**High Level Statistical Design for the Transformed Population and Social Statistics System**

1. **Introduction**

Statistical Design and Research (SDR) are responsible for delivering the statistical design for the 2021 Census and longer-term transition towards an admin led system for producing population and social statistics. Development and testing of the statistical design will provide evidence for the National Statistician’s Recommendation to Government in 2023 about the future population and social statistics system.

Timely delivery of the statistical design will be dependent on a range of products and services being delivered across ONS transformation teams. This paper outlines the high-level design principles underpinning work across SDR projects, with a particular focus on Population and Migration Statistics Transformation (PMST) and Social Statistics Admin First (SSAF). It also describes work being undertaken by the Census Statistical Design Team (CSD) to prepare for scenarios of low response to the 2021 Census.

A detailed statistical design has not yet been fully established for the future population and social statistics system. While methods continue to be researched and tested, this paper is intended to demonstrate how ONS projects fit together to support the overall design and highlights key methodological challenges that need to be progressed ahead of the 2023 recommendation.

1. **Key deliverables for the future population and social statistics system**

The future statistical design needs to continue delivering the existing range of outputs produced from censuses and social surveys where a clear need can be established. It also needs to make improvements to meet increasing user demand for more timely, frequent and granular statistics. In addition, the new statistical system needs to be flexible to deliver insights in expanding topic areas, for example those emerging as new priorities to support government policy.

Some of these demands for information can be met by making more integrated data available to approved researchers. Well-established and widely used sources of information, such as the Census Longitudinal Study will require redesign under a future admin-led system, however there is considerable potential to expand on the concepts underpinning this type of research within an admin-led system.

Key deliverables for the integrated statistical design are summarised below:

***Estimates of population size*:** Census and mid-year population estimates are produced at various levels of geography, down to output areas. The accuracy of mid-year population estimates is known to reduce during the period between censuses, however quality standards are set to ensure precision of national and local authority population estimates in census year[[1]](#footnote-1). The new system will aim to deliver local and national population estimates to prescribed levels of precision annually. The level of precision is yet to be determined and will be part of consultation with users regarding the success criteria for future population statistics.

***Estimates of population characteristics*:**Census collections enable detailed tabulations at small area level across the full range of census topics. This is uniquely available to users every 10 years and there is demand from users to increase the frequency and granularity of detailed small area statistics. The future population and social statistics system will need to deliver univariate and multivariate statistics for key population characteristics annually, as well as expand the range of topics beyond those traditionally collected from the Census.

***Statistics on households and housing*:** Censuses and social surveys are primarily designed as household collections and the statistics produced provide important information about housing demand in local areas. Censuses outputs include key statistics on the number, size and composition of households every ten years. ONS social surveys provide some statistics on households in the period between censuses, most notably the Labour Force Survey and Annual Population Survey. However, relatively small sample sizes prevent delivery of robust statistics at local area level. Information about housing conditions and property characteristics are partially collected across ONS surveys and the English Housing Survey, but with similar issues relating to sample size. Statistics on populations living in Communal Establishments are also limited in the period between Censuses. A future system led by admin data may require a departure from traditional ‘household’ definitions to meet user requirements for more frequent small area statistics.

***Measures of population change*:** The current population statistics system uses a *cohort component method* to produce annual mid-year population estimates in the years between censuses. Transition to an admin-led system may result in a departure from this method to ensure that there is coherence between annual estimates of population size and estimates of net international migration. ONS is also required to deliver more analysis about the impact of migration in UK society, requiring wider integration of data sources into the population statistics system.

***Responsive analytical service*:** ONS is increasingly called on to provide analytical insight in areas of evolving policy interest. Under current arrangements these are undertaken by ONS business areas as ‘ad-hoc’ projects, sometimes with limited availability of data. There is a need to streamline and capture the prioritisation of user needs within an admin first system through ONS wide engagement and consultation. The new population and social statistics system will be able to integrate and exploit existing and new information sources to maximise analytical value.

***Improved microdata and tools for users:***One of the advantages of an admin-led system is the potential for cross-sectional and longitudinal analysis. ONS will use privacy preserving approaches for linking data sources and develop methods to make datasets available to approved researchers as part of the Secure Research Service (SRS). The flexible table builder and disclosure control methods developed for the 2021 Census will have a legacy to support future admin-based population and social statistics.

