MARP Paper January 2023

A linkage project between the 2021 Census and Census Coverage Survey to the Demographic Index: Rationale and Research Questions

Authors: Elizabeth Pereira | Jack Rodgers | Emma Hand

Contributors: Daria Tkacz | Emilie Woodhall | Ffion Jones | Sally Mylles | Ayesha Barnes

Social Statistics Transformation, Analysis and Research

Purpose of paper

This paper will outline the planned work for the analysis of the linked data between the 2021 Census and Census Coverage Survey (CCS) and the Demographic Index (DI). We ask the panel to offer guidance and advice on the design principles and analysis plans.

This paper describes:

I. Background on Social Statistics Transformation Analysis and Research (SSTAR) programme
II. Background on data sources
III. Rationale for linkage
IV. Research questions and what they aim to inform
V. Design principles
VI. Assumptions
VII. Future work

Panel Ask

We ask Panel members to:

- Comment on the proposed research questions and their aims, in particular, are they appropriate, and advise of any other approaches or considerations we need to make in the design
- Review the current research questions and advise if there are more research questions we should consider
- Advise on the priority of the research questions

1. Background on Social Statistics Transformation Analysis and Research (SSTAR) Programme

ONS are transforming the way we produce population, migration and social statistics to:

- Produce more timely and regular population totals by age and sex at a local level
- Provide more timely and regular small-area multivariate outputs each year (ultimate aim: more topics than a census can provide once a decade)

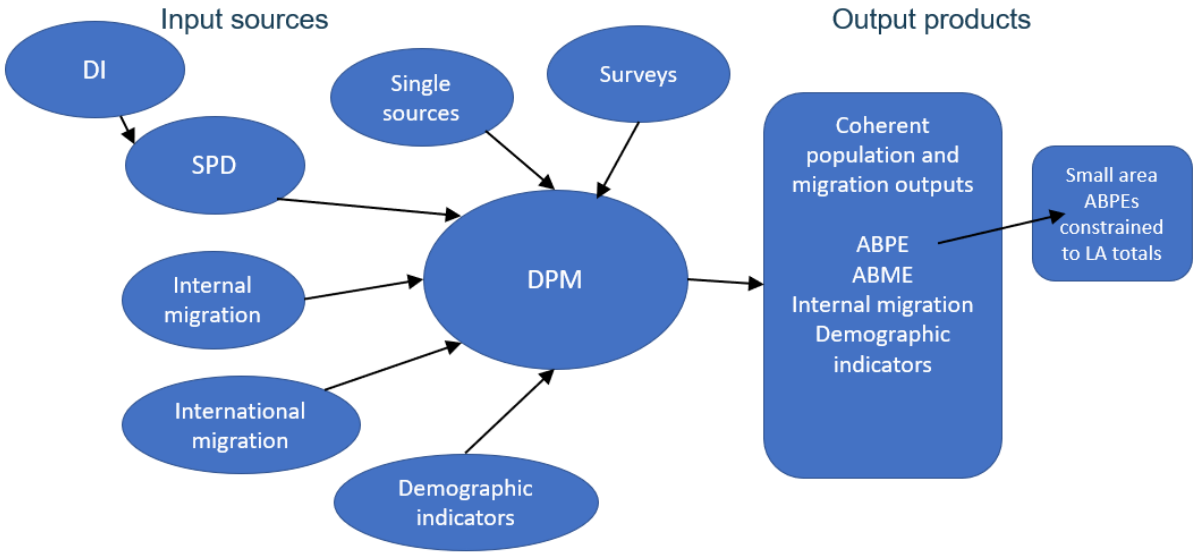
In 2023 the National Statistician will make a recommendation on the future of the census. The transformed system will continue to iteratively build on what ONS has achieved.

ONS has created Statistical Population Datasets[1] (SPDs). These use rules to combine administrative data (including the Demographic Index (DI)) to produce a record level dataset that attempts to include the usually resident population. Multiple versions have been produced using different rules for inclusion and priority based on the admin data source (see Appendix 1). Another version of the SPD is in development based on what has been learnt from the current versions, ongoing research and the potential from new data sources. We will develop the rules further based on findings from the linkage discussed in this paper.

ONS is also developing a Dynamic Population Model (DPM). This modelling framework will allow us to produce best-available, coherent and timely population statistics which will be official population estimates based on admin data (Blackwell 2021). Figure 1 shows the inputs into the DPM – including SPDs – and the outputs which will be generated.

The analysis of the linkage discussed in this paper will be used to inform the datasets used for the DPM and the rules for the latest set of SPDs, as well as the 2023 recommendation.

Figure 1: Dynamic Population Model (DPM) inputs and outputs



2. Background about Census 2021 and DI

[1] There was a time when SPDs were referred to as Admin Based Population Estimates (ABPEs) but they have been re-named to SPDs to avoid confusion with the planned output statistics.

