepresented groups and improves the accuracy of survey results. One such example comes from the Understanding Society Ethnic Minority Boost Sample (Berthoud, Fumagalli, Lynn & Platt, 2009), which used the 2007 Annual Population Survey (APS) to update 2001 Census estimates for geographical ethnic density by postcode area using regression modelling. Linked Census data was similarly used to enhance PAF and construct a frame that allowed over-sampling of both ethnic minorities and people born outside of the UK for the Understanding Society Immigrant and Ethnic Minority Boost sample (Lynn, Nandi, Parutis & Platt, 2018).

However, non-survey data sources may not be designed for research purposes, which can lead to coverage and measurement errors, especially for populations less likely to interact with public services, such as individuals with mental health issues or those in transient populations (Gasteen, 2022). Despite these challenges, integrating survey and non-survey data can improve the construction of sampling frames, ensuring more accurate and representative survey data.

1.2.2. Improving Responsive and Adaptive Designs

Non-response in surveys can lead to significant biases if not properly addressed. In social surveys, individuals who do not respond may systematically differ from those who do, based on key characteristics such as sex, gender, ethnicity, and socioeconomic classification. In longitudinal social surveys, individual unit non-response often accumulates over time (i.e., attrition). Selection bias typically occurs when units from the previous sweep of data collection are unable to be observed in the next sweep, as initially willing participants drop out of the study at later waves (Sakshaug, 2022).

Integrating non-survey data with survey data can help to better target survey data collection for cross-sectional surveys – or the first wave of longitudinal surveys – via the use of responsive or adaptive design (RAD) protocols, using more covariates to inform case prioritisation during the data collection process. (At subsequent waves of longitudinal surveys, integrated non-survey data is not required as RAD protocols can be based on survey data from earlier waves.) Responsive survey designs centre around the "phase capacity" of data collection protocols, beyond which additional data collection does not contribute to improvements in survey estimates. This may be cost-ineffective when compared to starting another phase of data collection with an updated protocol (Groves & Heeringa, 2006). Adaptive survey designs refer to the within-phase allocation of respondent groups to different data collection protocols to better target groups who are, for example, less likely to respond or who may be of greater interest to the survey design (Schouten, Peytchev, and Wagner 2017: for an overview of how this is done on UK surveys, see Sladka and Lynn, 2025). These two approaches are complementary and often used in tandem to focus data collection efforts on under-represented groups or to make the study more cost-effective (Groves & Heeringa, 2006). Adaptive and responsive survey designs can lead to a more representative respondent pool, thus reducing the differences between those responding and not responding to the survey. For example, the Community Life Survey (Verian, 2024) links geodemographic data from a commercial supplier to the PAF in order to be able to implement an adaptive design in which sample addresses are targeted based on expected age profile and deprivation index (Williams, 2024). The Italian Population Census has been enriched with a range of administrative data, illustrating how an adaptive design might be used to minimise and optimise CAPI interviews to compensate coverage errors (De Vitiis et al., 2024).

Through the use of RAD, resources can be more efficiently allocated (and reallocated) to capture the under-represented groups, leading to a more representative sample and requiring fewer weighting adjustments, thus leading to more statistically efficient estimates (Zhang & Wagner, 2022). However, the efficacy of RAD lies in its implementation; for example, implementing a single protocol for all cases may exacerbate existing biases. Further, the increased cost needed to implement a more granular approach may not be feasible (Tourangeau, Brick, Lohr & Li, 2017).

1.2.3. Monitoring and Adjusting for Non-Response Bias

If non-response biases persist (even after using RADs), then post-survey adjustments, such as weighting, can be applied to rebalance the sample. The integration of survey

and administrative data can help monitor, evaluate, and adjust survey data for selection bias, including coverage, sampling, and non-response biases. For example, non-response can be adjusted through the use of inverse-probability weighting, which includes modelling the probability of participation and applying the inverse of those probabilities as unit-level weights to correct for nonresponse bias (Mansournia & Altman, 2016).

Integrated survey and administrative data have been used to monitor and address non-response bias via multiple imputation and non-response weighting. For example, Rajah et al. (2023) used linked data from the National Child Development Study (NCDS) and Hospital Episode Statistics (HES) to evaluate and improve the selection of the integrated data. They identified ten HES variables, such as mental health treatment and hospital visits, that predicted non-response at age 55. These were included as auxiliary variables in multiple imputation models but only offered a minor improvement to representativeness beyond traditional survey predictors such as socio-economic background and cognitive ability. A limitation was that variables available in the administrative HES data did not necessarily capture factors influencing survey participation, limiting its effectiveness in reducing bias.

Another example using integrated survey and administrative data to improve non-response weighting comes from Integrated Employment Biographies (IEB) data, which have been linked to the National Educational Panel Study (NEPS) in Germany (Büttner, Sakshaug & Vicari, 2021). By incorporating administrative employment data such as job changes, unemployment spells, and income in the estimation of non-response weighting, findings show a modest reduction in attrition bias over eight waves of survey data. However, the effectiveness of this procedure varies across different survey estimates, and administrative data were only available for respondents who consented to linkage, introducing potential selection bias. Additionally, some key life events influencing non-response, like health changes, were not captured in employment records, limiting their impact on weighting adjustments.

1.2.4. Assessing Measurement Error

Combining survey and non-survey data can improve data quality by identifying and correcting measurement errors. Measurement error occurs when observed responses – in either survey or non-survey data – deviate from true values (Groves, 1989). Administrative data can provide detailed longitudinal information which can be used to validate and correct survey responses (Sakshaug & Antoni, 2018). Both survey and non-survey data can be distorted by different sources of measurement error, but by integrating and comparing measurement across data sources we also correct for them.

An example of measurement error evaluation comes from Jenkins & Rios-Avila (2023), who used linked data from the 2011/12 Family Resources Survey (FRS) and Pay As You Earn (PAYE) records to identify different types of error in the linked data. In survey data, measurement errors arise from inaccurate self-reporting, recall issues, and social desirability bias, while reference period errors occur when the period of time referenced for “annual” gross income is misaligned in the survey and administrative data sources

(for example, tax year and calendar year). Errors in administrative data include measurement errors, misreporting of employer payroll data, and linkage errors, which refer to missed or false matching of units between survey and administrative data sources.

To mitigate these sources of error, Jenkins & Rios-Avila (2023) advise the incorporation of covariates (e.g. job stability, work type) to model error variability, the calculation of inverse probability weighting to adjust for biases in data linkage and the estimation of reliability measures to assess consistency across data sources. Explicitly modelling different error types rather than assuming administrative data are a gold-standard error-free data source can also reduce sources of bias and help to improve the accuracy of model estimates.

1.2.5. Improving Estimation and Efficiency

The integration of survey and non-survey data can enhance estimation by increasing the effective sample size, including more cases and richer covariates than is feasible via survey data collection alone. This can lead to more precise statistical estimates with narrower confidence intervals and greater power to identify associations (Merkouris, Smith & Fallows, 2023). For example, sociodemographic characteristics from the Millenium Cohort Study (MCS) have been linked to records from the National Pupil Database (NPD) to address residual confounding and improve model estimation (Silverwood et al., 2024).

Several survey estimation operations (e.g., small area estimation) rely on the availability of population-level data, such as population census data. Access to aggregate level population totals or population microdata may be required depending on the target of estimation. However, lack of access to census microdata because of confidentiality constraints or lack of a recent census can limit the ability of researchers and organisations to produce estimates (Skinner, 2018).

Small area estimation (Rao, 2003; Rao & Molina, 2015) uses auxiliary information linked to the survey data, either at the individual or area level, to carry out model-based estimations. These involve combining direct estimates from the survey data with synthetic estimates obtained from regression modelling on a larger area, using the auxiliary information to inform the models. In newer approaches, particularly for countries focusing on developing an administrative-based census (e.g., the Netherlands), survey data is linked to combined administrative data sources and the gaps from the non-survey cases are completed using mass-imputation techniques to build a statistical register (De Waal & Daalmans, 2017).

The availability of geospatial data has enabled several applications in survey and official statistics, such as producing gridded population data (Stevens et al., 2015) and poverty mapping (Edochie et al., 2024). For example, Meta's Data for Good team has developed a public dataset of relative wealth index which provides micro-estimates of wealth and poverty for low- and middle-income countries at 2.4 km resolution which is integrated with large scale representative surveys (Demographic and Health Surveys) and remote

sensing data which can be used to add contextual data as predictors of wealth to inferential modelling techniques (Chi, Fang, Chatterjee & Blumenstock, 2022).

1.2.6. Enhancing Substantive Research

The integration of survey and non-survey data offers researchers access to a substantially larger pool of variables, allowing for the inclusion of exposure and outcome measures that the survey method did not collect. These measures can include health, education and employment data, in addition to place-based characteristics such as air quality, access to local assets and hazards (UK Longitudinal Linkage Collaboration, 2025) and measures of digital interactions like geolocation, activities, social interactions, and online behavior from respondent's smartphones (Smart Data Donation Service, 2025).

For example, linking individual-level educational attainment records from the National Pupil Database (NPD) with social survey data from Understanding Society allowed for an enriched analysis regarding social class inequalities (Stopforth, Gayle & Boeren, 2020). Another example, this time using geospatial data comes from the work of Baranyi et al. (2024), which linked data from the Scottish Longitudinal Study Birth Cohort of 1936 (SLSBC 1936), with historical, area-level air pollution data from EMEP4UK (Vieno et al., 2016) to estimate the effects of early-life air pollution exposure on limiting and long-term illness later in life. An example of linked survey and digital trace data comes from the Understanding Society Innovation Panel/Twitter linkage (University of Essex, 2024). This linked dataset was used to examine the socio-demographic patterning of social media usage during the COVID-19 pandemic (Wenz, Baghal, Sloan & Jessop, 2021).

1.3. How are Different Data Sources Integrated?

The integration of survey and non-survey data is most commonly performed via record linkage using identifiers unique to the element used for linkage (e.g., individuals). Unique identifying variables can vary depending on the level of linkage required (for example, name, address, national insurance number). Data can be integrated at different levels, given that the datasets contain appropriate identifiers. For example, records can be linked at the individual level, household level, or geographical level (i.e. postcode, lower super output area or geographical region). Due to the sensitive nature of unique identifiers, data integration is usually performed by a trusted third party or the non-survey data holders (Harron, 2022). Survey and non-survey data can alternatively be linked statistically, i.e. similar entities are linked on a set of variables of interest (for example, age, gender, income, occupation) through data integration approaches like statistical matching (including the use of propensity scores) and mass imputation which will be described below.

It is also important to consider that survey data is often linked to other sources of survey data, as non-survey data is often linked to other sources of non-survey data. An example of non-survey to non-survey data linkage comes from the ECHILD cohort (Education and Child Health Insights from Linked Data; Grath-Lone et al., 2022), which uses linked National Pupil Database (NPD) and Hospital Episode Statistic (HES) data to

construct an administrative cohort for all children and young people aged 0–24 years in England who were born between 1 September 1995 and 31 August 2020. An example of survey-to-survey data integration is the combination of surveys of different types, such as a smaller probability survey with a larger non-probability one, to improve estimator efficiency or facilitate small area estimation (Rao & Molina, 2015). One example comes from the German Internet Panel (GIP), which maintains a longitudinal probability survey. Alongside the 2015 wave, they collected in parallel eight independent non-probability panels, which are statistically integrated to evaluate the inference of small probability samples based on estimates from a larger non-probability sample (Sakshaug, Wiśniowski, Ruiz & Blom, 2019; Wiśniowski, Sakshaug, Ruiz & Blom, 2020).

1.3.1. Record Linkage

Record linkage seeks to match records or units across two or more data files, often using unique identifiers. Record linkage can be deterministic (via exact matches) or probabilistic (using statistical modelling to obtain the probability of a correct match). Record linkage applications depend on the data structure and goals of the researcher. The implications of each approach will be discussed in the following sections.

1.3.1.1. Deterministic Matching

If a survey and non-survey dataset share unique identifiers, then records can be matched using an exact matching procedure (i.e. via national insurance number; NINO). Deterministic matching is often carried out using combinations of non-unique identifiers and multiple respondent characteristics, such as NINO, sex and date of birth. Note that deterministic matching does not account for errors in data collection or processing (i.e. spelling errors), but steps can be included to introduce 'fuzziness' in the deterministic matching. For example, instead of linking on a fixed string, one can use string comparators and phonetic codes to allow for errors in the variables. The characters of linkage variables must be a one-to-one match to be considered valid and included in the linked dataset; as such, this method may lead to a higher rate of false negatives (or missed matches; Harron et al., 2017).

Examples of unique identifiers include the national insurance number (NINO) and NHS number, which are commonly used for English administrative data linkage, the Community Health Index (CHI), which is analogous to an NHS number for Scottish linked data, and the Anonymous Linking Field used to link respondents in the Welsh SAIL databank. However, deterministic matching can also be conducted using both unique and non-unique identifiers; for example, records may be linked on combinations of variables such as sex, date of birth or postcode. This option is often used when a unique ID such as NINO is not available, and this can lead to a lower certainty of a true record match (Harron et al., 2017). In cases of deterministic matching, iterations of exact and non-exact linkage are often used sequentially to increase the accuracy of the linkage process, for example initially linking on NINO, sex and date of birth, then NINO and sex and surname for missed matches and finally surname and first name and sex and date of birth (Rihal, Gomes & Henderson, 2021).