**2.1 Minimum requirements for an admin-based system from 2023 onwards**

SDR has not yet determined what minimum evidence is required to support a recommendation for an admin system after 2023. A document outlining key success criteria has been drafted and will be developed further once a statement of user need, as understood across PPP and PPA, is better understood. SDR are working closely with Census and Transformation Policy Team on a strategy for engaging our users in these discussions. To recommend switching to an admin led system, our current assumption is that by the 2023 recommendation we will need to have evidenced readiness to produce population and migration estimates from 2024 onwards and have the capacity to deliver key population characteristics that meet users’ needs by 2031.

1. **Functions supporting the future statistical design**

All supporting functions underpinning the future statistical design are already being progressed by ONS transformation teams with support from MDR and DST. While many of these functions are not being developed exclusively for the future population and social statistics design, requirements have been gathered from SDR and work is either already progressing or plans are in place for delivery. Successful transition to an admin led system will require these functions to have reached a certain level of maturity by the time of the recommendation in 2023. Some of the requirements for future design may conflict with urgent priorities in the short term, notably 2021 Census delivery. Decisions to de-prioritise work supporting the future statistical design will impact progress of research to support the 2023 recommendation. Key supporting functions are:

Reference Data Management Framework (RDMF): RDMF v0.1 was delivered in 2019, largely comprising of proxy versions[[2]](#footnote-2) of the key indexes to underpin ONS data architecture (address, demographic, business, geography, and classifications). Since then a first version of the full Demographic Index has been made available to PMST in DAP. RDMF v0.2 is underway and will include more complete versions of the indexes. To support a new admin-led system after 2023 the RDMF indexes will need to be periodically updated , with a system in place to assign linkage identifiers to all of the key datasets supporting the design.

Acquisition of admin and other non-survey data: Some of the key datasets considered to have broad coverage of the population are already supplied to ONS. We need to secure ongoing supply of these datasets and progress with new acquisitions to support methodological research. Critical data acquisitions that need to be progressed include those that show evidence of activity associated with an address, or key topic areas where we need to have shown progress in our research by 2023[[3]](#footnote-3).

Integrated Population and Characteristics Survey (IPACS): Based on current assessment of data sources in the UK, it is unlikely that the transformed population and social statistics system can be entirely based on admin sources alone. Known coverage errors and the absence of key topics from admin data requires that social surveys remain an integral part of the future statistical design. The proposed IPACS collection being led by Social Survey Transformation (SST) is partly designed to meet PMST requirements to assess coverage of admin-based population estimates (ABPEs) as well as collect data to meet PPA’s requirements for producing key population characteristics.

Data Access Platform (DAP): The future statistical design needs all datasets to be available in one place. PMST, SSAF and Census teams are currently split across ONS research environments. Significant upskilling of SDR personnel is needed to take advantage of cloud computing and open source software in DAP. DAP processes also need to be standardised to meet SDR requirements (e.g. data import, standardisation, validation, geo-referencing, linkage, statistical methods, disclosure control etc), all of which need to be part of the ONS cloud strategy.

Methodology:Methods for producing robust statistics with measurable quality need to be thoroughly tested and reviewed in advance of the 2023 recommendation. At this stage, projects within SDR are at different levels of maturity regarding methodological development, some because of how Methodology have to prioritise their workloads to balance the demands on them from Operations, Transformation and Census Business Areas. The Admin Methods Research Team has been set up in MDR to develop the framework for incorporating statistical methods and theory in the move towards using admin data in official statistics. Projects commissioned by this team are pursuing some of the recommendations made in the Skinner Review of the Beyond 2011 Programme, with a focus on assessing the quality of admin-based statistics. Key methodological challenges that need to be progressed as part of our evidence in 2023 include: (i) reconciling estimates of migration flows with annual estimates of population size, (ii) methods for producing multivariate statistics from combined sources, which include admin, survey and Census data.

Statements of User Need: ONS has a range of mechanisms for capturing user need for census and social survey collections. There are also forums for capturing user requirements to improve regular outputs and develop new or bespoke outputs. PPA are currently leading an exercise with topic leads and other business areas to review what is already known from engagement with users across the office. This will deliver a statement of user need to help shape the design of the future statistical system. SDR had originally intended to run a consultation in summer 2020 outlining a success criteria for switching to an admin led system, and are working with the Census and Transformation Policy team on the strategy for engaging users ahead on the 2023 recommendation.