The Demographic Index
The Demographic Index (DI) will be linked to the 2021 Census. The DI is a composite linked dataset, produced by linking together datasets from 2016 to 2021:

- Personal Demographic Service (PDS): National Health Service
- Higher Education Statistics Agency (HESA): Student enrolments for tertiary education – university students
- English School Census (ESC), and Welsh School Census (WSC) – school students not in private education
- Customer Information System (CIS): data from the Department for Work and Pensions, covering Pay As You Earn and benefits data (does not cover self-employed)
- The Births Register
- Individual Learner Record

The DI contains a single entity for what it believes to be a single individual, which may contain one or more records from any of the five datasets. Each entity is stored as an 'ONS ID', which can be used to cluster together all records that belong to that individual.

The DI does not seek to resolve ONS ID clusters into a single point-in-time record. Rather, the DI seeks to bring data together as efficiently as possible, and with the highest quality linkage as possible. However, it is difficult to define what "good quality" is for the DI, let alone measure it. The DI is a new type of data, which we have begun referring to as "composite", since it is the patchwork result of extensive linkage across both data sources and time.

We will be using Version 2.0 of the DI and keeping the most recent year of data available for each source: 2021 extracts of PDS, ESC and WSC and 2020 extracts of CIS and HESA. Four of the datasets are available to ONS in-the-clear, however the CIS dataset has all personal identifiable information (PII) hashed and has security restrictions in place that prevents other non-CIS PII being attached to it.

To be able to incorporate up to date information for students, especially new starters, we have developed a strategy to incorporate 2021 HESA data in our analysis. The HESA 2021 data could not be used in the linkage as it was not available in version 2.0 of the DI. An updated version of the DI (2.1) does include HESA 2021. We reviewed the records in DI 2.1 that were included in clusters with HESA 2021 and used in DI 2.0. This means that we can use these records as a bridge to connect clusters in DI 2.1 to clusters in version 2.0. HESA 2021 data can then be brought into our analysis via DI 2.0 ONS IDs which were successfully linked to the CC. It is important to note, HESA 2021 is therefore only available for those present in DI 2.0, meaning first year students not on another administrative source (i.e. HESA 2021 is their only source) will persist to be missing.

Census 2021
The 2021 Census is the England and Wales population-wide survey conducted in March 2021. The census is the largest statistical exercise that ONS undertakes, producing statistics that inform all areas of public life and underpin social and economic policy. Census had a household response rate of 97% across England and Wales, of which almost 90% were online responses (ONS 2021).

The census data used for this linkage, has gone through census processing including steps such as Remove False Persons (RFP) and Resolve Multiple Records (RMR) but has not yet undergone Edit and Imputation (E&I). The E&I processing would seek to fill missing variables in the census data which could cause issues when trying to match individuals. The analysis uses post E&I census, but removes all imputed variables and people, as we are only interested in records which went through the linkage exercise.

Census Coverage Survey
The CCS is a 1% sample survey carried out six to eight weeks after the census and is a fundamental part of ensuring that the 2021 Census statistics represent the whole population, not just those who completed a census return. CCS is linked to census so that dual system estimation can be used to estimate and adjust for the undercoverage and overcoverage of the census.

Census and CCS combined (CC)
Within this project, we use the census and the CCS as a combined dataset, The census to CCS linkage has been reviewed by clerical matchers and had strict linkage requirements (a false positive rate < 0.1% and a false negative rate < 0.25%) which were achieved. We therefore assume that the census to CCS links are correct. This combined dataset is referred to as CC throughout the document.

3. The rationale of the linkage

The rationale for linking the DI to the CC is to provide a rich dataset to use as an evidence base to inform decisions for the statistical transformation of population and migration outputs.

Once the linkage exercise is complete, the subsequent analysis will facilitate:

- **An understanding of the quality of the DI**

This project will enable understanding of the quality of the Demographic Index (DI) in its current state and provide evidence to inform improvements for future linkages. To validate the DI's quality, a high-quality linkage to high quality data, such as the census, is required.

Understanding the quality of the DI is critical as the DI is used as an input for both the SPDs and the DPM. Recognising any poor quality or characteristics which the DI struggles to match accurately is a valuable insight and a first step in the development of options to improve this.

- **Inform the Statistical Population Datasets (SPDs) and DPM**

This linkage will allow us to assess and refine the rules we use to include people in the SPD and to allocate them to addresses.

There have been three final versions of the SPD so far (SPD V1.0, SPD V2.0 and SPD V3.0), where the methods of defining the usual resident population differ between them. Recently, a fourth iteration of the SPD has been produced and this was taken to the panel in a (EAP180) paper and presentation in November. SPDv4.2 is the version used in this project. In the future, if needed, other versions can also be compared.

As the DI uses multiple sources of data, different sources may be matched to the same individual with differing address information. The SPDs then use rules to allocate individuals to areas. Doing this in the most accurate way is important, so when analysis is carried out by geography, it reflects the true picture of the usually resident population across the geography. Understanding where individuals are found geographically between the CC and the different admin sources gives an indication of the potential time lag of address information being correct. This is also important in understanding internal migration.