1.3.1.2. Probabilistic Matching

If a survey and non-survey dataset do not share a unique identifier or identifiers are subject to errors, then records may be matched using a probabilistic matching procedure, which estimates the probability that two records refer to the same entity (Fellegi & Sunter, 1969). Probabilistic matching can be used to match entities across data sources when the criteria for deterministic matching cannot be met. It involves the calculation of linkage weights, which are used to link units across data sources. These weights represent the match probability based on the overall agreement and disagreement of matching variables in both datasets. Typically, the linkage weights are ordered 0 to 1, with higher weights being indicative of correct matches and lower weights of non-matches. This approach will usually implement a threshold above which matches are classified as correct, and below which matches are incorrect (for example, 80%, 90% and 95% are commonly used depending on the amount of discriminating power inherent in the variables common to the records that need to be matched (Fellegi & Sunter, 1969)). The procedure generally requires some clerical review for those indecisive linkage weights. This method of data linkage may lead to a higher rate of false positives (or identified non-matches; Harron et al., 2017) as well as missed matches.

1.3.1.3. Statistical Matching

An alternative form of matching seeks to statistically match independent sample units rather than linking the same entity across data sources. Statistical matching techniques aim to link similar entities on a set of matching variables, and they are most commonly performed unit-to-unit (Rosenbaum & Rubin, 1983). For example, propensity score matching is a non-experimental causal inference technique in which pairs of "treated" and "untreated" respondents can be statistically matched based on shared confounding factors to make valid between-group comparisons (Rosenbaum & Rubin,1985). This approach mimics the design of a more traditional randomised controlled trial to estimate the effect of an exposure of interest (i.e. treatment or policy) on outcomes of interest (Austin, 2011). More recently, statistical matching has become a tool for integrating data using imputation techniques where datasets are merged according to a set of common auxiliary information. The techniques include regression modelling, predictive mean matching and hot-deck (see D'Orazio, Zio and Scano (2006) for more information on statistical matching). For examples of this procedure in the context of income and expenditure data, see Meinfelder and Schaller (2022) and Donatiello et al. (2022).

1.3.1.4. Mass imputation

When producing a statistical register based on the linkage of administrative data and survey data (either using statistical matching or record linkage techniques), there are gaps in some of the variables of interest for those cases that were not included in the survey. Under probability-based random sampling, the non-surveyed, missing cases may be assumed missing at random (MAR), where missingness is dependent solely on observed data, and hence, imputation processes can be performed to fill in the gaps. Given that there are many more cases that need imputation compared to the donor pool, this is known as mass imputation. Imputation techniques for item non-response

have been adapted for mass imputation, typically using model-based imputation approaches (Carpenter et al. 2023 and references therein).

1.4. Which Data Sources are Commonly Used for Data Integration?

Data integration combines survey and non-survey data to create richer datasets that enhance research and policymaking. This section outlines the most common data sources, detailing their characteristics, benefits, and challenges while providing examples of their integration in practice.

1.4.1. Sources of Survey Data

Survey data is collected directly from individuals through structured questionnaires, interviews, or self-administered forms. Surveys provide self-reported information on behaviours, attitudes, and socio-demographic characteristics. Probability surveys, such as the Labour Force Survey (LFS; Office for National Statistics, 2024a) and the Annual Population Survey (APS; Office for National Statistics, 2024b), use random sampling methods to produce representative estimates of the target population. For example, the LFS monitors employment trends across the UK, while the APS informs socio-economic planning at local and national levels. Non-probability surveys, on the other hand, rely on quota or convenience sampling to target specific groups, often addressing the challenges of underrepresented populations. An example is the Evidence for Equality National Survey (EVENS), which collected data from ethnic minorities in the UK during the COVID-19 pandemic (Finney et al., 2024).

Survey designs include cross-sectional surveys, which collect data at a single point in time, and longitudinal surveys, which track the same individuals over time. Cross-sectional surveys like the Crime Survey for England and Wales (CSEW; Office for National Statistics, 2021) provide a snapshot of societal attitudes and behaviours and can be repeated over time with different respondents. Longitudinal surveys, such as Understanding Society (UKHLS; University of Essex, 2024) and the 1970 British Cohort Study (BCS70), track within-individual changes over time, enabling researchers to study individual-level life course trajectories and long-term patterns (University College London, 2025).

A tool which allows the exploration of global survey data with integrated non-survey data is the Wellcome Atlas of Longitudinal Datasets (Atlas of Longitudinal Datasets, 2024). The platform provides information on longitudinal survey datasets, including design, number of participants, year of first data collection and countries covered, in addition to types of linked data available; administrative (healthcare, education and income and benefits data), geospatial (geographic, spatial and environmental data) and digital trace (social media & technology use data). However, the Wellcome Atlas primarily covers mental health data and only contains information on longitudinal population surveys, and while this resource links to data access options, researchers must request data access independently. Please see appendix A for a detailed overview table of flagship UK survey data with integrated non-survey data, provided by the Altas

of Longitudinal Datasets team. This table includes probability and non-probability surveys, in addition to clinical databases with survey data linkage.

1.4.2. What Can Non-Survey Data Add to Survey Data?

Non-survey data encompasses a wide array of sources collected independently of surveys. These include administrative data, geospatial data, and digital trace data, each offering unique advantages and challenges.

1.4.2.1. Linked Administrative Data

Administrative data is often the byproduct of administrative systems and is chiefly collected for routine, operational purposes. Administrative data is recorded when an individual interacts with (an often public) service and, as such, is often tied to an observed event or phenomena (Harron, 2022). Examples of survey to administrative data integration include:

- **Health Data:** NHS England, Scotland and Wales hospital episode statistics (HES): outpatient, admitted patient care and accident & emergency, and cancer and Office for National Statistics mortality records.

- **Education Data:** The National Pupil Database (pupil records in Scotland and Wales) and individualised learner records from the Department for Education.

- **Employment and Income Data:** Benefit receipt, tax credits from the Department for Work and Pensions and PAYE data from HM Revenue and Customs.

Researchers should bear in mind "research readiness" when looking to use integrated survey and administrative data (Grath-Lone et al., 2022). First, administrative data can lack the conceptual specificity of social surveys as this data is often not collected for research purposes and is not designed to capture attitudes and behaviours. Similarly, data quality can be a concern in administrative data, as missing data can occur because of incomplete recording, but also because of a failure to interact with a service (for example those who do not interact with secondary healthcare will not be present in the HES data, leading to a biased sample). The data integration process can further compound this missingness, as consent to link bias along with missed or incorrect linkages can introduce further errors in the dataset. Furthermore, administrative data benefits from frequent updates on new interactions, but data quality may suffer when outdated information is not deleted appropriately. This typically leads to over-coverage; for example, the same people may be registered at different addresses, and deaths may not be deleted.

1.4.2.2. Linked Geospatial Data

Geospatial data includes location information in the form of coordinates, allowing observations to be mapped to specific geographic locations. This type of data can be

linked to various geometries such as points, lines, polygons, and grids. In the context of social science research, geospatial data is used to enhance survey data with contextual data that acts as a proxy for household or individual characteristics. For example, WorldPop (2025) at the University of Southampton produces datasets which use geospatial data to output global gridded population estimates. Gridded population sampling approaches can also be used to supplement survey design where census data is out of date or absent (Edochie et al., 2024; Appendix B).

The majority of longitudinal population studies in the UK have robust geospatial linkages in place, using output areas from the 2001 and 2011 UK Census, which can be linked to various geospatial characteristics. Geospatial data is collected via satellite imagery or sensors and can be processed to produce area-level statistics for a given zone, for example, Government region, Middle/Lower Super Output, Area (M/LSOA) Postcode, km x km grid and Respondent unit. These variables can be linked at the selected spatial scale with survey data to add contextual geospatial variables for each respondent. Via the UK Longitudinal Linkage Collaboration, there are a number of permitted linkages of geospatial characteristics to UK longitudinal population studies at the LSOA, postcode level, for example:

- **Air Quality:** Nitrogen dioxide (NO2) and fine particulate matter (PM2.5) via the Department for Environment, Food & Rural Affairs.

- **Access to Healthy Assets and Hazards (AHAH):** Retail environment, health services, physical environment and Air quality (NO2, PM10, SO2) via the Consumer Data Research Centre)

- **Energy Performance Certificates:** Energy efficiency; average energy efficiency ratings, energy use, carbon dioxide emissions, fuel costs, average floor area sizes and numbers of certificates recorded via the Department for Levelling Up, Housing & Communities)

Geospatial data can enrich the survey design and analysis stages with environmental and geographical context. However, challenges inherent to both survey and geospatial data remain. For example, there can be temporal inconsistencies between the survey and geospatial datasets; geographical data may consist of annual averages and minimum or maximum values depending on the application, while survey data is often a snapshot of respondent attitudes and behaviours. Similarly, survey-to-geospatial linkages are often cross-sectional due to the complexities of longitudinal geospatial data, and geospatial data is also often historical, and the reliability of estimates may change over time as measurement technologies improve (Jutila et al., 2025). Further, the selected spatial scale may lead to a loss of information. For example, grid cells are often smaller than the selected enumeration area, resulting in information being lost when aggregating, a process which increases as the spatial scale increases (Edochie et al., 2024). Further, administrative boundaries may introduce statistical bias by using arbitrarily classified units to report spatial patterning (Openshaw, 1984).

1.4.2.3. Linked Digital Trace Data

Digital trace (or footprint) data is information generated as a byproduct of an individual's interaction with digital services and environments. Digital trace data encompasses a wide range of user activities and is inherently tied to observed online behaviours and interactions, capturing temporally linked events and trends. Digital trace data is derived from interactions with digital platforms and is well-suited to capturing real-time behaviours, attitudes and trends and is often integrated with sociodemographic variables for substantive and methodological research (Cernat, Keusch, Bach & Pankowska, 2024). Sources of digital trace data commonly linked with survey data include:

- **Social media:** Platform-level data such as posts, likes, shares and follows, in addition to post-level sentiment, syntax and lexical variables (University of Essex, 2024).

- **Digital transactions:** Banking information/transactions, loyalty card data (Wenz et al., 2023).

- **Health data:** From wearable trackers, for example, actigraphy and accelerometery data (Keusch, Struminskaya, Eckman, & Guyer, 2024; Dobson et al., 2023).

- **GPS data:** Real-time information from geographical positioning systems (Bähr, 2019).

- **Sensor information:** For example, air quality captured by sensors worn by individuals (Schulte, 2022).

To be linked with survey data, digital trace data needs to be identifiable and is often collected or donated from a subsample of survey respondents. Types of digital trace data can be collected, accessed and integrated with survey data from a number of sources, including:

- **APIs and web scraping:** Application programming interface that allows data to be collected directly from applications, such as Twitter (X) and Facebook APIs (Baghal, Wenz, Sloan and Jessop, 2021)**.**

- **Smart tracker apps**: URLs, apps usage, geolocation (Vermeulen & Gutiérrez Amaros, 2024; Silber et al., 2022).

- **Document scanning** via a mobile receipt-scanning app (Wenz et al., 2023; Jäckle et al.,2021).

- **Data donation**: data downloaded by survey respondents and donated to researchers (Boeschoten et al., 2022; Carrière et al., 2024). You can refer to the Smart Data Donation Service for a novel initiative for UK smart data donation and integration with survey data (Smart Data Donation Service, 2025).

The integration of survey and digital trace data can also present challenges for data quality, such as noise, data sparsity, and non-response bias. For example, recent restrictions on platform access, such as Twitter's (X) API paywall, further complicate its integration with survey data (Davidson et al., 2023). Digital trace data can also be difficult to estimate survey weights for, as the target population of, for example, social media users is often unknown, so correcting for non-response and associated errors can be difficult. Further, issues such as measurement error can be difficult to assess, as similar social media and survey data may not capture the same underlying construct; linked digital trace and survey data often show a low correlation due to trait differences, method effects and random error (Cernat, Keusch, Bach & Pankowska, 2024).

1.5. Accessing Integrated Data

In the United Kingdom, access to survey-to-non-survey integrated data is often available via the data holder's secure access service. Specific requirements vary depending on the data holder, and jurisdictions, but comprehensive training and approvals are often required for researchers to access the data via secure physical or virtual environments; Trusted Research Environments (TREs) (Harron, 2017). The TRE used will vary depending on the linkage needed, for example the UKDS facilitates access to ONS data via the UKDS SecureLab.

In the context of the United Kingdom, researchers are required to obtain accredited researcher status under the Digital Economy Act (2017), with outputs adhering to statistical disclosure requirements to maintain confidentiality in highly sensitive linked data (i.e. cannot be used to identify individual entities). While often potentially disclosive in nature, linked datasets should, as far as possible, adhere to the characteristics encapsulated within FAIR principles (data should be (Findable, Accessible, Interoperable and Reusable; Wilkinson et al., 2016).

The most prominent Trusted Research Environments include:

- **The UK Data Service (UKDS)** which holds integrated datasets for a large proportion of the UK's social surveys, including Understanding Society (University of Essex, 2023), the Labour Force Survey (Office for National Statistics, 2024a) and a range of cohort studies such as the Next Steps cohort study (University College London, 2024). Once applications have been approved, researchers can access controlled data via application to UKDS.