Business Change: Implementation of the future statistical design will require recruitment and development of new skills at ONS. Some of the roles in existing business functions will need to be adapted to support the new statistical system. Work The target state will need to be quantified in terms of resource requirements, estimated costs associated with the new system, and the realisation of benefits to support the economic business case.

1. **Data Architecture**

The Reference Data Management Framework (RDMF) is delivered and maintained by ONS Data Architecture. It provides the base infrastructure supporting the new population and social statistics system. A brief summary of how the RDMF indices support the statistical design is summarised below:

Geographic Index: The geographic index will be used to produce standardised statistical outputs at all levels of ONS geography. Should 2021 be the last Census collection, alternative data sources will need to be used to update ONS statistical geographies, including output areas, daytime/workday populations, workplace zones and urban/rural classifications.

Classification Index: Classifications will need to be maintained and updated to ensure standardised coding of variables for consistent output production, for example Standard Industrial Classifications (SIC) and Standard Occupational Classifications (SOC).

Business Index: The business index will be an important source for identifying and accurately classifying CE’s on AddressBase. It will also provide a link between employers and employees, allowing development of future admin-based outputs (for example travel to work statistics). The link between employers and employees will also provide opportunities for new analysis, for example gender pay gap analysis, and social mobility.

Address Index: The address index will be used as the frame for sampling addresses in the IPACS collection, which will be an ongoing collection sampled quarterly. Combining information from across the address index and the demographic index will also be the basis for producing future household statistics from an admin led system.

Demographic Index:The demographic index is the key architecture for producing underlying population counts in the future population statistics system. It is also needed to support the integration of admin records to support 2021 Census estimates. The first version of the Demographic Index is now available in DAP, with PMST and SSAF teams building on research previously undertaken in the Secure Research Environment (SRE). The next section gives a brief summary of how the demographic index is used to support admin-based population estimates (ABPEs).

1. **Demographic Index**

Up until recently PMST have based all research on a prototype ‘population spine’, based on encrypted data and held in the SRE. This has since been significantly redesigned by ONS Data Architecture to improve quality and meet wider requirements across ONS business areas. Sources being used in the construction of the Demographic Index include: NHS Personal Demographic Service (PDS), DWP Customer Information System (CIS), Higher Education Statistics Agency (HESA) Data, and English and Welsh School Censuses (SC). Additional sources may be added to the demographic index depending on the assessment of coverage, and the requirements of other ONS business areas.

The wider requirement across ONS is for the demographic index to provide a comprehensive list of ‘ever-registered’ persons appearing on these data sources. It needs to have high coverage of the population (past and present) and minimal duplications arising from linkage error[[4]](#footnote-4).

All datasets supporting the production of future population and social statistics need to be linked at record level to the demographic index and assigned with an ONS statistical identifier. ONS Data Architecture will undertake all matching exercises associated with the construction of the demographic index, and the referencing of ONS statistical identifiers by linking all sources to the demographic index[[5]](#footnote-5).

In the longer term (provisionally January 2022 onwards) the demographic index will need to be updated and available in DAP on a quarterly basis to support the future methodology for producing population estimates.

**5.1 Admin Based Population Estimates (ABPEs)**

Since the demographic index is designed for the ‘ever-registered’ population, it cannot be used to produce population counts directly. PMST have developed business rules for determining whether to include records on the demographic index in the Admin Based Population Estimates (ABPEs). This is done by linking a range of ‘activity indicators’ to records on the demographic index using ONS statistical identifiers. The business rules continue to develop as new data sources are made available. The most up to date version (ABPE v3.0) has been designed to minimise over-coverage by using a strict set of business rules, i.e. recent evidence of interaction with government services. ABPE v3.0 has considerable under-coverage (approx. 10% of population), and it is still likely that a proportion of the records still represent over-coverage that needs to be adjusted for.

The ABPE is a building block for estimating population size and compiling information to produce detailed estimates of population characteristics. Statistical methods need to be developed to:

* Assess the level of under-coverage and over-coverage on the ABPE to produce population estimates at local area level by age and sex
* Reconcile differences between two consecutive years of mid-year population estimates and net migration estimates
* Supplement the ABPE with additional information from admin, survey and census data to produce key population characteristics that are benchmarked to population totals

Comparisons with 2021 Census estimates will provide key benchmarks for evaluating the quality of a future admin led system. Plans to integrate the use of administrative data to enhance Census estimates raises challenges regarding the circularity of data being used in this comparison. This needs to be taken into consideration as part of success criteria developed for the 2023 recommendation.