This linkage should also provide insight into under and overcoverage in the SPDs by different characteristics, such as age and sex, which will inform the rules.

As the DPM relies on many of the datasets within the DI, it will also be informed by analysis of this linkage. The linkage will inform the DPM through understanding the quality of the DI, particularly coverage gaps. A specific DPM concern is around the lag of admin data, so research will be designed to make use of the “address from 1 year ago” census variable and comparing which DI addresses match to this address and which match to the “usual” census address.

Collectively, the linkage should enable analysis which allows us to define the estimation problem remaining after an SPD has been produced using rules to most closely replicate the usual resident population. We will then be able to assess the best way to achieve this within the DPM modelling framework.

- **Evidence of admin data quality for specific populations**

Some specific populations, i.e., Communal Establishment residents, special populations (Appendix 2) and vulnerable populations, are particularly difficult to identify and place in admin data. This linkage will allow the specific populations in both sources to be compared and the quality of what is found in the admin sources to be reviewed. This evidence will inform whether more focused sources are needed to target these populations and how those sources might be integrated successfully into the SPD approach.

- **Evidence to improve future linkages between the census and admin data**

This linkage will also provide insights for future linkages of the census data to other admin sources and offer evidence on how to improve or conduct future linkages using the DI. In particular, it will inform deterministic and probabilistic methods for future linkages to the census, in addition to how the linked outputs can be utilised. The clerical matching exercise, while focused on complex clusters and CCS2 areas, will inform future iterations of census linkage to the DI.

- **Evidence to support research on multivariate characteristics**

Linking the CC to the DI allows onward linkage to further datasets containing information about a range of population characteristics. Comparisons between census and admin data characteristics can be made to assess the relationship between to two data sources and inform work to develop new statistics and insight on those topics. These comparisons will highlight where there are gaps in our admin data, whether that be specific variables or understanding the representativeness of the variables we do have. This will inform next steps on ensuring multivariate characteristics can be represented in the work of ONS.

4. Research Questions and what they aim to inform

To meet these rationales, it is our intention to answer the following questions using this linkage. Appendix 3 includes a table which demonstrates how each research question ties back to the rationales. The research questions have been developed through engagement with relevant stakeholders and aim to meet the needs of the ONS and 2023 Recommendation. We have worked with our stakeholders to establish the priority of each research question, defining what analysis needs to be completed for the 2023 Recommendation and what can be reviewed at a later date.

Research questions are set out in four themes: geographic location, DI coverage, SPD comparison and Specific populations.

It is worth noting that the analysis will focus within a sample of CCS areas known as CCS2. This is because the clerical matching within the linkage design will focus on a 50% sample of CCS areas (referred to as CCS2) due to cost, time and resource. This is discussed in more detail in the design principles.

Geographic location

a) Within CCS2 postcodes, what percentage of CC usual residents have DI records in the same geography?

This question aims to analyse how well our administrative data geography compares to census. We plan to analyse a few different scenarios for this:

- How this varies for different levels of geography, e.g. How many have the same Unique Property Reference Number (UPRN)/Postcode/Output Area /Local Authority on DI and census
- How this varies by age and sex (and stage of life, such as student/working age/etc.)
- How this varies depending on the sources within a DI cluster

- How this varies by other CC characteristics (ethnicity, country of birth, religion, etc.)

b) Within CCS2 postcodes, which DI addresses align best with CC?

i) Establish a hierarchy of DI addresses, based on how accurate the geography on each source is compared to census by different levels of geography.

The main aim is to establish which admin data source has the most accurate geography for different population groups. For example, is it the case that HESA has the most accurate geography for students, PDS for pension-age, ESC for school-age etc. A hierarchy of the addresses can be established for the population groups so that if the most accurate source isn't available, another source can be chosen. Geography refers to UPRN (only available on PDS and ESC), postcode and LA.

This analysis can be done by age, sex, and LA initially, but also by other characteristics to see if that affects the accuracy of their geography. Examples of these characteristics are Hard to Count (HtC) areas, born in the UK, ethnicity, full time students and if they're new arrivals into the UK.

ii) Within CCS2 postcodes, for CC usual residents who have DI records in the same geography, what percentage of their DI records agreed with their CC geography. By different geographies (UPRN, postcode and LA)

This question aims to investigate what percentage of DI addresses match to the CC usual address, for CC usual residents who have DI records in the same geography. For example, if there were three DI records for a CC usual resident, and the postcode(s) for two of those DI records agreed with the CC postcode(s), that would be a 66.7% agreement. The average agreement percentages for different demographic characteristics, different combinations of DI sources, and the number of DI records can then be compared, to see if there is variation in these agreement percentages.

Note that for all geography analysis, HESA 2021 will be used, not HESA 2020. This is to ensure the most up-to-date geography information is used in comparisons.

