MARP Paper 01 March 2022

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

Authors: Elizabeth Pereira | Emma Hand | Jack Rodgers

Social Statistics Transformation, Analysis and Research

Purpose of paper

This paper will outline the planned work for linking the 2021 Census and Census Coverage Survey (CCS) to the Demographic Index (DI). At present the methodology and the analysis plan is in the development stage and we seek guidance and advice from this Panel on our research questions and any notable unexplored areas.

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 will inform
V. Design principles
VI. Assumptions

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 regular (monthly) and timely 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)

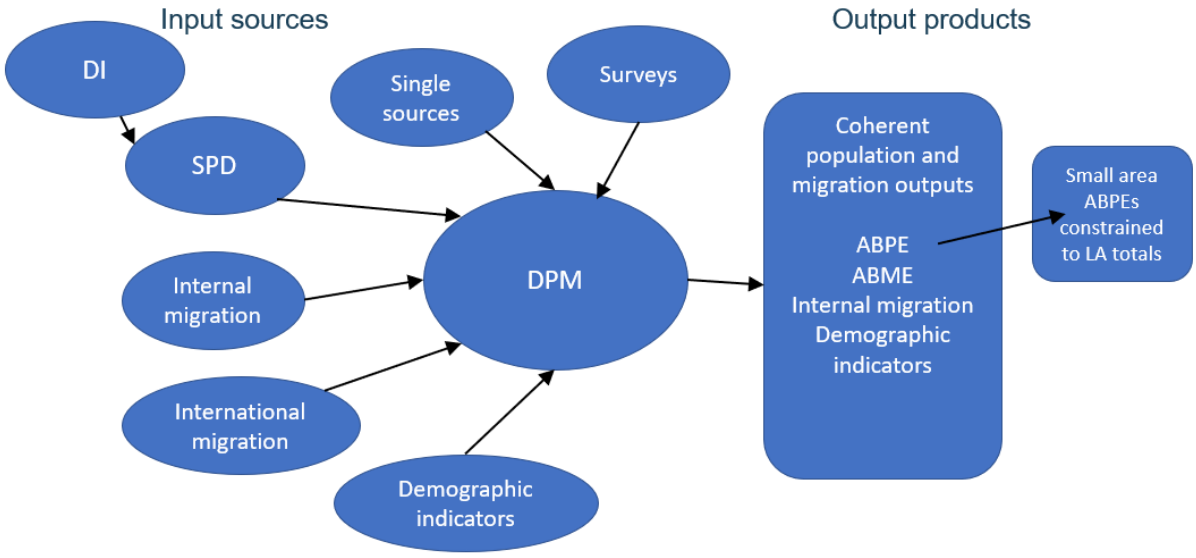
In 2023 the National Statistician will make a recommendation on what's needed in future. The transformed system will continue to iteratively build on what ONS have achieved.

ONS have created Statistical Population Datasets[1] (SPDs[2]). These use rules to combine admin 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 further develop the rules based on findings from the linkage discussed in this paper.

ONS are also working on 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.

[2] SPDs have recently been referred to as Admin Based Population Estimates (ABPEs) but they have been re-named to SPDs to avoid confusion with the planned output statistics which will now be derived from the dynamic population model as shown in Figure 1.

The Demographic Index

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

- 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 of Work and Pensions, covering Pay As You Earn and benefits data[1]
- The Births Register
- Individual Learner Record

[1] Does not cover self-employed

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.

We will be using Version 2.0 of the DI which contains 2021 extracts of PDS, ESC and WSC and 2020 extracts of CIS and HESA (the most recent data available). 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.

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 as high a quality of 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.

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. 97% of households across England and Wales responded to the Census 2021 of which almost 90% did so online (ONS 2021).

Census Coverage Survey

The Census Coverage Survey (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.

3. The rationale of the linkage

The rationale for linking the DI to the census/CCS 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 also provide evidence to inform improvements for future linkages. In order to validate the DI's quality, a high-quality linkage to high quality data such as the census is required.

Having an understanding of the quality of the DI is critical as the DI is used as an input for both the SPDs and the DPM. Understanding any poor quality or characteristics which the DI struggles to match accurately is 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 versions of the SPD so far (SPD V1.0, SPD V2.0 and SPD V3.0), where the methods of replicating the usual resident population differ between them.

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 on characteristics, it reflects a true picture of the usual resident population across the geography. Understanding where individuals are found geographically between the census/CCS 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 to understand under and overcoverage in the SPDs by different characteristics, 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. However, a specific DPM concern is around the lag of admin data and 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 will allow 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., special populations (Appendix 2) and vulnerable populations, are particularly difficult to identify and place in admin data. This linkage will allow these 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 for 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 insight into future linkages of the census data to other admin sources and provide evidence of 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 but also in how to utilise the linked outputs. The clerical matching exercise, while focused on complex clusters and CCS areas, will inform future iterations of census linkage to the DI.

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

Linking the census/CCS to the DI allows onward linkage to further datasets containing information about a range of population characteristics. Comparisons between census characteristics and admin data characteristics can be made to assess the relationship 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 variables or on the representativeness of the variables we do have, this will, in turn, inform next steps on ensuring multivariate characteristics can be represented in the work of ONS.

4. Research Questions and what they aim to inform

In order to meet these rationales, it is our intention to answer the following questions using this linkage. In Appendix 3 there is a table which demonstrates how each research questions 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. As we design the analysis plan, the questions will be further refined and developed in more detail. We are working with our stakeholders to establish which research questions are the priority and to define a sequence in which the questions need to be addressed. The linkage strategy and outputs from the linkage are being designed in a way that should offer the flexibility to explore these questions fully. Both the linkage strategy and the analysis plan will be presented to MARP in the coming months.