SDR are working with the Census and Transformation Policy Team to outline options for the 2023 recommendation. These options include variants of Census taking (for example, online only in 2031) if our evidence suggests that an admin-based system supported with surveys will not meet user requirements. To make a recommendation to an admin-led system without the provision of future censuses, we are currently assuming that we will need to have proven the following:

* Readiness to produce annual estimates of population size with specified levels of quality, and coherent migration statistics from 2024 onwards.
* A methodological framework that demonstrates our capacity to meet user requirements for small area and key multivariate statistics by 2031.
1. **Production of Population and Migration Statistics**

PMST are working with MSD and MDR to deliver an admin-based system for producing population ‘stocks’ (annual mid-year population estimates) as well as estimates of ‘flows’ (including international migration). The current system is largely driven by a flows-based approach (cohort component method), which is rebased with Census estimates of population size (stocks) once every ten years.

**6.1 Estimating Population Stocks**

A major aim of PMST is to deliver annual estimates of population size with consistently high levels of quality over time. While the aim is to continue delivering annual population estimates down to small areas, the overall design is based on meeting key quality standards at local authority level. The methodological design is still being researched through simulations, but the current assumption is that the ABPE needs to be constructed quarterly to support the estimation framework. The estimation framework is based on assessing coverage of the ABPE using the IPACS survey, which provides an independent source of information about individuals that are missing from the ABPE. By collecting IPACS on a continuous basis, annual mid-year population estimates can be produced independently each year.

The approach is similar in principle to methods already established for measuring Census non-response using the Census Coverage Survey (CCS). The estimation framework is being developed around Dual System Estimation (DSE). The estimation framework needs to be adapted to produce a time series of estimates throughout the year with a smoothing method to produce annual mid-year population estimates.

**6.1.1 IPACS requirements**

Research has been undertaken across PMST and SST to integrate admin data coverage survey requirements with redesign of the ONS Labour Force Survey and Annual Population Survey collections. The design for IPACS centres on the proposal to collect a continuous ‘master wave’ questionnaire with an issued sample size of 500,000 households per annum. This sample size is driven by the assumption that we are aiming to produce estimates of population size that have equivalent levels of precision as current Census population estimates. Social Survey Transformation (SST) are in the process of evaluating the cost and viability of running a survey of this scale. This sample size can potentially be reduced if we can demonstrate that the quality of population estimates from the new system is an improvement on mid-year estimates produced in between Censuses.

We do not anticipate response rates to the IPACS master wave to be as high as the CCS. Based on operational testing of mixed mode approaches (online first, face to face and paper), we assume a response rate in the region of 60% can be achieved from the voluntary IPACS master wave.

PMST have developed a stratified sampling approach similar to the CCS that optimises allocation of sample to areas where there is higher prevalence of under and overcoverage on the ABPE. This is equivalent to the CCS Hard-to-Count (HtC) index and requires that the address frame, which accurately identifies all residential addresses is periodically augmented with admin data to draw the sample.

The IPACS master-wave is primarily designed to measure under-coverage on the ABPE. The design PMST is working towards assumes that methods and sources will improve over time to detect and remove all instances of over-coverage from the ABPE.

**6.1.2 Over-coverage on the ABPE**

While PMST continue researching methods for detecting over-coverage, there is still a risk that residual over-coverage will continue to persist on the ABPE. This would require that the IPACS adjustment for under-coverage will also need to be supplemented with an additional over-coverage adjustment. This might be achieved through model-based estimation, or the collection of additional data specifically aimed at measuring over-coverage. An initial proposal to test a survey based on ‘dependent interviewing’ (verifying admin records in the field) was rejected following advice from NSDEC[[6]](#footnote-6).