- **The Office for National Statistics (ONS)** operates the Secure Research Service (SRS) and is currently transitioning to the government-wide Integrated Data Service (IDA). This service provides researchers with access to Census data, along with a wide range of survey, administrative and geospatial datasets, including the Annual Population Survey (APS; Office for National Statistics, 2024b), Annual Survey of Hours and Earnings (ASHE; Office for National statistics, 2025) and the Crime Survey for England and Wales (CSEW; Office for National Statistics, 2021). Many of these datasets are also available via the UKDS. However, specific linkages can only be accessed through the ONS SRS.

- **The UK Longitudinal Linkage Collaboration (UKLLC)** is a national trusted research environment which provides researchers with remote access to UK longitudinal population surveys, including Understanding Society (University of Essex, 2024) and The English Longitudinal Study of Ageing (Banks et al., 2024), linked with NHS England and Wales administrative data, and geospatial data from the Department for Environment, Food & Rural Affairs (UK Longitudinal Linkage Collaboration, 2025).

- **The Secure Anonymised Information Linkage (SAIL) Databank** provides researchers with access to a range of Welsh survey and linked administrative data (Lyons et a., 2009). This includes The Welsh Health Survey, Welsh Census data, and healthcare records from NHS Wales. The SAIL databank does not directly contain traditional social survey data but offers comprehensive administrative and geospatial data, which survey data is often linked to.

- **Research Data Scotland's (RDS) Research Access Service** enables researchers to access linked data from nine of Public Health Scotland's most frequently accessed datasets, including morbidity and birth registrations, mental health and cancer records, and accident and emergency and prescription information (Research Data Scotland, 2025). The RDS datasets are comprised of various administrative sources, which researchers and survey practitioners can link to their respective survey datasets.

1.6. Summary and Examples

Integrating survey and non-survey data enables researchers to leverage the strengths of each data source while helping to address their limitations. For instance, linking administrative health records with survey data on behaviours can identify determinants of health outcomes, while merging geospatial and economic indicators reveals regional disparities. While the benefits of integration include richer analyses and improved selection, challenges such as ensuring data compatibility, managing privacy concerns, and addressing biases require careful consideration. With appropriate methods and safeguards, data integration provides a robust framework for advancing research and informing evidence-based decision-making (Harron et al., 2022).

Table 1 includes some examples of integrated surveys with non-survey datasets. This table covers some of the flagship longitudinal panel studies in the United Kingdom, and linked data available and deposited the UK Data Service and permitted via UK Longitudinal Linkage Collaboration (2025).

Table 1. Example integrated survey and non-survey datasets available via the UK Longitudinal Linkage Collaboration and the UK Data Service

Data Type	Dataset	Avon Longitudinal Study of Parents & Children	1970 British Cohort Study	English Longitudinal Study of Ageing	Millennium Cohort Study	National Child Development Study	Next Steps	Understanding Society
Administrative	**NHS England**	UK LLC	UKDS & UK LLC	UK LLC	UKDS & UK LLC	UKDS & UK LLC	UKDS & UK LLC	UK LLC
	NHS Wales	UK LLC (forthcoming)	UK LLC (forthcoming)	X	UKDS & UK LLC (forthcoming)	UK LLC (forthcoming)	UK LLC (forthcoming)	UK LLC (forthcoming)
	NHS Scotland	UK LLC (TBC)	UKDS & UK LLC (TBC)	UK LLC (TBC)	UKDS & UKLLC (TBC)	UKDS & UKLLC (TBC)	UK LLC (TBC)	UK LLC (TBC)
	E.g. NHS hospital episode statistics: outpatient, admitted patient care and accident & emergency, and cancer and Office for National Statistics mortality records							
	Department for Education (DfE)	UK LLC (TBC)	X	X	UKDS & UKLLC (TBC)	X	UKDS & UKLLC (TBC)	UKDS & UKLLC (TBC)
	E.g. the national pupil database (pupil records in Scotland and Wales) and individualised learner records							
	Department for Work and Pensions (DWP)	UK LLC (forthcoming)	UK LLC (forthcoming)	UK LLC (forthcoming)	UK LLC (forthcoming)	UK LLC (forthcoming)	UK LLC (forthcoming)	UK LLC (forthcoming)
	E.g. Records of benefits receipt							
	HM Revenue and Customs (HMRC)	UK LLC (forthcoming)	UK LLC (forthcoming)	UK LLC (forthcoming)	UK LLC (forthcoming)	UK LLC (forthcoming)	UK LLC (forthcoming)	UK LLC (forthcoming)
	E.g. Tax credits, earnings and employment data							
Geospatial	**Neighbourhood-level Geographies (e.g. lower layer super output areas)**	UK LLC	UKDS & UK LLC	UKDS & UK LLC	UKDS & UK LLC	UKDS & UK LLC	UKDS & UK LLC	UKDS & UK LLC
	Postcode-level Geographies	UK LLC	UKDS	UKDS	UKDS	UKDS	UKDS	UKDS
	E.g. Annual averages of NO2 and PM2.5, noise exposure, green space, healthy assets and hazards and energy performance certificates							
Digital Trace	**Social Media Data**	X	X	X	X	X	X	UKDS
	E.g. Posts, likes, shares and follows, in addition to sentiment, syntax and lexical variables.							

Note: UK LLC = UK Longitudinal Linkage Collaboration; UKDS = UK Data Service

1.7. Future Directions and Recommendations

1.7.1. Optimising Questionnaire Design

- *Survey questionnaire design can be optimised by asking for consent to link to alternative data sources, rather than asking for respondents to provide this data themselves.*

Respondent burden can impact the subsequent quality of survey data. This can be caused by the length of the interview, amount of effort, frequency of interviews, and the stress from the content (Bradburn, 1979). These survey features can be more or less burdensome depending on the characteristics of the respondent, but have been repeatedly associated with respondent motivation, and consequently, with an increased level of non-response (i.e. missing data), and a potentially less valid and accurate response from participants (Data Quality Hub, 2020; Wenemark, Frisman, Svensson and Kristenson, 2010).

By collecting rich behavioural and attitudinal data and supplementing survey questions with non-survey data, researchers can reduce the length of the interview, the respondent effort and potentially the stress from the interview content. Alternatively, with respondent consent, researchers can source the data directly from other sources and skip parts of the questionnaire. For example, in the Canadian census, respondents are not required to answer income-related questions, as Statistics Canada retrieves this information from personal income tax and benefit records provided by the Canada Revenue Agency (CRA) (Statistics Canada, 2023). Further, data donation and document scanning can enable survey practitioners to integrate data which would otherwise be unfeasible to collect. This approach has been extended to web-browsing habits (Bach & Wenz, 2020), real-time geolocation (Bähr et al., 2019) and scanned expenditures (Wenz et al., 2023).

1.7.2. Updating Sampling Frames

- *Sampling frames may be more effectively updated through the use of gridded-sampling approaches, updating geospatial information by integrating richer covariates available via survey data.*

Accurate population estimation at a small geographic scale (e.g., LSOA) is fundamental for the effective design of survey sampling frames. However, the use of census data for sampling frame construction can become outdated due to changes in the population. Traditional methods often lack the granularity or timeliness required leading to challenges in resource allocation and case prioritisation. To address this, the Office for National Statistics (ONS) conducted a proof-of-concept study exploring geospatial approaches for producing top-down household population estimates at the LSOA level (Office for National Statistics, 2021).

The integration of survey and geospatial data can enhance the construction of survey sampling frames. Social survey data are collected more frequently than census data,

meaning that survey practitioners can include more timely sociodemographic covariate information from social surveys to produce more up-to-date population estimates. Incorporating updated and optimised covariate data in small area estimation can enable more effective targeting of the sample population (Newhouse, 2023). In the UK context, this approach can be particularly relevant for areas with a high level of migration and population turnover, such as high-population-density city centres.

1.7.3. Improving Measurement and Estimation

- *Measurement and estimation can be improved by integrating measures from administrative, geospatial and digital trace sources to better address residual confounding, improve survey weighting procedures and better target adaptive survey designs.*

Accurate and timely data is necessary to improve measurement, estimation and consequent inference. However, due to time and budget constraints, it is often not feasible to collect high-frequency survey data. The integration of digital trace data with survey data can provide a number of benefits to measurement and estimation, including real time geolocation, financial information, and digital interactions data.

The integration of various forms of non-structured, non-survey data with structured survey data can help to enhance measurement and estimation through improving residual confounding, calculating more efficient survey weighting, and more effectively targeted responsive and adaptive survey design. In addition to collecting complementary data, the alternative forms of data can also be used to estimate and correct for selection and measurement errors. This can inform future designs and enhance secondary analysis.

2. Case Studies

This section of the practitioner guide spotlights instances of data integration at different levels and using various methodologies. The first example comes from the integration of the Next Steps cohort and administrative data from the student loans company, and the second example illustrates the linkage of the Family Resources Survey data with census data and a range of publicly available geospatial data. The benefits and challenges of each data source and integration method are discussed.

2.1. Centre for Longitudinal Studies

2.1.1. Study Purpose

The purpose of this study was to examine the representativeness and quality of the Next Steps-Student Loans Company (SLC) linked data (Booth et al., 2024). The dataset used in this study includes linked social survey data from the Next Steps cohort with administrative data provided by the Student Loans Company. The linked dataset is available via the UK Data Service under Secure Access (SN 8848). The following sections of this case study use information from the user guide (Rihal, Gomes and Henderson, 2021) and from the representativeness and data quality analysis by Booth et al. (2024). This case study aims to describe the characteristics of the data sources used, illustrate the data linkage process, and evaluate the challenges and opportunities that survey-to-administrative data linkage offers.

2.1.2. Survey and Non-Survey Data Sources

2.1.2.1. Next Steps Survey

Next Steps (University College London, 2024) is a longitudinal cohort study comprised of around 16,000 respondents born in England between the years 1989 and 1990. The study was previously known as the "Longitudinal Study of Young People in England" (LSYPE) and was managed and funded by the Department for Education. Data collection began in 2004 when respondents were aged 14 years and followed up annually until age 20 (2010). In 2015 (age 25), the study was relaunched under the management of the Centre for Longitudinal Studies (CLS) and was funded by the Economic and Social Research Council (ESRC), with sweeps at age 25 (2015; fieldwork conducted by NatCen), and age 32 (fieldwork conducted by IPSOS).

The Next Steps age 25 survey included 7,707 respondents collected via sequential mixed mode methods, involving online, telephone and face-to-face data collection. This sweep of data collection focussed on describing the health, labour market relations, attitudes and political beliefs of the sample. In addition, consent for various data linkages was collected at the age 25 sweep, including consent for record linkage to Student Loans Company records (Rihal, Gomes & Henderson, 2021).

2.1.2.2. Student Loans Company Datasets

The Student Loans Company (SLC) is a non-profit, government-owned organisation that administers grants and loans to students in further and higher education in the United Kingdom. For the purposes of integration with the Next Steps survey, four datasets

covering England were provided by the SLC: 'Applicant', 'Payments', 'Repayments' and 'Overseas'. The datasets include individual-level student loan records for the years 2007 to 2021. Further detailed information can be found in the Next Steps-SLC user guide (Rihal et al., 2021) and data quality report (Booth et al., 2024). The datasets were:

- **The SLC Applicant dataset** consists of individual-level records of student loan applications made between 2007 and 2020 (regardless of whether any payment was actually made), including, for example, the academic year, institution name, course name, mode of study, and household income (for the purpose of means-testing).
- **The SLC Payments dataset** covers student loan payments that were made to students between 2007 and 2021, including, for example, the total amount paid to students by financial year for all loan products, excluding non-repayable products such as grants, stipends and allowances.
- **The SLC Repayments dataset** contains individual-level records of any repayments made to the SLC between 2009 and 2021, including, for example, any voluntary repayments or obligatory repayments made via PAYE or self-assessment.
- **The Overseas dataset** details cohort members who have moved overseas, including the date and country of residence.

2.1.2.3. Data Integration Steps

The linkage between Next Steps and SLC data was carried out in June 2021. In this case study, the data integration procedure was handled by the SLC and followed these steps:

1. In 2019, the CLS contacted the SLC to link all respondents in the Next Steps cohort who provided consent to the student loans records held in the Student Finance Database.

2. Of the 7,707 respondents at the age 25 sweep of Next Steps, 4,501 consented to the data linkage (58%).

3. CLS provided the SLC with a matching file including a proxy ID, first name, surname, sex, date of birth, address, and National Insurance Number (NINO) for the 4,501 consenting respondents.