It is worth noting that the research questions focus within CCS areas, this is because the clerical matching within the linkage design will focus on CCS areas due to cost, time and resource. This is discussed in more detail in the design principles.

Research questions, by theme[3] :

Geographic location

1. Within CCS postcodes, what percentage of census usual residents have DI records in the same geography? Patterns by demographic variables and geography

2. Within CCS postcodes, which DI addresses align best with census?

These two research questions aim to inform:

- How we develop the SPDs and their inclusion rules
- How well the DI captures the usually resident population by age/sex/Hard to Count areas/other variables
- Understanding lag in admin data
- Understanding of university students and other people living between addresses and where they appear in admin data
- Decision making about preferred address for the DI/SPD

DI Quality and coverage

3. Quality of the DI

a. **How much false positive matching within the DI identities?**
b. **How many false negatives are there in the DI's current method?**

This question will aim to inform:

- Future DI matching strategy i.e. change matching methods
- Understanding of the DI's quality
- DPM and SPD (who use the DI as a direct input) use of the DI
- Future linkages between DI and census

4. What are the coverage gaps in the DI?

a. **Within CCS postcodes, who has a census return but isn't on the DI?**
b. **What are their characteristics and geographies?**

This question will aim to inform:

- Undercoverage in the DI
- Understanding if there is a relationship between undercoverage and lag

[3] Note that the themes are not mutually exclusive and some questions might provide insight to more than one theme, for simplicity research questions are only listed once.

- Evidence to support if further admin sources are needed to do research on multivariate characteristics

5. Within CCS postcodes, how many records were found in the DI and the SPD but not found by census?

a. How many can be explained by census undercount? What are their characteristics and geography?

6. Within CCS postcodes, how many records were in the DI that are not included in the SPDs, and not captured by the census?

These two questions will aim to inform:

- Understanding of people in "the wrong" place on the admin data
- Understanding of people in the admin data that were not in census (NOTE limited characteristics will be available for these)
- Understanding of addressing to inform SPD and Admin Based Household Estimates
- Validation of rules for SPD
- If there are certain populations the DI overcounts

SPD inclusion rules

7. Within CCS postcodes, how many records were present according to the census/CCS and DI linkage, but not included in the SPD and the reason they were not included?

This question will aim to inform:

- SPD rule development
- Understanding SPD undercoverage

DI/SPD Capture

8. How well do the census/DI/SPD capture specific[4] populations?

This question will aim to inform:

- Comparisons and quality review of vulnerable 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 approach.

9. Within CCS postcodes, where census found non-DI residents within an address, how different are the census and DI demographics of residents at that address?

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

This question will aim to inform:

- The extent to which household age/sex/relationship structures remain intact even when we don't have the right people.
- How accurate our aggregate outputs are

5. The design principles of the linkage strategy

Based on the research questions and rationale outlined above, the design principles of the DI-census/CCS linkage strategy have been listed below. Note that the linkage methodology is designed as they start to use the data and so in coming months this will be presented to CRAG and MARP.

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. To restrict the clerical matching to CCS postcode clusters to limit scale due to costs and resource, and because we have both census and CCS responses here. The CCS has been designed to be stratified by Hard-to-Count score (HtC: areas at risk of census non-response) and is conducted by door-to-door interviewers who, unlike the census, are not reliant on a list of addresses but rather have to explore the area to find all residential addresses within a postcode. Therefore, using the CCS areas increases the likelihood that we are capturing more of the usual residences than using census alone.

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 offers the ability for 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. There are time lags in admin data and so using census address one year ago helps us understand this lag. The DI is made up of several sources which may have differing address information, using all these sources and their address information 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 CCS areas. This is because individuals in CCS areas on either source may be found in non-CCS areas on the other source so prevents false positives being matched.

6. Comprehensive flagging of results to support detailed analysis, which is currently in development and we can bring to MARP with our analysis plans at a later date.

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/or 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 as follows:

- 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 be given a chance 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 in the census), 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 makes the assumption 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 for the clerical matchers 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.

6. Assumptions

2020 versions of HESA and CIS will be sufficient

We are unable to speed up or change the frequency of delivery of some admin sources to meet our current timeline, so extracts of HESA and CIS will be from 2020 rather than 2021. 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.

This means that the more up-to-date sources in the DI have the potential to have more up to date address information. We intend to assess the implications of these reference point inconsistencies for our analysis and adapt it appropriately. In particular, we will work through how these impact on any analysis of students and how to interpret this.

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

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 within 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.

We are using the optimal version of census data and census-CCS links are correct
The census cut 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. If a variable has been imputed but is included in the matching exercise then this could cause false matches to be made.

The census-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-CCS links are correct.

References

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

Blackwell Louisa (2021) Integrated statistical design for the transformed population and social statistics system- Bayesian methods for demographic estimation. MARP paper December 2021

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 (2021), 'The Census Coverage Survey', November 2021

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
1. Within CCS postcodes, what percentage of census usual residents have DI records in the same geography? Patterns by age, sex, geography.							
2. Within CCS postcodes, which DI addresses align best with census?							
3.Quality of the DI a. Is there any false positive matching within the DI identities? b. Is there any false negatives the DI's current method failed to match							
4. What are the coverage gaps in the DI?							
5. Within CCS postcodes, how many records were found in the DI and the SPD but not found by census?							

6. Within CCS postcodes, how many records were in the DI that are not included in the SPDs, and not captured by the census?							
7. Within CCS postcodes, how many records were present according to the census/CCS and DI linkage, but not included in the SPD and the reason they were not included?							
8. How well do the census/DI/SPD capture vulnerable populations?							
9.. Within CCS postcodes, where census found non-DI residents within an address, how different are the census and DI demographics of residents at that address?							