**6.1.3 Statistical testing**

The first statistical test for producing IPACS adjusted admin-based population estimates is intended as part of quality assurance for the 2021 Census. PMST and MDR have recommended use of CCS data to support production of admin-based estimates on Census day. This would enable the start date for collection of IPACS to be deferred until July 2021. In order to demonstrate readiness to produce admin-based population estimates from reference year 2022 onwards, this would be the latest start date for collection of IPACS data, as a 12-month time series of IPACS data is required to support the methodology. There will not be sufficient time to produce IPACS adjusted admin-based population estimates for reference year 2023 in time for the recommendation.

To ensure successful development of a new system to estimate population size, SDR require the following to be in place. All of these are being progressed with the necessary business areas, and processes are in place to capture additional requirements as they emerge.

**Table 1: Requirements to support the statistical design for estimates of population size**

|  |  |  |
| --- | --- | --- |
| **Requirement** | **Purpose** | **Owner(s)** |
| Indexing of all person level datasets used to estimate population size (Mar 2020) | To support ongoing development of ABPEs including new data sources only available in DAP | Data Architecture |
| Quarterly updating of the Demographic Index in DAP from Mar 2020 onwards | To support the time series methodology for producing pop estimates under the new system.  | Data Architecture |
| Estimation framework for producing Census day population estimates from ABPE and CCS  | To provide independent Census day estimates to support national adjustment. | Methodology |
| IPACS sample design (wave 1) for local authority estimates | To deliver robust population estimates at local authority level (ref mid-year 2022) | Methodology |
| Commence IPACS collection (wave 1), starting July 2021, 30K addresses sampled per quarter up to end of 2021. | To collecting data for time series modelling and develop iterations of the stratification index. | SST |
| Increase scale of IPACS collection to 60K per quarter throughout 2022, and 125K per quarter from 2023 onwards | To collect data on sufficient scale to produce LA level population estimates for ref year 2022, and LA estimates by age sex for ref year 2023 | SST |
| Systems in place for processing IPACS data in DAP (July 2021), as well as in-house systems for collecting IPACS data in the longer term. | To enable IPACS data to be processed and linked for production of population estimates | Census Processing/SST/DST |
| Method for linking IPACS to ABPE (inc. clerical matching), and optimised collection of IPACS data fields (names, dates of birth) to improve linkage quality. | Improve existing methods to minimise errors and meet DSE assumptions | Methodology/PMST |
| Cognitive testing of IPACS questions (inc. any questions designed to collect data on over-coverage)  | Finalise questions to be included in wave 1 of IPACS | SST/Methodology |
| Acquisition of datasets to improve the quality of ABPEs (i.e. address specific activity – Royal Mail, DVLA, Electoral Register, Council Tax). | To improve methods for detecting and removing over-coverage from the ABPE | DaaS |
| Further research into model-based approaches for estimating residual over-coverage on ABPEs (longer term) | Additional adjustments to future population estimates may be required if over-coverage cannot be removed from the ABPE | Methodology |

**`6.2 Estimating Population Flows**

PMST are working with Migration Statistics Division (MSD) to deliver a method that places administrative data at the core of migrations statistics. Key to the design is ensuring that annual estimates of net international migration (when combined with births and deaths) are coherent with the differences observed between two consecutive years of population stock estimates.

In the short term, MSD are focusing on incorporating the use of admin data to make improvements to the existing system for producing migration statistics, which is primarily based on International Passenger Survey (IPS) data. Some adjustments have already been made to experimental statistics using DWP and Home Office Exit Checks data. Work is ongoing in MSD to publish a back series using this method and incorporate changes into the next round of mid-year population estimates.

In the longer term PMST are considering how to design a system that produces coherent population stocks and flows. The ideal admin-based system would enable changes at record level to be observed on the ABPE over two consecutive years, including capture of all migration activity into and out of the country. In some countries with register-based censuses, the population statistics system is driven entirely by admin data. Population registers are believed to have timely and complete coverage of the population, including border flows.

In the UK context, there are additional complexities underlying an admin-based system. First is the need to assess the ABPE for coverage error (as described in 6.1), and second is the difficulty of correctly identifying migrants, even for those that do appear on the ABPE. In earlier publications, PMST have outlined three options for producing coherent estimates of population stocks and flows:

* A ‘stock-driven’ system - international migration estimates are constrained to the difference between two consecutive population stocks.
* A ‘flows-driven’ system – population size is estimated each year by rolling forward births, deaths and migration estimates. This is similar to the current cohort component method but underpinned with record-level data.
* A hybrid option - estimates for stocks and flows are derived from the ABPE/surveys independently and statistical methods are used to reconcile inconsistencies between the two.