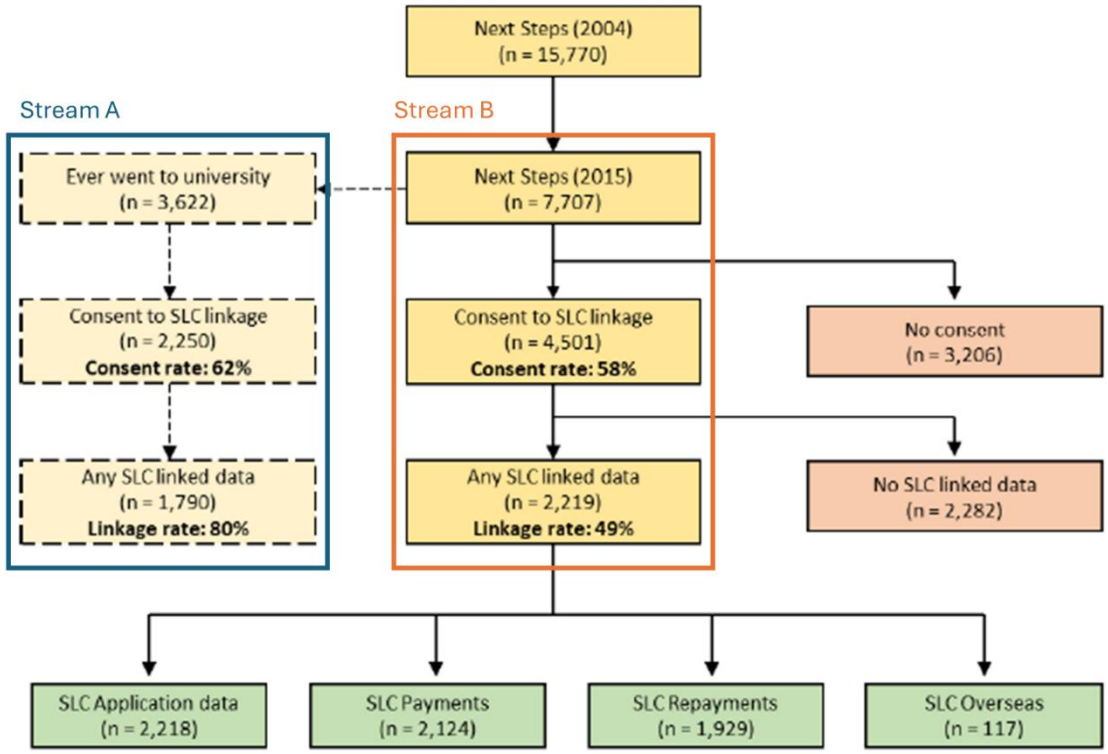
4. Following this, deterministic matching using at least three of the markers provided was conducted. The SLC integrated student loans records to the matching file using the following criteria:
    a. Match 1: NINO and sex and date of birth.
    b. Match 2: NINO and sex and surname.
    c. Match 3: Surname and first name and sex and date of birth.

5. Overall, a total of 2,219 respondents were successfully linked with SLC data representing 49% of consenters (n=4,501) and 29% of the Next Steps age 25 sample (n=7,707).

6. Sample sizes differ between the four datasets provided by the SLC, depending on data availability. Of the 2,219 total linked respondents:
    a. N=2,218 were linked to SLC application data.
    b. N=2,124 were linked to SLC payments data.
    c. N=1,929 were linked to SLC repayment data.
    d. N=117 were linked to SLC overseas data.

7. The linked datasets were returned to CLS in June 2021 and can be accessed via the UK Data service via the study number (SN)8848 (UK Data Service, 2024).

Figure 4. Participant flow chart for the Next Steps/Student Loans Company data linkage (Adapted from Booth et al., 2024)



2.1.3. Benefits and Challenges

Booth et al. (2024) highlight several conceptual and methodological challenges encountered throughout the data integration process, including identifying potential sources of bias in representation due to coverage, sampling and non-response errors (Groves & Lyberg, 2010).

2.1.3.1. Defining the population of interest

A distinct challenge in this case study was defining the population of interest and, consequently, the linkage rate. Student loan records were not available for every respondent in the Next Steps sample, which is primarily contingent on whether or not the survey respondent had ever attended university. As such, when working with linked data, it is necessary to redefine the population of interest to more accurately capture valid linkage rates. Booth et al. (2024) achieved this by conditioning the linked sample on survey respondents who were identified as 'ever attending university' in the survey data. However, this approach was unable to capture other forms of higher education which would have been eligible for a student loan and cannot account for subsequent biases in consent to linkage.

2.1.3.2. Consent bias

A mechanism of non-response error in this case study comes from the differences in socio-demographic characteristics between those who consented to data linkage and those who did not. Booth et al. (2024) highlight that Next Steps respondents from minoritised ethnic groups and lower socio-economic backgrounds were less likely to consent to data linkage. Of those who 'ever attended university', when compared to non-consenters, those who consented to data linkage were more likely to be from a White ethnic background (83% of consenters compared to 68% of non-consenters), and to have a parent with a university degree (31% of consenters compared to 24% of non-consenters). This led to an underrepresentation of these groups in the linked dataset.

However, after further comparison between consented linkages, the full Next Steps sample and national statistics from external data sources, the authors note that sample composition between data sources was broadly similar (Booth et al., 2024). For example, the proportion of those attending Russell group institutions was 26% in both the Next Steps SLC linkage and in external data from the Higher Education Statistics Agency (HESA), suggesting that while a consent bias does exist in this linkage, the effect is minimal.

2.1.3.3. Linkage error

Another important area of data linkage in which bias and error can be introduced is in the matching procedure used to link cases across data sources. Linkage error refers to missed or false matches in the linked data, which can be observed by a discrepancy between the number of respondents who consented to data linkage and reported receiving a student loan and those who appeared in the linked data. Booth et al. (2024) highlight that 15% of Next Steps respondents who consented to have their data linked and reported having taken out a student loan were unable to be matched to SLC records, which suggests a relatively high number of false negatives or missed matches.

This case study used exact, deterministic record linkage procedures, which matched cases based on NINO and iterations of sociodemographic characteristics (as discussed in Figure 4). However, exact matching procedures can be subject to errors (e.g. misreporting of NINO, misspellings in surname and forenames), and while the sequential approach taken in this case study is designed to ameliorate these effects as far as possible, inaccuracies and errors may still remain. Booth et al. (2024) advise on the use of probabilistic matching procedures which estimate the probability that two records refer to the same entity using a variety of predictors (Fellegi & Sunter, 1969). However, this approach may lead to a higher rate of false positives and incorrect matches and is often used for fringe cases which are unable to be matched deterministically.

2.1.3.4. Data Quality

This linkage showed a high level of agreement across shared variables, suggesting that SLC income, loans and repayment data may be used to effectively supplement data in social surveys. Further, despite an underrepresentation of minoritised ethnic groups, those from disadvantaged backgrounds and lower earners, the sociodemographic

characteristics of this linkage are in line with the characteristics of the Next Steps sample and of similar linkages. This suggests that the coverage errors previously explored have had a minimal impact on any selection/sample biases for the linked dataset.

2.1.3.5. Concepts and Methods

Another distinct benefit of this linkage is access to sensitive, more accurate data from administrative sources, which can help to overcome the recall biases sometimes found in survey data (especially for highly variable measures such as income; Prati, 2017). When coupled with detailed data on beliefs and values from the Next Steps survey, this linkage may allow for the development and exploration of new research questions, an example of which is investigated in the full paper (see Booth, Crawford, Rajah, Silverwood and Henderson, 2024).

2.1.4. Reflections and Opportunities

The authors note a few opportunities to improve the integration of administrative and survey data, the most salient being an improvement to the matching procedure. In this analysis, the matching of survey to administrative data was reliant on the respondent's NINO, with variables such as sex and date of birth being used to supplement this match (see point 4 of "*Data Integration Steps*"). To reduce the number of missed matches in the dataset, Booth et al. (2024) note that more detailed supplementary variables may be used in the deterministic matching phase or that future research may implement "fuzzy" or probabilistic matching techniques.

2.1.5. Summary

This case study illustrates some of the most salient benefits and challenges of working with integrated survey and administrative data. Please refer to the full article by Booth et al. (2024) for more information on how the representativeness of this integrated dataset was assessed, and a novel, policy relevant example of how this linkage could be used in substantive research. Please also refer to the user guide by Rihal et al. (2021) for more information on each data source, and linkage methods.

2.2. Office for National Statistics

2.2.1. Linkage Purpose

Model-based and model-assisted survey estimation is commonly implemented with the aid of population data, such as census data. When interest is in estimating complex parameters, i.e., not means and totals but non-linear statistics, analysis may require access to population-level microdata, for example, from a census. The first challenge is that accessing census microdata is difficult because of confidentiality constraints. In addition, although in the most developed countries, censuses are updated every ten years, they are much less frequent in many countries in the global south. The lack of easy access and frequent updating of population microdata motivates the need to look for alternative data sources that are freely accessible and frequently updated. Using data from alternative data sources has been explored in recent work by private firms producing survey-type estimates (https://dataforgood.facebook.com/dfg/tools). Challenges in survey design, including increasing survey non-response, means that we can no longer afford to ignore the role that alternative data sources can play in survey estimation and design.

In this case study, we focus on the use of geospatial data. Geospatial data have global coverage and are frequently updated. Advances in the availability and processing of geospatial data have created renewed interest in their use as predictors (auxiliary variables) in model-based estimation. Geospatial data have been used in small area poverty mapping in countries that lack frequent collection of census data. Despite acting only as proxies to household characteristics, results from using geospatial data are encouraging. Small area estimates using geospatial data are well correlated with design-unbiased direct estimates and with "gold standard" model-based estimates that use up-to-date census data. In addition, using geospatial data offers an approach to updating estimates in off-census years, hence improving the timeliness of the estimates.

The application we present here is motivated by recent collaborative work with the World Bank in Mozambique and in several countries in Sahel (Edochi et al., 2024). In Mozambique, we find that using geospatial data instead of census data in small area models leads to estimates that are comparable to gold-standard estimates produced with recent census microdata. Using outdated census data leads to the overestimation of poverty rates in urban areas. This is most likely caused by changes in household characteristics in urban areas during the intercensal period, which we don't capture with outdated census data. This illustrates the importance of having access to frequently updated data sources. The application in Sahel also demonstrates the added value of using geospatial data in small area estimation. The research findings from the use of geospatial data are relevant in other countries with frequently updated population data. Here, we explore how the use of geospatial data can be adapted to the UK context to assist the estimation of small area estimates of income variables. This work is relevant to the ongoing discussion in the UK about reducing reliance on census data. In collaboration with the UK Office for National Statistics, we produce research estimates of income deprivation for middle super output areas and local authority districts using data from the UK Family Resources Survey (FRS) integrated with

geospatial data. Estimates with geospatial data are compared to estimates using industry standard methods (e.g. the Empirical Best Predictor) that require access to the latest UK census microdata. The purpose of this data integration exercise is to assess the merits of using alternatives to census data to produce small area estimates, therefore reducing the reliance on census data. You can check Appendix B for more technical information on geospatial data.

2.2.2. Choice of Survey and Non-Survey Datasets

2.2.2.1. Survey Data

We use data from the UK Family Resources Survey over several years. The choice of years is such that datasets both in the middle of the intercensal period and closer to the latest census year are available.

Since the main objective of this case study is to evaluate the performance of alternative data sources against census data, we obtained two sets of variables from the UK Family Resources between 2018 and 2021. The first set includes a comprehensive range of variables that intersect with census variables. This set enables us to estimate the mean income in target areas using industry-standard small area methods. In this group of variables, we have the primary four income variables (i.e., total income, net income, equivalised income before housing costs, and equivalised income after housing costs) and numerous household characteristics (e.g., education levels and gender proportions, among others). The second set of variables includes only zonal statistics of the geospatial data. In addition to these variables, the locations of the households in the different target areas and additional administrative boundaries are available. The target small areas are MSOAs and LADs. However, since we are working with geospatial data, we also use a grid with a cell size of 100^2 meters. Therefore, we have an anonymised identifier of the cell for each household. This grid is the base for obtaining zonal statistics of the geospatial variables for all cell grids in England and Wales. The number of households in the two initial surveys (2018-2019 and 2019-2020) exceeded 13,000; however, in 2020-2021, it dropped to just 7,600 households due to COVID-19. In the following year, 2021-2022, it rose to over 12,280 households. The average sample in each MSOA varied between 5.1 and 5.3 households, except for 2020-2021, when it averaged 3.5 households. Similarly, the average number of households per LAD ranged from 39.5 to 45 households, with 24.1 for 2020-2021.

2.2.2.2. Census Data

For comparison reasons we are also working with data from the 2021 census to estimate small area estimates using industry-standard methods. The set of census variables to be considered in small area models includes those that intersect with the FRS data, primarily demographic characteristics for example, ethnicity, age groups, and education, as well as aspects related to housing and health conditions.

2.2.2.3. Geospatial and Administrative Data

Given the temporal differences between the census and survey data, publicly available geospatial variables were used to create zonal statistics, which were used as predictors in the models. In particular, ONS generated a 100-meter grid for England and Wales with

nearly 15.5 million cells. A zonal statistic at the cell level was extracted for each layer of geospatial data. Using the generated grid, ONS was able to match the location of the household with each cell to obtain zonal statistics associated with each household. Table 2 contains the geospatial variables used in the case study.