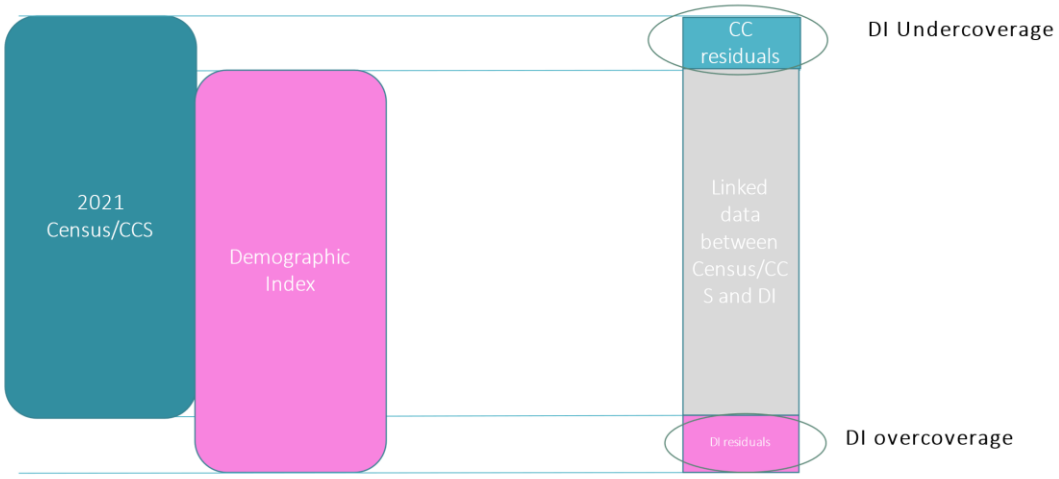
These research questions aim to inform:

- Decision making about preferred address for the DI/SPD
- How we develop the SPDs and their inclusion rules
- Understanding lag in admin data (through using the census address 1 year ago variable)
- Understanding of university students and other people living between addresses and where they appear in admin data

DI Coverage

What undercoverage and overcoverage does the DI have when compared to CC?

Figure 2: Populations of interest when comparing the coverage of the DI once linked to the CC



a) DI undercoverage: Within CCS2 postcodes, who has a CC return but isn't on the DI?

i. Grouped by age, sex (individually and then cross-tabulated by each characteristic that follows), LA, country of birth, nationality, ethnicity, students, non-English speakers, short-term migrants, armed forces, disability, religion, sexual orientation, employment status, carer status and qualifications

b) DI overcoverage: Within CCS2 postcodes, who is on the DI but not on CC

i. Breakdown the individuals classed as DI overcoverage by age, sex, LA
ii. Investigate which data source(s) led to their inclusion in the DI

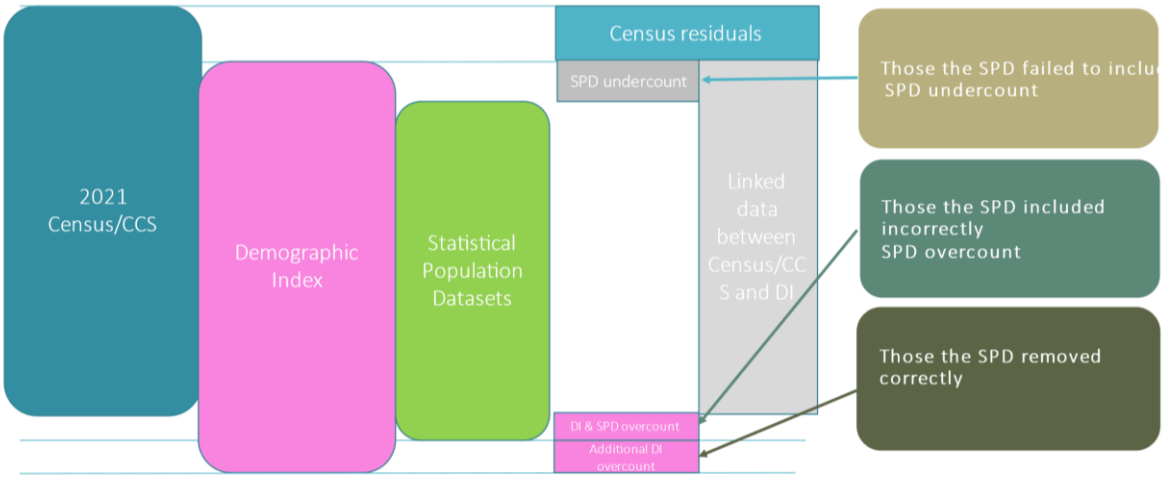
These questions aim to inform:

- Undercoverage in the DI
- Overcoverage in the DI
- Understanding if there is a relationship between undercoverage and lag
- Evidence to support whether further admin sources are needed to do research on multivariate characteristics

SPD comparisons

The three SPD questions aim to review those the SPD failed to include (i.e., in the DI and on CC but not included in the SPD), those the SPD included incorrectly (i.e., were not found on CC) and those the SPD removed correctly (i.e., were found in the DI but weren't retained in the SPD).