The consensus view from initial engagement with users (RSS roundtable events, Census roadshows) is that ONS should progress research on the hybrid option to produce the most accurate statistics for population stocks and flows. A method for producing coherent population stocks and flows needs to be in place by 2023, allowing for at least two consecutive mid-year population estimates (reference years 2022 and 2023). The following requirements need to be in place to support research to deliver admin-based migration statistics within the new statistical system. These requirements are all captured in MDR’s plans and being progressed as far as possible.

**Table 2: Requirements to support the statistical design for migration estimates in the new system**

|  |  |  |
| --- | --- | --- |
| Requirement | Purpose | Owner(s) |
| Acquisition of key admin datasets capturing migration activity on the ABPE (Home Office Exit Checks and Semaphore, DWP RAPID system) | These sources need to be thoroughly researched to assess whether they are of sufficient quality to support a system that places admin data at the core of international migration statistics. | DaaS |
| Methods that measure uncertainty in international migration estimates. | Measures of uncertainty are needed to meet user needs and support the method for reconciling estimates of population stocks and population flows.  | Methodology |
| A method that reconciles differences in net migration estimates when comparing two consecutive years of population stocks.  | To ensure that the population statistics system is coherent and fit for purpose. | PMST/MSD/Methods |

**6.3 Supporting 2021 Census Population Estimates in scenarios of low response**

The standard design for producing Census day population estimates relies on adjusting for non-response using the Census Coverage Survey (CCS). While the DSE framework is well-established, it may be necessary to make a similar national adjustment, as was the case in 2011, using methods outside of the standard design. ABPE estimates that have been assessed for coverage (as described in 6.1.3) will provide an independent source for quality assuring provisional Census estimates and have potential to enhance the strategy for making any national adjustments. To assess coverage of the ABPE on Census day we are currently proposing to use the CCS, as collection of IPACS data on a sufficiently large scale is not planned to start until July 2021. Methodological assurance regarding use of the CCS to assess coverage of the Census and the ABPE is being sought from Southampton University.

SDR are also planning for a range of scenarios including that 2021 Census response rates fall below a level that would prevent the standard design from producing reliable population estimates. This occurred in the New Zealand 2018 Census, where methods had to be developed to supplement census counts with administrative records.

* + 1. **Potential Low response scenarios**

CSD have identified potential scenarios for low response in 2021:

* Localised geographical area issue, for example missing a care home or student hall, flooding, foot & mouth outbreak
* Population group specific issue, for example young men not captured by either Census or CCS, population subgroups boycotting Census, persons more likely to be missed within households (e.g. babies)
* Low response rate overall, not achieving the critical success factors across the country
* Issues resulting from design/collection, for example fake returns due to authentication approach, individual forms returned without household forms, not enough paper anticipated if there’s an online security scare

In some of these scenarios it may be possible to rely on ABPE estimates directly as a substitute for Census data. For example, in cases where care homes or student halls are entirely missing from the Census collection, it may be preferable to recover the missing data directly from admin records. In scenarios where non-response is more prevalent across the board however, the standard design may not be able to meet critical success factors for high quality population estimates. It is in these scenarios that additional research needs to be undertaken to enable census counts to be supplemented with administrative records to support the estimation framework.

Due to the prevalence of over-coverage the ABPEs cannot be relied upon as a substitute data source for counting persons in households that have not responded to Census. While the rules used to produce ABPE v3.0 have focused on activity data to confirm usual residency, there is limited evidence to confirm current address.

PMST are working with Census Statistical Design (CSD) to pre-populate the Census Address Frame with administrative data, using an ‘address-centric’ approach. This will be based on longitudinal information about residents registering at the address over time, relationships between residents, and evidence of activity associating individuals with the address. The aim is to produce a probabilistic model that accurately predicts likelihood that individuals are resident in non-responding addresses. The model will be assessed using Census rehearsal data and will only be used to support Census population estimates if response rates are too low to be supported using the standard design.