Table 2. Geospatial data sources, variables, measures and years that were obtained for the case study

Source	Variables	Measures	Years
MetOffice	Temperature, rain, wind, sun, humidity, vapour pressure, sea pressure, frost, snow,	centroid value, mean, minimum, maximum	2021, 2022
DEFRA	PM2, PM10, nitrous oxide, sulphur dioxide, benzene, ozone, flood risk	centroid value, mean	2021, 2022, 2023
Ordnance Survey	Terrain	centroid value	2023
ESRI	Night-time lights, land cover classification	centroid value	2021, 2022
Address Index	Residential addresses (approved and occupied), non-residential addresses (approved and occupied)	count	2021
OS Open Roads	Distance to main roads, distance to nearest road, road link in cell	Geodesic distance indicator	
World Cover	Distance to nearest water body, distance to nearest inland water body	Geodesic distance	
VIIRS Night-time lights 2.1	Night-time lights radiance	Median	2021,2022
Global Human Settlement Layer	Human settlement layer built-up	Centroid value	2020, 2030, 2050
WorldPop	Distance to coastline	Distance	2020

Some geospatial variables provide annual information because they depend on satellite images captured throughout the year, enabling the construction of a time series. Other geospatial variables, however, only offer measures that rely on the moment of extraction, such as distance to main roads. When possible, the measures were obtained for each survey year; otherwise, the most recent information was used.

The selection of variables was informed by similar applications in other countries. However, we also acquired access to additional administrative data because the initial results show that in the UK, models that include only geospatial covariates have low predictive power (lower than in similar applications in other countries). Administrative data are derived from multiple sources, but these are aggregated into higher administrative boundaries due to confidentiality constraints. Therefore, similar to the

geospatial data, these variables are treated as contextual variables, with the only household-level observations being the income variables from the FRS.

Table 3. Administrative data variables, measures and years are taken from multiple sources.

Source	Variables	Measures	Years
Land Registry Prices Paid	Prices paid per property type	sum, mean, median (OA, LSOA, MSOA)	2018-2022
Index of Multiple Deprivation	Income and Employment scores and ranks	Scores and ranks	2019
Planning Data	Green Belt	Indicator	2025
Consumer Data Research Centre	Census Area Classification	Class of: supergroup, group, subgroup	2021
Census	Age bands, ethnicity	Proportion	2021
DWP	Disability living allowance, employment support allowance, universal credit, pension credit	mean	2018-2022

2.2.2.4. Data Integration Steps

The integration process of the FRS, as well as the geospatial and administrative data, was performed internally by the ONS to preserve confidentiality. As mentioned before, a grid was generated to cover England and Wales using squared cells of 100 meters. Each raster of geospatial data was then projected onto the grid using different measures. For instance, if the original raster has a finer resolution than the generated grid, then we could take the average of the values inside the generated grid's cells. As mentioned in Table 2, other measures to generate geospatial-based zonal statistics include the minimum, maximum, and median, among many others. In more technical terms, we are masking the original raster into the newly generated grid.

After all the rasters are masked into the generated grid, the raster values based on the households' locations are extracted. It is important to mention that households located within the same cell will have identical values for the different geospatial covariates, but each household will have a unique income value. The result of the linkage process for the survey dataset is a table that lists each household as a row, along with its income and values on geospatial covariates, where the latter is repeated for all households in the same cell. On the other hand, the administrative data are aggregated at a higher administrative level, where households are located.

When using geospatial data, the equivalent dataset to the census dataset used by the industry-standard small area estimation methods is the generated grid for all cells in England and Wales. In Table 2, the Address Index source measures the count of households living in residential areas within the different cells. Such information is used to aggregate the cell-grid level predictions to the target geographical areas. In other words, we can use the survey dataset to estimate a model that explains the relationship

between the income and the covariates (from geospatial and administrative data), then predict the values for each cell in the generated grid and finally aggregate the result to the target area using the estimated population in each cell.

Other alternatives exist for integrating the household survey with the geospatial data. The choice mainly depends on what we know about the households' locations. One possibility is to use the household's georeferenced location and calculate the zonal statistics (i.e., mean, median, maximum, etc.) within a buffer zone around the household. This method should reveal more variability in the geospatial data, as we are likely to have different values for households in the same cells. Alternatively, if we don't have the georeferenced location of the household but instead its location within administrative boundaries, we could also obtain the zonal statistics for these boundaries. However, in this case, we will observe less variability since we are aggregating and, hence, losing information, with all households within the same administrative boundary having the same values.

2.2.3. Benefits and Challenges

Geospatial data offers significant opportunities to enhance surveys. However, there are challenges for data integration. First, the challenges arise from the source itself, and second from the information needed to establish the linkage.

Geospatial data is created through various transformations of the original satellite imagery. The initial imagery contains measurement errors. For example, we often encounter images obscured by clouds, preventing us from obtaining ground-level information. These limitations result in either missing data or inaccurate measurements. Additionally, some variables rely on human classifications of imagery (e.g., buildings), which may vary from expert to expert, introducing a degree of uncertainty. Finally, when integrating geospatial data with household surveys, it is crucial to recognise that we are observing proxy measures for household characteristics rather than the household characteristics directly.

The second challenge arises from the information we have about the location of the household. Having access to household georeferenced information represents the ideal scenario, as it enables us to determine the exact location and derive geospatial data without losing information. When we have information about the household's location within administrative boundaries, the linkage process depends on the ability to obtain zonal statistics at finer resolutions. In this case, we lose information because we aggregate the data to a higher administrative unit. The focus of current research is on how the aggregation of data impacts the precision of estimates.

2.2.4. Reflections and Opportunities

Geospatial data offers significant opportunities to utilise freely available, frequently updated data with global coverage in survey estimation. Private firms already employ alternative data sources to construct complex data pipelines and provide their clients

with new, insightful information. Meanwhile, research in countries with limited data resources has encouraged the use of geospatial data, resulting in positive research outcomes. Integrating alternative data into data-rich contexts offers significant opportunities for enhancing survey data. It also paves the way for reconsidering survey design and data collection.

2.2.5. Summary

In this case study, we have explored the integration of survey and geospatial data for model-based survey estimation. Despite the positive research findings in other countries, our current work shows that these findings are not immediately reproducible in the UK. Models with geospatial predictors estimated with UK data have lower power to predict economic deprivation than similar models estimated with data from other countries. We are currently exploring the use of alternative geospatial data and administrative data. In addition, we are assessing the performance of models for different types of areas, such as urban and rural areas. This is because initial results show that in the UK, using alternative data sources may work better for urban than rural areas.

References

1. *ADR UK - Administrative Data Research UK. Data-driven change*. (n.d.). ADR UK. https://www.adruk.org/
2. Al Baghal, T., Wenz, A., Sloan, L., & Jessop, C. (2021). Linking Twitter and survey data: asymmetry in quantity and its impact. *EPJ Data Science*, *10*(1). https://doi.org/10.1140/epjds/s13688-021-00286-7
3. Atlas of Longitudinal Datasets. (2024). *Atlas of Longitudinal Datasets*. [online] Available at: https://atlaslongitudinaldatasets.ac.uk. Accessed on 4th April 2025
4. Austin P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate behavioral research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786
5. Bach, R. L., & Wenz, A. (2020). Studying health-related internet and mobile device use using web logs and smartphone records. PLOS ONE, 15(6), e0234663. https://doi.org/10.1371/journal.pone.0234663
6. Bähr, S., Haas, G.-C., Keusch, F., Kreuter, F., & Trappmann, M. (2019). The IAB-SMART App Measurement quality in mobile geolocation sensor data, The European Survey Research Association https://www.europeansurveyresearch.org/conf2019/uploads/521/2233/29/ESRA19_baehr.pdf
7. Banks, J., Batty, G. David, Breedvelt, J., Coughlin, K., Crawford, R., Marmot, M., Nazroo, J., Oldfield, Z., Steel, N., Steptoe, A., Wood, M., Zaninotto, P. (2024). English Longitudinal Study of Ageing: Waves 0-10, 1998-2023. [data collection]. 40th Edition. UK Data Service. SN: 5050, DOI: http://doi.org/10.5255/UKDA-SN-5050-27
8. Baranyi, G., Buchanan, C.R., Conole, E.L.S. *et al.* Life-course neighbourhood deprivation and brain structure in older adults: the Lothian Birth Cohort 1936. *Mol Psychiatry* **29**, 3483–3494 (2024). https://doi.org/10.1038/s41380-024-02591-9
9. Berthoud, R., Fumagalli, L., Lynn, P., and Platt L.(2009) Design of the Understanding Society ethnic minority boost sample, *Understanding Society Working Paper 2009-02*, Colchester: University of Essex
10. Bensmann, F., Heling, L., Jünger, S., Mucha, L., Acosta, M., Goebel, J., Meinel, G., Sikder, S., Sure-Vetter, Y., & Zapilko, B. (2020). An Infrastructure for Spatial Linking of Survey Data. Data Science Journal, 19. https://doi.org/10.5334/dsj-2020-027
11. Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., & Oberski, D. L. (2022). A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*, 4(2), 388–423. https://doi.org/10.5117/CCR2022.2.002.BOES
12. Booth, C., Crawford, C., Rajah, N., Silverwood, R. J., & Henderson, M. (2024). Examining sample representativeness and data quality in the linked Next Steps survey and Student Loans Company administrative data - UCL Discovery. Ucl.ac.uk. https://discovery.ucl.ac.uk/id/eprint/10188335/1/CLS-Working-

Papers-2024-1-Sample-representativeness-and-data-quality -in-the-linked-Next-Steps-survey-and-Student-Loans-Company-administrative-data.pdf
13. Bradburn, N. M., Sudman, S., & Blair, E. (1979). *Improving interview method and questionnaire design*. Jossey-Bass.
14. Büttner, T. J. M., Sakshaug, J. W., & Vicari, B. (2021). Evaluating the Utility of Linked Administrative Data for Nonresponse Bias Adjustment in a Piggyback Longitudinal Survey. *Journal of Official Statistics*, *37*(4), 837–864. https://doi.org/10.2478/jos-2021-0037
15. Carpenter, J. R., Bartlett, J. W., Morris, T. P., Wood, A. M., Matteo Quartagno, & Kenward, M. G. (2023). Multiple Imputation and its Application. John Wiley & Sons.
16. Carrière, T.C., Boeschoten, L., Struminskaya, B. Janssen, H.L., de Schipper, N.C. & Araujo, T . Best practices for studies using digital data donation. *Qual Quant* **59** (Suppl 1), 389–412 (2025). https://doi.org/10.1007/s11135-024-01983-x
17. Cernat, A., Keusch, F., Bach, R. L., & Pankowska, P. K. (2024). Estimating Measurement Quality in Digital Trace Data and Surveys Using the MultiTrait MultiMethod Model. *Social Science Computer Review*, *0*(0). https://doi.org/10.1177/08944393241254464
18. Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(3), e2113658119. https://doi.org/10.1073/pnas.2113658119
19. Data Quality Hub. (2024, November 7). GOV.UK. https://www.gov.uk/government/organisations/government-data-quality-hub
20. Davidson, B.I., Wischerath, D., Racek, D., Parry, D.A., Godwin, E., Hinds, J., van der Linden, D., Roscoe, J.F., Ayravainen, L., Cork, A.G. (2023). Platform-controlled social media APIs threaten Open Science. Nature Human Behaviour.
21. De Waal, T & Daalmans, J. (2017). Mass imputation for Census estimation: Methodology.
22. Digital Economy Act (2017), Parliamentary Bills - UK Parliament. (2017). Retrieved from https://bills.parliament.uk/bills/1859
23. Dobson, R., Stowell, M., Warren, J., Tane, T., Ni, L., Gu, Y., McCool, J., & Whittaker, R. (2023). Use of Consumer Wearables in Health Research: Issues and Considerations. *Journal of medical Internet research*, *25*, e52444. https://doi.org/10.2196/52444
24. Edochie, I., Newhouse, D., Tzavidis, N., Schmid, T., Foster, E., Hernandez, A. L., Ouedraogo, A., Sanoh, A., & Savadogo, A. (2024). Small Area Estimation of Poverty in Four West African Countries by Integrating Survey and Geospatial Data. *Journal of Official Statistics*, *41*(1), 96-124. https://doi.org/10.1177/0282423X241284890 (Original work published 2025)

25. Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049
26. Finney, N., Nazroo, J., Becares, L., Kapadia, D., Shlomo, N., Ellingworth, D., & Begum, N. (2022). EVENS questionnaire.
27. Gasteen, A., Douglas, E., & Bell, D. (2022). *Linking longitudinal studies of ageing with administrative data First interim report ESRC Centre for Population Change Report Summary of health & retirement studies' current data linkages*. Retrieved from https://www.cpc.ac.uk/docs/2022_Linking_Longitudinal_Studies_of_Ageing_with_Admin_Data_Report1.pdf
28. GOV.UK. (2021). *Country and regional analysis: guidance*. https://www.gov.uk/government/statistics/country-and-regional-analysis-2021/country-and-regional-analysis-guidance
29. Grath-Lone, L. M., Jay, M. A., Blackburn, R., Gordon, E., Zylbersztejn, A., Wiljaars, L., & Gilbert, R. (2022). What makes administrative data "research-ready"? A systematic review and thematic analysis of published literature. *International journal of population data science*, *7*(1), 1718. https://doi.org/10.23889/ijpds.v6i1.1718
30. Grath-Lone, L. M., Jay, M. A., Blackburn, R., Gordon, E., Zylbersztejn, A., Wiljaars, L., & Gilbert, R. (2022). What makes administrative data "research-ready"? A systematic review and thematic analysis of published literature. International journal of population data science, 7(1), 1718. https://doi.org/10.23889/ijpds.v6i1.1718
31. Groves, R. M. (1989). Survey Errors and Survey Costs. In *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc. https://doi.org/10.1002/0471725277
32. Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065
33. Hamer, M., Gale, C. R., Kivimäki, M., & Batty, G. D. (2020). Overweight, obesity, and risk of hospitalization for COVID-19: A community-based cohort study of adults in the United Kingdom. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(35), 21011–21013. https://doi.org/10.1073/pnas.2011086117
34. Han, Y., & Lahiri, P. (2018). Statistical Analysis with Linked Data. *International Statistical Review*. https://doi.org/10.1111/insr.12295
35. Harron, K. (2022). Data linkage in medical research. *BMJ Medicine*, *1*(1), e000087. https://doi.org/10.1136/bmjmed-2021-000087
36. Harron, K. L., Doidge, J. C., Knight, H. E., Gilbert, R. E., Goldstein, H., Cromwell, D. A., & van der Meulen, J. H. (2017). A guide to evaluating linkage quality for the analysis of linked data. *International journal of epidemiology*, *46*(5), 1699–1710. https://doi.org/10.1093/ije/dyx177