Figure 3: A diagram to illustrate the different populations of interests when comparing the SPD to the linked CC/DI data



a) SPD undercount: Within CCS2 postcodes, how many records were present according to the CC and DI linkage, but not included in the SPD and the reason they were not included? What are their characteristics and geography?

b) SPD overcount: Within CCS2 postcodes, how many records were found in the DI and the SPD but not found by CC?

i. How many can be explained by census undercount? What are their characteristics (age and sex), geography and which sources are they on?

c) Those the SPD removed correctly: Within CCS2 postcodes, how many records were in the DI that are not included in the SPDs, and not captured by the census? What are their characteristics (age and sex) and geography and which sources are they on?

These questions aim to inform:

- Understanding of SPD undercoverage
- Understanding of SPD overcoverage
- Understanding within LAs the level of undercoverage and overcoverage, by intersecting characteristics
- Validation of rules for SPD and rule development
- Understanding SPD undercoverage

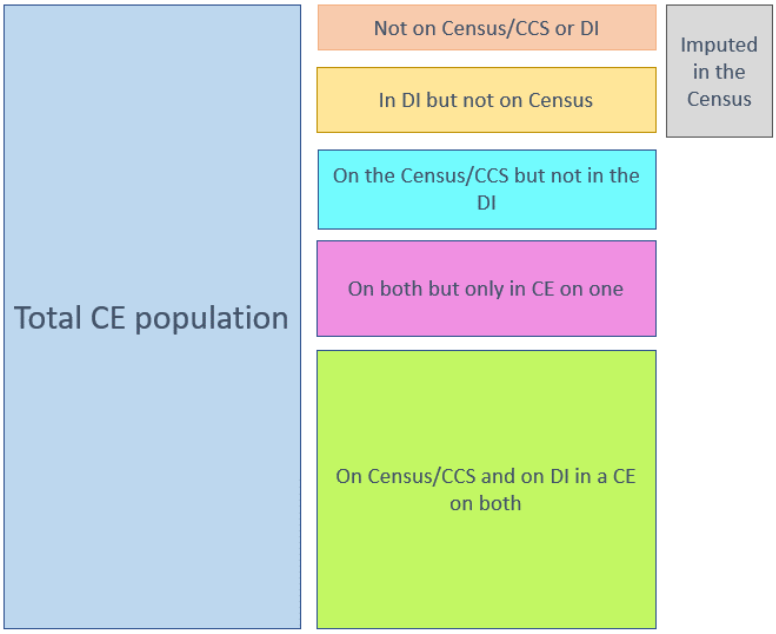
Specific Populations

A) Within CCS2 postcodes, how well do the CC/DI/SPD capture specific[2] populations?

[2] The 'Specific populations' to be analysed will cover the needs of our stakeholders. The question will be refined through further discussions.

This analysis focuses on looking broadly at all types of Communal Establishments (CEs) by comparing how well they are captured across the CC, DI and the SPD.

Figure 4: The different groups those in CEs could be identified in



To assess how well those in CEs are captured across CC and the DI, we will look at agreement status (whether a record is a CE resident on the two sources) and compare geography agreement (UPRN, postcode, and LA) by age, sex, and CE type. We will also review these against the SPD inclusions/exclusions.

This analysis aims to inform:

- Comparisons and quality review of specific populations in all sources
- Evidence to support if further admin sources are needed to do research on multivariate characteristics for specific populations
- How those sources might be integrated successfully into the SPD and DPM approach

5. The design principles of the linkage strategy

Based on the research questions and rationale outlined above, the design principles of the DI-CC linkage strategy have been listed below.

1. To conduct high-quality record linkage because the detailed findings will have important statistical and operational implications, outlined in the rationale for the linkage section.

2. To use clerical matching to achieve best possible links and to evaluate the automatic linkage.

3. Note that due to time and cost restraints, clerical work cannot be completed for the entire CC and DI. We are therefore restricting clerical work to those DI or CC records that have a postcode in a subsample of CCS postcodes. This subsample is approximately half the CCS postcodes and was selected by including CCS postcodes as follows:

- If there is only one CCS postcode in an output area, then select it
- If there are two or three CCS postcodes in an output area, then randomly select one of them
- If there are four or five CCS postcodes in an output area, then randomly select two of them

This method of selecting the subsample should mean that the subsample is stratified in the same way as the original CCS sample and will include postcodes from across England and Wales. To be within a CCS2 area, either an ONS ID must contain one source in a CCS2 area, or the CC usual or alternative address is in a CCS2 postcode.

We reviewed the HtC group representation in the subsample compared to the original CCS sample. Table 1 below demonstrates that the percentages of HtC are similar, though our sample is slightly skewed towards the hardest to count groups, where group 5 is the hardest to count group and 1 the easiest (Dini 2018).

Table 1: Percentage of postcodes in HtC groups by the original CCS sample and the subsample used for the clerical matching in the DI-CC linkage project

Hard to Count groups	Full CCS sample	Sample used for clerical (~50% of CCS)
	Percentage in each HtC group	
1	28.8%	27.05%
2	40.2%	39.53%
3	12.4%	13.22%
4	14.1%	15.28%
5	4.4%	4.91%

This concept was agreed at the Census Research Assurance Group (CRAG) so has been discussed at assurance groups but also with the linkage and estimation expert groups. Ideally, we would have then gone on to produce weights to enable us to make inferences about the whole population of England and Wales, but due to the complex sampling for the CCS and the additional sampling of the CCS postcodes for CCS2 areas, this will not be feasible in our current timescales. In the future, this work can be further scoped out if it is deemed necessary, but for now we will have to review our analysis while considering the implications that the sample is skewed to include more census HtC populations.

4. To use all possible census and DI addresses, to utilise as much of the data as possible to find matches across sources.

The census questionnaire allows respondents to provide an alternative address, which could validly match to an admin record. For example, the impact of the

pandemic on the enumeration of students means that we may link a student at either their usual census address or their alternative census address and including both in our linkage will help capture both of these scenarios. Using census address one year ago helps us understand time lags in admin data. The DI is made up of several sources which may have differing address information. Using all these sources and their address information within the linkage will help inform how we can best use the address information in the future.

5. To extend automatic linkage to the entire census and DI, not only CCS2 areas. This is because individuals in CCS2 areas, on either source, may be found in non-CCS2 areas on the other source so prevents false positives in the matching.

6. Comprehensive flagging of results to support detailed analysis, including flags which detail the way in which clusters were matched. This will allow us to utilise the knowledge that was produced through the linkage, rather than taking the links and not accounting for the varying quality of matches made at different stages of the process (i.e., links can be made automatically, clerically, inferred).

7. We propose the use of non-greedy matchkeys, which means that all records are put though every matchkey and then conflicts are clerically resolved and the 'best' match is chosen. In other words, we will not remove a record from the 'pot for matching' once a match has been found. There are several reasons why we prefer the non-greedy approach:

- Suppose a record matches on a relatively loose matchkey at the preferred address and to a different record on an exact matchkey at the non-preferred address, causing a conflict. Both matches could be correct, but the exact match is arguably stronger. In a greedy approach, only the fuzzy match would be found.
- The answer to the research question 'Within CCS postcodes, which DI addresses align best with census' will be biased if we do not look for a match once the preferred address match has been found, as no other addresses will have potential to match.
- Using non-greedy matchkeys will also help us to quantify the number of duplicates in the DI as conflicts can be created (i.e., two DI records matching to a single census response).
- There are cases of duplicates in the census (in 2011 ~ 350,000 persons were recorded multiple times), and the use of greedy matchkeys may cause links to duplicates to not be found, despite being correct links. Many duplicates have differing addresses, and so linking all instances of a person on the census to the DI will help in understanding the quality of the DI addresses.
- If greedy matchkeys are used, the order of the matchkeys affects the outcome of the linkage. Accepting only the first link made assumes that the ordering is correct.
- Linking all records future-proofs the matching and will enable flexibility at the analysis stages. For example, suppose it is decided after the matching is completed that the SPD definition of preferred address is not correct. This would require re-running the matching if greedy matchkeys are used since the hierarchy of the matchkeys would be incorrect. However, with the proposed non-greedy strategy, making this change would only involve changing the flags that say which address is the preferred address.

8. For clerical matchers to be able to ‘merge’ DI clusters where they find individuals who they believe to be the same and to also be able to split DI clusters and remove sources they do not believe belong to the individual. This will enable us to understand matching failure in the current DI and learn how to improve the DI’s matching method in the future.

6. Assumptions

2020 versions of HESA will be sufficient in the linkage

We are unable to change the frequency of delivery of some admin sources to suit our current timeline, so extracts of HESA included in the linkage will be from 2020 (2019-2020 academic year) rather than 2021 (2020-2021 academic year). If we were to wait for the more up-to-date data sources, it would delay the start of matching by a minimum of four months. It is important that these linkage plans are done at pace, to allow enough time to implement improvements into the transformation work currently underway for the 2023 recommendation.

To understand the potential impact of using HESA 2020 data instead of HESA 2021 data, analysis was carried out exploring the quality of address information in both extracts. 2020 HESA was linked to and compared against 2021 HESA to assess: how many of those found in HESA 2021 will be missing in the DI-CC linkage analysis; how many students left in HESA 2020 would be found in the linked data and what the geography of those on both extracts looked like. HESA 2020 was also compared to 2021 PDS data to assess the accuracy of the geography of student leavers (e.g., how many students update their address on administrative data within a year of finishing university?). This can help us to understand how many people are being incorrectly included in analyses on students, as well as how they might act in administrative data, providing context to conclusions drawn for these populations.