37. Jenkins, S. P., & Rios-Avila, F. (2023). Reconciling reports: modelling employment earnings and measurement errors using linked survey and administrative data. *Journal of the Royal Statistical Society Series A: Statistics in Society*. https://doi.org/10.1093/jrsssa/qnac003
38. Jutila, O. I., Mullin, D., Vieno, M., Tomlinson, S., Taylor, A., Corley, J., Deary, I. J., Cox, S. R., Baranyi, G., Pearce, J., Luciano, M., Karlsson, I. K., & Russ, T. C. (2024). Life-course exposure to air pollution and the risk of dementia in the Lothian Birth Cohort 1936. *Environmental epidemiology (Philadelphia, Pa.)*, *9*(1), e355. https://doi.org/10.1097/EE9.0000000000000355
39. Keusch, F., Struminskaya, B., Eckman, S. & Guyer, H. (2024). Data Collection with Wearables, Apps, and Sensors. Bookdown.org. https://bookdown.org/wasbook_feedback/was/
40. Lynn, P., Nandi, A., Parutis, V. & Platt, L. (2018) Design and implementation of a high quality probability sample of immigrants and ethnic minorities: Lessons learnt' *Demographic Research*, 38, 513-548. https://doi.org/10.4054/DemRes.2018.38.21.
41. Lyons, R.A., Jones, K.H., John, G. *et al*. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* **9**, 3 (2009). https://doi.org/10.1186/1472-6947-9-3
42. Mansournia, M., & Altman, D. (2016). Inverse probability weighting. BMJ, 352, i189.
43. Marcello D'Orazio, Marco Di Zio, & Scanu, M. (2006). *Statistical Matching*. https://doi.org/10.1002/047002355
44. Meinfelder, F., & Schaller, J. (2022). Data Fusion for Joining Income and Consumtion Information using Different Donor-Recipient Distance Metrics. *Journal of Official Statistics*, *38*(2), 509-532. https://doi.org/10.2478/jos-2022-0024 (Original work published 2022)
45. Merkouris, T., Smith, P. A., & Fallows, A. (2023). Combining National Surveys with Composite Calibration to Improve the Precision of Estimates from the United Kingdom's Living Costs and Food Survey. *Journal of Survey Statistics and Methodology*, *11*(3), 713–741. https://doi.org/10.1093/jssam/smad001
46. Mooney, S. J., & Garber, M. D. (2019). Sampling and Sampling Frames in Big Data Epidemiology. *Current epidemiology reports*, *6*(1), 14–22. https://doi.org/10.1007/s40471-019-0179-y
47. Newhouse, D. (2023). Small Area Estimation of Poverty and Wealth Using Geospatial Data: What have We Learned So Far? Calcutta Statistical Association Bulletin, 76(1), 7–32. https://doi.org/10.1177/00080683231198591
48. Office for National Statistics (2021). Geospatial methods for Small Area Population Estimates: proof of concept - Office for National Statistics. Ons.gov.uk. https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/geospatialmethodsforsmallareapopulationestimatesproofofconcept

49. Office for National Statistics. (2021). Crime Survey for England and Wales. [data series]. 3rd Release. UK Data Service. SN: 200009, DOI: http://doi.org/10.5255/UKDA-Series-200009
50. Office for National Statistics. (2024a). *Labour Force Survey*. [data series]. *11th Release*. UK Data Service. SN: 2000026, DOI: http://doi.org/10.5255/UKDA-Series-2000026
51. Office for National Statistics. (2024b). Annual Population Survey. [data series]. 8th Release. UK Data Service. SN: 200002, DOI: http://doi.org/10.5255/UKDA-Series-200002
52. Office for National Statistics. (2025). Annual Survey of Hours and Earnings, 1997-2024: Secure Access. [data collection]. 26th Edition. UK Data Service. SN: 6689, DOI: http://doi.org/10.5255/UKDA-SN-6689-25
53. Openshaw, S. (1984). Ecological Fallacies and the Analysis of Areal Census Data. *Environment and Planning A*, *16*(1), 17-31. https://doi.org/10.1068/a160017 (Original work published 1984)
54. Prati, A. (2017). Hedonic recall bias. Why you should not ask people how much they earn. *Journal of Economic Behavior & Organization, 143,* 78–97. https://doi.org/10.1016/j.jebo.2017.09.002
55. Rajah, N., Calderwood, L., De Stavola, B. L., Harron, K., Ploubidis, G. B., & Silverwood, R. J. (2023). Using linked administrative data to aid the handling of non-response and restore sample representativeness in cohort studies: the 1958 national child development study and hospital episode statistics data. *BMC medical research methodology*, *23*(1), 266. https://doi.org/10.1186/s12874-023-02099-w
56. Rao, J. N. K. & Molina, I. (2015). *Small area estimation*. John Wiley & Sons, Inc.
57. Rao, J. N. K. (2003). *Small Area Estimation*. https://doi.org/10.1002/0471722189
58. Research Data Scotland (2025). Research Data Scotland, Research Access Service. https://www.researchdata.scot/
59. Rihal, S., Gomes, D., Henderson, M. (2021) Next Steps: Linked Student Loans Company administrative datasets User Guide (2nd edition, October 2021). London: UCL Centre for Longitudinal Studies.
60. Rihal, S., Gomes, D., Henderson, M. (2021) Next Steps: Linked Student Loans Company administrative datasets User Guide (2nd edition, October 2021). London: UCL Centre for Longitudinal Studies.
61. Robert M. Groves, Steven G. Heeringa, Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs, *Journal of the Royal Statistical Society Series A: Statistics in Society*, Volume 169, Issue 3, July 2006, Pages 439–457, https://doi.org/10.1111/j.1467-985X.2006.00423.x
62. Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.2307/2335942
63. Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity

Score. *The American Statistician*, *39*(1), 33–38. https://doi.org/10.1080/00031305.1985.10479383
64. Sakshaug J. W. (2022). Reducing Nonresponse and Data Linkage Consent Bias in Large-Scale Panel Surveys. *Forum for health economics & policy*, *25*(1-2), 41–55. https://doi.org/10.1515/fhep-2021-0060
65. Sakshaug, J. W., & Antoni, M. (2018). Evaluating the Utility of Indirectly Linked Federal Administrative Records for Nonresponse Bias Adjustment. *Journal of Survey Statistics and Methodology*, *7*(2), 227–249. https://doi.org/10.1093/jssam/smy009
66. Sakshaug, J. W., Wiśniowski, A., Ruiz, D. A. P., & Blom, A. G. (2019). Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach. Journal of Official Statistics, 35(3), 653–681. https://doi.org/10.2478/jos-2019-0027
67. Schouten, B., Peytchev, A., & Wagner, J. (2017). Adaptive Survey Design (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315153964
68. Schulte, K. (2022). 'Real-time' air quality channels: A technology review of emerging environmental alert systems. *Big Data & Society*, *9*(1). https://doi.org/10.1177/20539517221101346 (Original work published 2022)
69. Shlomo, N., & Goldstein, H. (2015). Editorial: Big data in social research. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *178*(4), 787–790. https://doi.org/10.1111/rssa.12144
70. Silber, I., R.C. Jackson, A.M. Fridlind, A.S. Ackerman, S. Collis, J. Verlinde, and J. Ding, 2022: The Earth Model Column Collaboratory (EMC^2) v1.1: An open-source ground-based lidar and radar instrument simulator and subcolumn generator for large-scale models. *Geosci. Model Dev.*, **15**, no. 2, 901-927, doi:10.5194/gmd-15-901-2022.
71. Silverwood, R. J., Baranyi, G., Calderwood, L., Ploubidis, G. B., Harron, K., De Stavola, B. L., (2024). Using Linked Cohort Data to Help Address Residual Confounding in Analyses of Population Administrative Data. Society for Longitudinal and Lifecourse Studies Annual Meeting, University of Essex. https://hubble-live-assets.s3.eu-west-1.amazonaws.com/slls/file_asset/file/1118/2024_SLLS_CONFERENCE_ABSTRACT_BOOK.pdf
72. Skinner, C. (2018). Issues and Challenges in Census Taking. *Annual Review of Statistics and Its Application*, *5*(1), 49–63. https://doi.org/10.1146/annurev-statistics-041715-033713
73. Sladka, V. & Lynn, P. (2025) Targeted procedures for tackling survey non-response: Evidence Review. Survey Futures Report No. 5. https://surveyfutures.net/reports/
74. Smart Data Research UK (2025). Retrieved from https://sdds.ac.uk/assets/SDDS_brochure.pdf
75. Statistics Canada. 2023. Census Profile. 2021 Census of Population. Statistics Canada Catalogue number 98-316-X2021001. Ottawa. Released November 15,

2023. https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=E (accessed April 4, 2025)
76. Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLOS ONE*, *10*(2), e0107042. https://doi.org/10.1371/journal.pone.0107042
77. Stopforth, S., Gayle, V., & Boeren, E. (2020). Parental social class and school GCSE outcomes: two decades of evidence from UK household panel surveys. *Contemporary Social Science*, *16*(3), 309–324. https://doi.org/10.1080/21582041.2020.1792967
78. Tourangeau, R., Brick, J. M., Lohr, S., & Li, J. (2017). Adaptive and responsive survey designs: a review and assessment. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *180*(1), 203–223. http://www.jstor.org/stable/44682558
79. *UK Longitudinal Linkage Collaboration*. (2025). UK LLC. https://ukllc.ac.uk/
80. University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2025). Next Steps: Sweeps 1-9, 2004-2023. [data collection]. 18th Edition. UK Data Service. SN: 5545, DOI: http://doi.org/10.5255/UKDA-SN-5545-10
81. University College London, UCL Social Research Institute, Centre for Longitudinal Studies. (2025). *1970 British Cohort Study*. [data series]. *12th Release*. UK Data Service. SN: 200001, DOI: http://doi.org/10.5255/UKDA-Series-200001
82. University of Essex, Institute for Social and Economic Research. (2024). *Understanding Society: Innovation Panel Twitter Study, 2007-2023*. [data collection]. UK Data Service. SN: 9208, DOI: http://doi.org/10.5255/UKDA-SN-9208-1
83. University of Essex, Institute for Social and Economic Research. (2024). Understanding Society: Waves 1-14, 2009-2023 and Harmonised BHPS: Waves 1-18, 1991-2009. [data collection]. 19th Edition. UK Data Service. SN: 6614, http://doi.org/10.5255/UKDA-SN-6614-20.
84. Verian (2024) Participation Survey 2023 to 2024: Annual technical report. https://www.gov.uk/government/statistics/participation-survey-2023-24-annual-publication
85. Vermeulen, W. and F. Gutierrez Amaros (2024), "How well do online job postings match national sources in European countries?: Benchmarking Lightcast data against statistical and labour agency sources across regions, sectors and occupation", *OECD Local Economic and Employment Development (LEED) Papers*, No. 2024/02, OECD Publishing, Paris, https://doi.org/10.1787/e1026d81-en.
86. Vieno, M., Heal, M. R., Williams, M. L., Carnell, E. J., Nemitz, E., Stedman, J. R., & Reis, S. (2016). The sensitivities of emissions reductions for the mitigation of UK

$PM_{2.5}$. *Atmospheric Chemistry and Physics*, *16*(1), 265–276. https://doi.org/10.5194/acp-16-265-2016

87. Vitiis, C. D., Stefano Falorsi, Alessio Guandalini, Inglese, F., Righi, P., & Terribili, M. D. (2024). Adaptive sampling design for the Italian social sample surveys: an application on the population census. *METRON*, *82*(1), 19–35. https://doi.org/10.1007/s40300-023-00262-3
88. Wenemark, M., Frisman, G. H., Svensson, T., & Kristenson, M. (2010). Respondent satisfaction and respondent burden among differently motivated participants in a health-related survey. *Field Methods, 22*(4), 378–390. https://doi.org/10.1177/1525822X10376704
89. Wenz, A., Baghal, T., Sloan, L., & Jessop, C. (2021). Twitter use during the Covid-19 pandemic: Results from two UK studies linking Twitter and survey data. Understanding Society Survey Methods Conference, (virtual conference), September 27th to September 30th, 2021
90. Wenz, A., Jäckle, A., Burton, J., Couper, M., & Read, B. (2023). *Quality of expenditure data collected with a mobile receipt scanning app in a probability household panel Understanding Society Working Paper Series*. https://www.understandingsociety.ac.uk/wp-content/uploads/working-papers/2023-02.pdf
91. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3, 160018. https://doi.org/10.1038/sdata.2016.18
92. Williams, J. (2024) DCMS Participation Survey: selected design features. Paper presented at the 3rd Survey Futures Survey Practice Forum. https://surveyfutures.net/events/2024/01/17/survey-practice-forum-3/
93. 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 (General)*, *150*(3), 230–240. https://doi.org/10.2307/2981474
94. Wiśniowski, A., Sakshaug, J. W., Ruiz, D. A. P., & Blom, A. G. (2020). Integrating Probability and Nonprobability Samples for Survey Inference, *Journal of Survey Statistics and Methodology*, Volume 8, Issue 1, https://doi.org/10.1093/jssam/smz051
95. Worldpop. (2025). *WorldPop*. https://www.worldpop.org/
96. Zhang, S., & Wagner, J. (2022). The Additional Effects of Adaptive Survey Design Beyond Post-Survey Adjustment: An Experimental Evaluation. *Sociological Methods & Research*, 004912412210995. https://doi.org/10.1177/00491241221099550

Appendix

Appendix Item A: List of UK survey to-non-survey data integration data sources.