Initial analysis found that, using 2020 HESA instead of 2021 HESA will not affect the DI-CC linkage analysis greatly, as less than half of people who appear on 2020 HESA, but not on 2021, updated their address on the PDS within a year of leaving university. However, there are some exceptions. When conducting analysis on 18–20-year-olds, it should be noted that there are many who appear on 2021 HESA, but not on 2020 HESA, and will therefore be missed in our analysis if they are not present on another DI source. Consequently, they may have less accurate geography within their ONS ID cluster. When conducting analysis on 22–25-year-olds, it should be noted that there are many who appear on 2020 HESA, but not on 2021 HESA, so are being incorrectly captured in analyses involving student populations. We could therefore see records being incorrectly matched on geography data or having presence in the DI where they should not.

As noted previously, for our CEs and geography work we will be integrating HESA 2021 where possible. This means that even though the HESA 2021 wasn’t used in the linkage, for most cases we will be able to use the updated HESA 2021 information to compare with Census.

Using CIS 2020 in the DI build but CIS 2021 in the linkage is acceptable

The DI was built using CIS data from 2011-2020. This data was removed and completely resupplied prior to this analysis, thus the linkage was conducted on the

2021 re-bulk which contained the 2011- 2020 data, although further cleaning and edits may have been made to the data before reaching ONS. We have reviewed the implication of how the newly supplied CIS data may have changed the construction of the ONS IDs within the DI (i.e., new information or better-quality cleaning could lead to ONS IDs splitting or merging) and we found less than 0.05% of CIS master keys linked to a different ONS ID in the DI2.1 compared with DI2.0.

We will account for the effects of COVID-19 on admin data

A key consideration ahead of the analysis of the linked DI-CC is how the coronavirus pandemic may have impacted the quality of both the administrative data sources within the DI as well as the Census and CCS. The widespread displacement of people throughout the pandemic period could have led to inaccurate address information on both administrative and census data. We have scrutinised the effects on both the DI and the CC:

- GP registrations on the PDS are higher in December 2020 compared to previous years, reflecting the start of the vaccination programme. Registrations remained higher than previous years until August 2021 when the trend aligns more with previous years. The PDS used in the linkage was the June 2021 extract and will have benefitted from potentially more up-to-date information than usual.
- WSC had lower registrations initially but as the year went on registrations increased, balancing out the effect.
- For HESA, when comparing counts of activity in April to previous years, a decline is evident, thus records are more likely to be out of date. HESA issued exceptional guidance for the academic year 2020/21 that providers did not have to return a term time postcode if it was especially burdensome to collect.
- Census Field Operations suggests that there is little evidence that COVID caused issues with household response rates, though there is concern around the impact on communal establishments. Many establishments had their own COVID protocols, in particular those with vulnerable residents (e.g. care homes), thus were operating under strict guidelines. This resulted in difficulties in gaining access to these establishments and therefore impacted the interviewers' ability to capture a response.
- Furthermore, student displacement as a result of COVID was a flagged as a major issue throughout the campaign. Census instructed students to include their term-time address as their usual address, however as many students remained at their non-term-time address during the academic year it is likely that this guidance was not universally followed.
- Similarly, many military bases were closed as a result of the pandemic, thus military personnel were enumerated elsewhere in households. This caused return rates in these locations to be lower than they otherwise would be.

Note that these are summaries and we've considered more beyond this.

Clerical matching being targeted to CCS2 areas is acceptable

It is critical that the quality of the linkage is optimised, including clerical matching and review. Because high quality linkage is expensive, we have decided to focus on a

sub-sample of census records for this analysis. These specific areas are targeted in CCS2 areas because we can utilise having both census and CCS responses. In addition, it will allow us the opportunity to use the linked data in CCS areas to test a dual system estimation approach if we decide, post analysis, that this is an appropriate method to explore.

This relies on the assumption that analysis done on CCS2 areas is useful for areas beyond CCS areas. CCS areas are specifically chosen to be representative of the entire population, but also capture areas declared as Hard to Count. Large Communal Establishments are not covered by the CCS. We have reviewed whether large CEs were present in the CCS2 postcode areas and whether they covered the different types and identified that US military bases were not included. Additional clerical review was undertaken to ensure they were represented.

When using the outputs of the analysis we will review the impact of the clerical matching being targeted in CCS2 areas only. Use of clerical in CCS2 areas will increase precision but will also mean that outside of CCS2 areas the recall (missed matches) will be higher than where clerical matching has been used. Therefore, our analysis is focused on CCS2 areas where the best quality of linkage has occurred. However, we will have to be mindful of how much generalisation is acceptable.

We can only do what is within our capability with hashed data

The CIS data received for the linkage will be hashed, blocking out any Personal Identifiable Information (PII). This means that the linkage method for CIS is limited, and we have no ability without the PII to design a bespoke linkage, assess the quality or carry out any clerical on CIS data. Therefore, less analysis can be conducted on the CIS data. However, the number of records in the DI with only CIS addresses is very small so the quality of the linkage on most records will still hold. We intend to assess the implications of using hashed CIS data for our analysis and adapt it appropriately.