Name	Acronym	URL	Number of participants at first data collection	Profile Paper DOI	Data access	Year of first data collection	Data linkage types	Country
Our Future	LSYPE2	https://www.gov.uk/government/publications/longitudinal-study-of-young-people-in-england-cohort-2-wave-1	13,100 (participants)	https://www.gov.uk/government/publications/longitudinal-study-of-young-people-in-england-cohort-2-wave-3	Accessible via study website, data sharing platform etc.	2013	Education data, Healthcare data, Tax, income & benefit data, Police & judicial system data, Other government data	England, United Kingdom of Great Britain and Northern Ireland
Avon Longitudinal Study of Parents and Children	ALSPAC	https://www.bristol.ac.uk/alspac/	14,541 (mothers),14,062 (children)	https://doi.org/10.1093/ije/dys064	Contact study team for access	1990	Education data, Healthcare data, Social media & technology use data, Geographic, spatial & environmental data	United Kingdom of Great Britain and Northern Ireland, England
TwinsUK: The UK Adult Twin Registry	TwinsUK	https://twinsuk.ac.uk/	> 16,000 (twins)	https://doi.org/10.1017/thg.2019.65	Contact study team for access	1992	Healthcare data, Mortality data, Education data, Geographic, spatial & environmental data	United Kingdom of Great Britain and Northern Ireland, England, Northern Ireland, Scotland, Wales
Northern Ireland Longitudinal Study	NILS	https://nils.ac.uk/	508,000 (participants)	https://doi.org/10.1093/ije/dyq271	Contact study team for access	1981 (linked Census data),2006 (NILS baseline)	Census data, Mortality data, Medical birth registry, Geographic, spatial & environmental data, Other government data	Northern Ireland, United Kingdom of Great Britain and Northern Ireland
Born in Bradford	BiB	https://borninbradford.nhs.uk/	13,818 (births),3,448 (partners),12,453 (mothers)	https://doi.org/10.1093/ije/dys112	Contact study team for access	2007	Healthcare data, Education data, Geographic, spatial & environmental data	United Kingdom of Great Britain and Northern Ireland, England
Lothian Birth Cohort of 1921	LBC1921	https://www.ed.ac.uk/lothian-birth-cohorts/	550 (participants)	https://doi.org/10.1093/ije/dyy022	Contact study team for access	1932 (Scottish Mental Survey 1932),1999 (LBC1921)	Healthcare data	Scotland, United Kingdom of Great Britain and Northern Ireland
UK Biobank	UKB	https://www.ukbiobank.ac.uk	500,000	https://doi.org/10.1371/journal.pmed.1001779	Accessible via study website, data sharing platform etc.	2006	Healthcare data, Geographic, spatial & environmental data, Mortality data	United Kingdom of Great Britain and Northern Ireland, England, Scotland, Wales

Airwave Health Monitoring Study		https://police-health.org.uk/	53,228 (participants)	doi.org/10.1016/j.envres.2014.07.025	Contact study team for access	2006	Healthcare data	England, Scotland, Wales, United Kingdom of Great Britain and Northern Ireland
Twins Early Development Study	TEDS	https://www.teds.ac.uk/	13,694 (twin pairs)	https://acamh.onlinelibrary.wiley.com/doi/full/10.1002/jcv2.12154	Contact study team for access	1995	Healthcare data, Social media & technology use data, Geographic, spatial & environmental data, Education data	United Kingdom of Great Britain and Northern Ireland, Wales, England
Northern Ireland Cohort for the Longitudinal Study of Ageing	NICOLA	https://www.qub.ac.uk/sites/NICOLA/	8,478 (participants)	https://doi.org/10.1093/ije/dyad026	Contact study team for access	2014	Mortality data, Healthcare data	Northern Ireland, United Kingdom of Great Britain and Northern Ireland
Understanding Society, The UK Household Longitudinal Study	UKHLS	https://www.understandingsociety.ac.uk/	39,802 (households)	http://dx.doi.org/10.14301/llcs.v3i1.159	Accessible via study website, data sharing platform etc.	2009 (UKHLS households),1991 (BHPS households)	Education data, Geographic, spatial & environmental data, Healthcare data, Tax, income & benefit data	United Kingdom of Great Britain and Northern Ireland, England, Wales, Scotland, Northern Ireland
Lothian Birth Cohort of 1936	LBC1936	https://www.ed.ac.uk/lothian-birth-cohorts/	1,091 (participants)	https://doi.org/10.1093/ije/dyy022	Contact study team for access	1947 (Scottish Mental Survey 1947),2004 (LBC1936)	Healthcare data	Scotland, United Kingdom of Great Britain and Northern Ireland
1970 British Cohort Study	BCS70	https://cls.ucl.ac.uk/cls-studies/1970-british-cohort-study/	17,198 (participants)	https://doi.org/10.1093/ije/dyac148	Accessible via study website, data sharing platform etc.	1970	Healthcare data, Tax, income & benefit data	United Kingdom of Great Britain and Northern Ireland, England, Scotland, Wales, Northern Ireland, Jersey, Guernsey, Isle of Man
#BeeWell		https://beewellprogramme.org/	20,241 (participants)	https://doi.org/10.1186/s13034-023-00687-8	Contact study team for access	2019	Education data, Geographic, spatial & environmental data, Tax, income & benefit data, Other government data	United Kingdom of Great Britain and Northern Ireland, England
Aberdeen 1936 Birth Cohort Study	ABC1936	https://www.abdn.ac.uk/achds/environment/birth-cohorts/1936-birth-cohort-316.php	498 (participants)	https://doi.org/10.1016/j.maturitas.2011.05.010	Contact study team for access	1947 (Scottish Mental Survey),1999 (ABC1936)	Education data	Scotland, United Kingdom of Great Britain and Northern Ireland
Scottish Longitudinal Study	SLS	https://sls.lscs.ac.uk/about/	270,385 (participants)	https://doi.org/10.1093/ije/dyn087	Contact study team for access	1991	Census data, Education data, Mortality data, Medical birth registry, Mortality data, Geographic, spatial & environmental data, Healthcare data, Other government data	Scotland, United Kingdom of Great Britain and Northern Ireland

National Survey of Health and Development	NSHD	https://nshd.mrc.ac.uk/	5,362 (participants)	https://doi.org/10.1093/ije/dyi201	Accessible via study website, data sharing platform etc.	1946	Mortality data, Healthcare data, Geographic, spatial & environmental data	England, Scotland, Wales, United Kingdom of Great Britain and Northern Ireland
Healthy Ageing In Scotland	HAGIS	https://www.hagis.scot/	1,000 (participants at pilot)	https://doi.org/10.1136/bmjopen-2017-018802	Accessible via study website, data sharing platform etc.	2017 (pilot study)	Healthcare data, Tax, income & benefit data, Education data, Social care data	Scotland, United Kingdom of Great Britain and Northern Ireland
English Longitudinal Study of Ageing	ELSA	https://www.elsa-project.ac.uk/	11,391 (participants),708 (partners)	doi.org/10.1093/ije/dys168	Accessible via study website, data sharing platform etc.	2002	Mortality data, Healthcare data, Tax, income & benefit data, Healthcare data	England, United Kingdom of Great Britain and Northern Ireland
Growing Up in Scotland	GUS	https://growingupinscotland.org.uk/	5,217 (participants)		Accessible via study website, data sharing platform etc.	2005	Education data, Healthcare data	Scotland, United Kingdom of Great Britain and Northern Ireland
Next Steps		https://cls.ucl.ac.uk/cls-studies/next-steps/	15,770 (participants)	https://doi.org/10.5334/ohd.16	Accessible via study website, data sharing platform etc.	2004	Education data, Healthcare data, Tax, income & benefit data, Other government data	United Kingdom of Great Britain and Northern Ireland, England
National Child Development Study	NCDS	https://cls.ucl.ac.uk/cls-studies/1958-national-child-development-study/	17,415 (participants)	https://doi.org/10.1093/ije/dyi183	Accessible via study website, data sharing platform etc.	1958	Healthcare data, Mortality data, Other government data	United Kingdom of Great Britain and Northern Ireland, England, Scotland, Wales, Isle of Man, Jersey, Guernsey
Aberdeen 1921 Birth Cohort Study	ABC1921	https://www.abdn.ac.uk/achds/environment/birth-cohorts/1921-birth-cohort-314.php#panel310	275 (participants)	https://doi.org/10.1016/j.maturitas.2011.05.010	Contact study team for access	1932 (Scottish Mental Survey),1997 (ABC1921)	Education data	Scotland, United Kingdom of Great Britain and Northern Ireland
Hertfordshire Cohort Study: The 1930's Cohort	HCS	https://generic.wordpress.soton.ac.uk/herts/	3,225 (participants)	https://doi.org/10.1093/ije/dyi127	Contact study team for access	1998 (HCS baseline),1931 (birth records)	Healthcare data, Mortality data	England, United Kingdom of Great Britain and Northern Ireland
Whitehall II		https://www.ucl.ac.uk/epidemiology-health-care/research/epidemiology-and-public-health/research/whitehall-ii	10,308 (participants)	https://doi.org/10.1093/ije/dyh372	Accessible via study website, data sharing platform etc.	1985	Education data, Healthcare data, Mortality data, Geographic, spatial & environmental data	England, United Kingdom of Great Britain and Northern Ireland
Millennium Cohort Study (UK)	MCS	https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/	18,818 (children)	https://doi.org/10.1093/ije/dyu001	Accessible via study website, data sharing platform etc.	2001	Education data, Healthcare data, Medical birth registry, Mortality data, Tax, income & benefit data, Geographic, spatial & environmental data	United Kingdom of Great Britain and Northern Ireland, England, Scotland, Wales, Northern Ireland, Isle of Man, Jersey, Guernsey

Appendix Item B: Further details on geospatial-to-survey data integration methodologies.

Geospatial data includes location information in the form of coordinates, allowing observations to be mapped to specific geographic locations. This type of data can be linked to various geometries such as points, lines, polygons, and grids. In the context of social science research, geospatial data is used to enhance survey data with contextual data. For example, in recent work, geospatial data has been used to estimate poverty measures (Edochie et al., 2024).

Geospatial data is often publicly available and becoming more accessible to researchers. The Google Earth Engine (GEE) provides a comprehensive repository of geospatial data. Other repositories are at NASA and Copernicus, part of the European Union's space programme. In addition, there are other sources of data, such as previous studies that compile geospatial data. For example, datasets produced by WorldPop (2024) at the University of Southampton use geospatial data to output gridded population estimates globally.

The curation of remote sensing data generally goes through multiple processes. Figure 5 illustrates a simplified process for obtaining geospatial data from different repositories. The image from the satellite is processed to decompose the different bands of light depending on the researchers' main interest. Usually, the next step is to train a model or algorithm to predict the target variable. The final product is an amalgamation of processes that reflects a summary of information (zonal statistics) for each cell, e.g., in a grid.

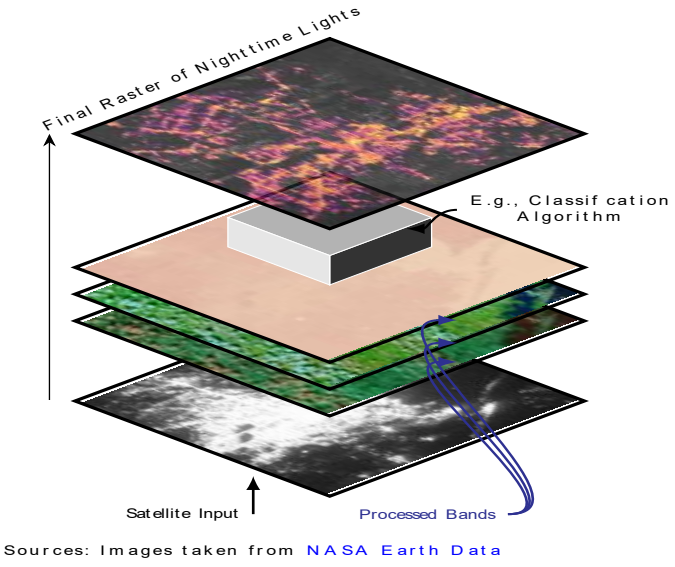


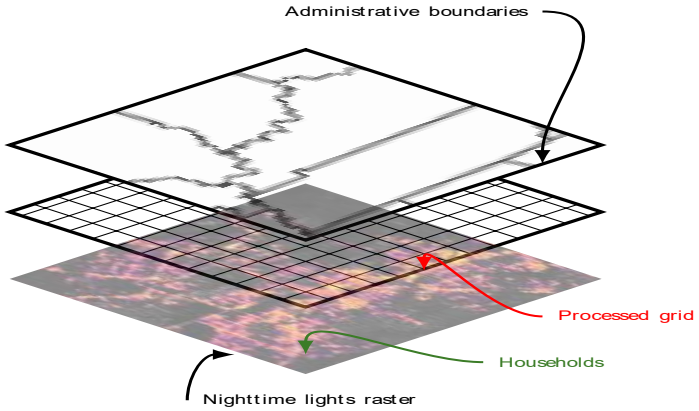
Figure 5: Illustration of the process to generate geospatial data

Linking geospatial data with survey data presents several challenges. First, legal restrictions, particularly concerning the sensitive and potentially disclosive level of visibility that geographical data can provide, may limit the accessibility and sharing of detailed geospatial information due to the risk of re-identifying survey participants.