7. Future Work

The work we have detailed in this paper covers our priority analysis, but further work has been planned, including:

- Reviewing those who are found in both the admin data and the CC
- Household analysis and comparisons to the Admin Based Household Estimates
- Further work to look at Specific Populations focusing on students, boarding school pupils and UK Armed Forces
- Reviewing the agreement between postcodes within a DI cluster
- Reviewing our geography questions with census address 1 year ago to better understand the quality of address information and lagging in administrative data, to improve our methods of producing admin-based internal migration estimates
- A quality assessment of how using clerical matching in this project (and the ability to spit up and merge ONS ID) can help us understanding the quality of the DI's current matching method and also improve the DI's future matching

method. Much of this work has been outlined already by Rosalind Archer et al (2022) at MARP in November
- A review comparing CC residuals with DI residuals, including imputed census records to understand the impact of census undercount
- Further analysis to provide evidence for the coverage adjustment of the SPDs for the DPM

References

Archer R, Kenning E, Lloyd S and Vellanki P (2022) 'Evaluating Statistical Quality in the Demographic Index' November 2022. Archer R, Kenning E, Lloyd S and Vellanki P (2022) 'Evaluating Statistical Quality in the Demographic Index' EAP 182 November 2022.

Blake A (2020) 'Developing our approach for producing admin-based population estimates, subnational analysis for England and Wales: 2011', July 2020

Black A (2021) 'Developing admin-based population estimates, England and Wales: 2016 to 2020' November 2021

Blackwell Louisa (2021) Integrated statistical design for the transformed population and social statistics system- Bayesian methods for demographic estimation. MARP paper EAP 174 December 2021
Blackwell Louisa (2021) 'Dynamic population model for local authority case studies in England and Wales: 2011 to 2022'

Dini, E (2018) 'Hard to Count Index for the 2021 Census MARP Paper' EAP 102 April 2018

McNally J (2022) 'Statistical Population Dataset version 4: Research to Date and Future Developments' EAP 180 November 2022

Office for National Statistics (2019) 'Transforming population and migration statistics: Research into developing an alternative approach to producing administrative data-based population stocks and flows', January 2019

Office for National Statistics (2021), 'Digital take up of Census 2021 beats targets', October 2021

Office for National Statistics (2022) 'Linkage methods for Census 2021 in England and Wales'

Appendix 1 – Statistical Population Dataset (SPD) rules

Version	Sources used	Rules applied
SPDv1	GP records (PR), National Insurance records (CIS), HESA data for students	1) Must be on PR and CIS 2) Address allocated 50:50 if sources don't agree *unless* on HESA (in which case students allocated 100% to student location)
SPDv2	PR, CIS, Eng/Welsh School Census (SC), HESA	1) Must be on 2 of 4 sources 2) School-aged children allocated to SC address 3) Students allocated to HESA address 4) "Activity" info from DWP and health data (PDS) used to resolve conflicting addresses for adults
SPDv3	PDS, CIS, Benefits and Income data, School Census, HESA	1) "Hierarchy of belief" model – give preference to source most likely to cover a particular population group 2) Use intelligence from all sources to act as "signs of life" (intended to only keep "active" records, and to assign geographically) 3) Objective: remove overcoverage (strict rules)

Appendix 2 – List of Communal Establishments and Special Population Groups

Communal Establishments:

APPROVED PREMISES
BOARDING SCHOOL
CARE HOME
EDUCATION OTHER
HALL OF RESIDENCE
HIGH SECURE MENTAL HEALTH
HOSPICE
HOSPITAL
HOSTEL
HOTEL
IMMIGRATION REMOVAL CENTRE
LOW/MEDIUM SECURE MENTAL HEALTH
MILITARY SLA (Barracks)
MILITARY US SLA (Barracks)
PRISON
RELIGIOUS COMMUNITY
RESIDENTIAL CHILDRENS HOME
ROUGH SLEEPER
STAFF ACCOMMODATION
YOUTH HOSTEL

Special Population Groups:

CARAVAN
EMBASSY
MARINA
MILITARY SFA (Houses behind the wire)
MILITARY US SFA (Houses behind the wire)
ROYAL HOUSEHOLD
TRAVELLING PERSONS

Appendix 3 – Research questions and which rationale they correspond to

	Rationale the research questions supports						
Question	DI Quality	Inform SPD	Geographic location	Evidence of admin data quality for special populations	Evidence to improve future linkages	Evidence to support research on multivariate characteristics	Inform the DPM
Within CCS2 postcodes, what percentage of census usual residents have DI records in the same geography? Patterns by age, sex, geography.							
Within CCS2 postcodes, which DI addresses align best with census?							
What undercoverage and overcoverage does the DI have when compared to CC?							
Within CCS2 postcodes, how many records were found in the DI and the SPD but not found by CC?							
Within CCS2 postcodes, how many records were in the DI that are not included in the SPDs, and not captured by the CC?							
Within CC2S postcodes, how many records were present according to the CC and DI linkage, but not included in							

the SPD and the reason they were not included?							
How well do the census/DI/SPD capture specific populations?							