Additionally, data sources may be incomplete or temporally inconsistent across different regions, and so further, integrating geospatial data with other data sources (e.g., administrative data and survey data) requires careful handling to avoid mismatches in spatial resolution and alignment (Bensmann et al., 2020).

Modern technology allows sample surveys to obtain georeferenced coordinates of the participating units (e.g., households). Because of confidentiality constraints, national statistical offices (NSOs) do not make the households' geolocations publicly available. Instead, NSOs typically use a process of aggregation under which households are placed at the centre of a grid cell or administrative unit. The grid cell within which a household lies is then used to link information collected through the sample survey to grid cell-level zonal statistics of geospatial variables.

Figure 6 illustrates a case in which the analyst has the location of each household. In this example, the analyst has a raster of the nighttime lights. The green dots denote households' true locations. The analyst can process the original raster to a set of grid cells (processed grid) or at the original resolution. Using the processed grid and the location of the household, the analyst can obtain the values of zonal statistics for the night lights for cells in the processed grid. Households located in the same cell will have the same values of zonal statistics.



Sources: Images taken from NASA Earth Data

Figure 6: Illustration of location of households within grid-cells and administrative boundaries

Using the exact household's location is possible when the analyst has no confidentiality constraints. However, it is more difficult for a secondary analyst to access these detailed data in a census. Therefore, a more realistic data access scenario is one where the analyst has access to the georeferenced data from the survey but lacks access to exact location data in the census. In this case, geospatial covariates can be aggregated at some administrative level (i.e. higher than the grid level) if the secondary analyst has access to the location of the households at this level of geography. For example, in Figure 6, the analyst could use the administrative boundaries of the lowest possible level available, such as enumeration areas (EAs). Ideally, the analyst would also have access to the location of the households in the EAs and the number of households in

each census EA. Hence, the analyst can still use the geospatial covariates even with higher administrative levels if access to the exact household locations in the survey and census data are not available. If aggregation to a higher administrative level is used, geospatial zonal statistics must be produced at this level. This can be done by using weighted summary statistics, where the weights are defined by the fraction of the higher administrative unit each cell covers. Figure 7 illustrates the aggregation of the nighttime lights variable in the Ka Mpfumo district in Mozambique. It is important to note that the cells are often smaller than an EA, which translates into a loss of information due to the aggregation process. As the size of the administrative boundary increases, the loss of information is greater, and more households will be assigned the same value of the geospatial variable.

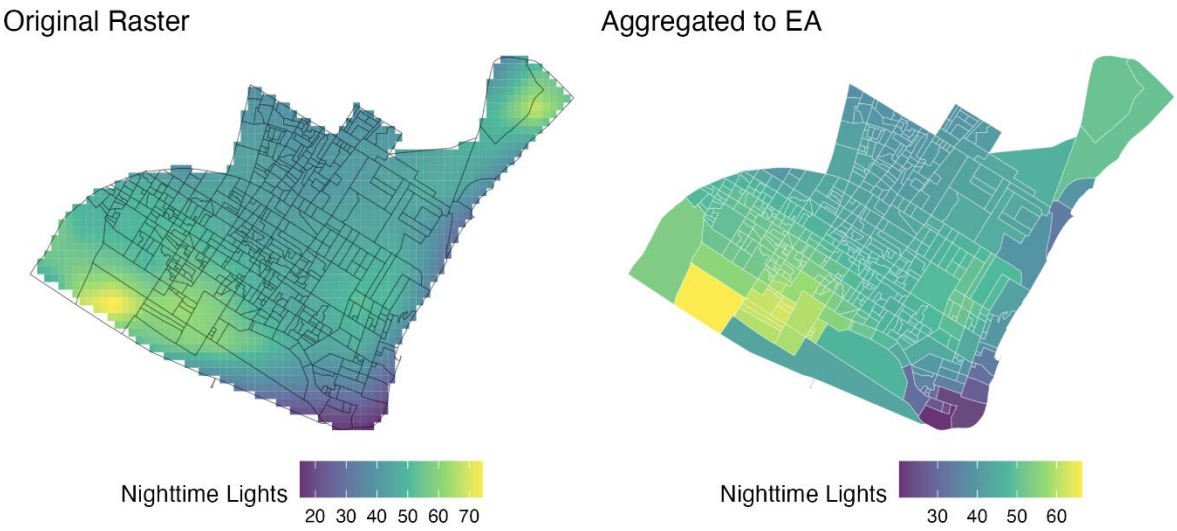


Figure 7: Nighttime lights - Original raster and aggregated zonal statistics in Ka Mpfumo

One recurrent problem with the use of geospatial data is cloud coverage, which introduces noise to the algorithms and may generate some data with errors. There are additional problems in other cases; for example, the South Atlantic Anomaly impedes satellites and spacecraft from obtaining correct information when passing through this area.[1]. In some cases, geospatial data are affected by missing data. This was the case, for example, in a recent application in Mozambique. Figure 8 shows the missing data in the province of Cabo Delgado (grey area on the left-hand side of the map). This is because the Google Buildings V3 variable contains missing information for a large portion of this province. However, the Microsoft Building Footprint 2023 dataset contains information for this province. The map on the right-hand side shows the Microsoft Footprints variable with coverage over most EAs in Cabo Delgado. This demonstrates the importance of combining information from several sources of geospatial data to mitigate issues with missing data.

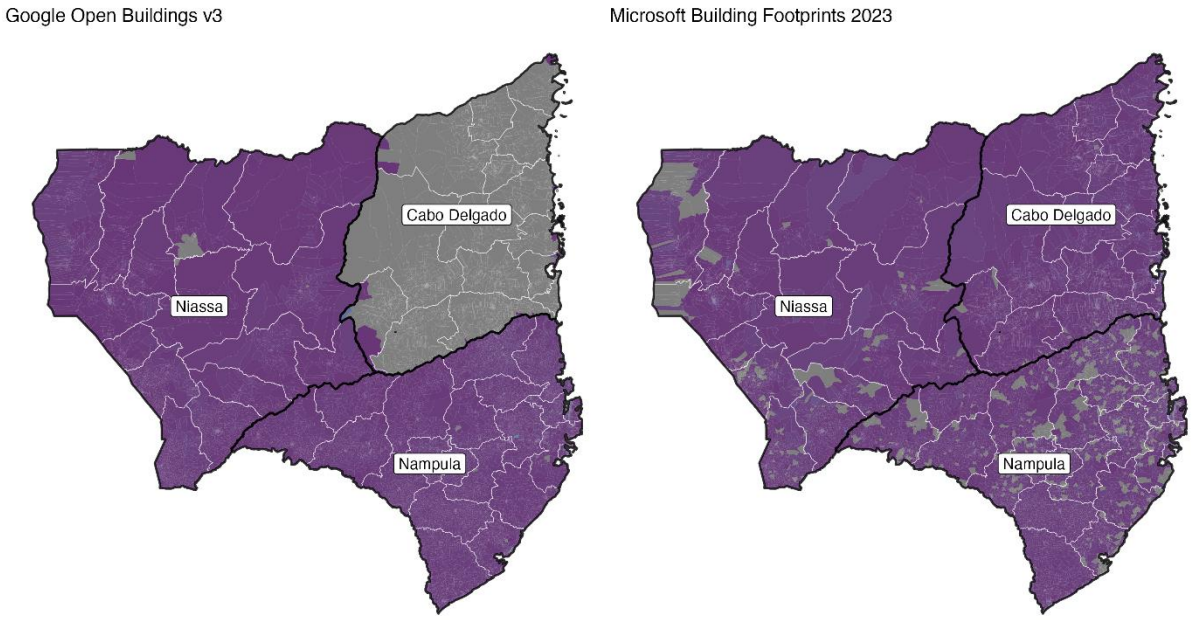
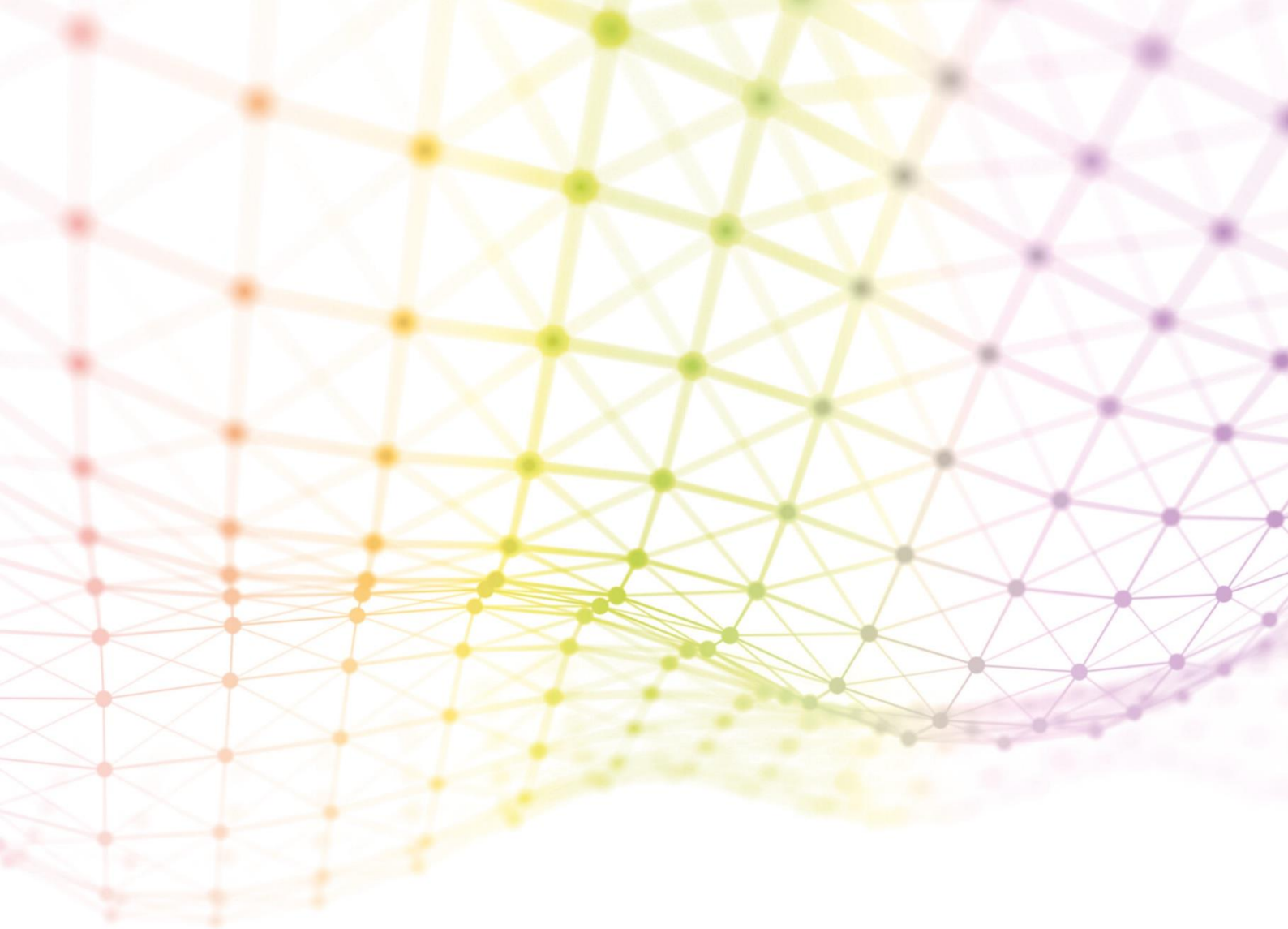


Figure 8: Map of Cabo Delgado, Nampula, and Niassa. Google Buildings V3 and Microsoft Footprints data.



University of Essex

University of Southampton

UKRI Economic and Social Research Council

NCRM National Centre for Research Methods

Office for National Statistics

UCL

Warwick The University of Warwick

Manchester 1824 The University of Manchester

City University of London Est 1894

National Centre for Social Research

LSE The London School of Economics and Political Science

Ulster University

Unil UNIL | Université de Lausanne

Ipsos

verian Formerly Kantar Public

www.surveyfutures.net