To ensure that necessary preparations are in place for low response scenarios in 2021, SDR require the following:

**Table 3: Requirements to support statistical design in scenarios of low Census response**

|  |  |  |
| --- | --- | --- |
| **Requirement** | **Purpose** | **Owner(s)** |
| ABPEs (v2.0) referenced to January 2021 and available in advance of Census collection | These estimates will be used to check the plausibility of Census response rates and detect areas on low response for targeted follow up | PMST / Data Architecture |
| Estimation framework for producing Census day population estimates from ABPE (v3.1) and CCS data | Produce independent Census day estimates at national level (by age and sex), to support national adjustment. | Methods |
| Methods for linking ABPE v3.1 to CCS records, including clerical matching and the need to measure quality of linkage to hashed DWP data.  | Quality of linkage for DSE needs to be improved on what has previously been delivered in the SRE. | Methods/SDR |
| Development of a Census address frame pre-populated with admin records, including likelihood of residency at the address. | For potential use in blended census / admin counts if response rates to Census are lower than necessary to support the standard design.  | PMST/CSD/PSD |
| An estimation framework that can incorporate the use of admin records to supplement Census counts in population estimation  | A fully researched and tested method needs to be ready for implementation if the standard design for Census population estimates fails to deliver required quality. | SDR/Methods |

1. **Household Statistics and Communal Establishments (CEs)**

Censuses and social surveys are largely designed as household collections. Similar to the framework for producing population estimates, the standard census design relies on DSE to estimate numbers of households for local authorities. Additional statistical adjustments include imputation of missing relationships between household residents. This enables estimates of household size and composition to be produced at small area level.

A future population and social statistics system driven by admin data will have limited access to information that has been collected on households. In recent years PMST have produced a series of admin-based household statistics based on an alternative definition of ‘occupied address’. These include:

* Estimates for number of occupied addresses[[7]](#footnote-7) – A DSE approach based on ABPE and survey counts. IPACS is a key data source to adjust for occupied addresses that are missing from the ABPE.
* Household size – estimates of the number of occupants in addresses have been produced by modelling ABPE counts with distributions from the Annual Population Survey (APS). This method was based on Structure Preserving Estimation (SPREE).
* Household composition – relationships between residents in addresses on the ABPE have been derived from birth records and benefit claims. Where information was unavailable, linked APS data was used to impute relationships.

These admin-based household statistics are broadly comparable with those produced from Census and survey estimates. Where APS data has been used to improve on ABPE estimates, we intend to use IPACS data for similar adjustments in future.

Due to limited information about relationships between residents on admin data, PMST have not yet been able to produce estimates of household compositions across all categories currently produced in Census outputs. PMST are in the process of developing a ‘relationships matrix’ which supplements admin records with 2011 Census data. This will maintain a record of relationships over time, minimalizing the number of relationships that need to be imputed on the ABPE. Further engagement with users is needed to establish the level of support for changing from traditional ‘household’ definitions to alternatives used in some register-based countries. An example is Finland’s *household-dwelling unit.*

**7.1 Communal Establishments**

Research on producing admin-based statistics on populations in communal establishments (CEs) is in very early stages of development. We anticipate that the RDMF address index will be able to identify large sized CEs on the ABPE. There will still be the need to accurately classify smaller CEs by type and estimate for CEs that are missing from the ABPE. IPACS is not currently being designed to collect data from CEs. A strategy for measuring the coverage and accuracy of CEs on the ABPE is needed to support the future system. Reviewing approaches taken in other countries to undertake surveys in CEs may offer a useful starting point, as well as our own experiences of using admin data for large CE’s (i.e. 50+ bed spaces) for the 2021 Census. A detailed proposal for producing statistics on CEs is unlikely to be available ahead of the 2023 recommendation, however we should be able outline options for meeting user requirements in future.

**Table 4: Design requirements for household and CE statistics**

|  |  |  |
| --- | --- | --- |
| **Requirement** | **Purpose** | **Owner(s)** |
| User engagement regarding alternative household definitions | Need to understand what users perceive to be losing from a change in definition from a policy and decision-making perspective, e.g. overcrowding analysis. Also, implications of a break in time series from current statistics produced from Census, LFS and APS data. | PMST / PPA |
| Establish how the relationships matrix being developed in PMST is incorporated into ONS data architecture | Historic relationships between family members need to be stored in the ONS data architecture to support statistics on household composition and inter-generational research. | Data Architecture |
| Assess the accuracy of the address index to identify and classify CEs  | Need to show users the potential for providing information on CE populations in future. Previous research that concluded low potential for identifying CE’s on address frames needs to be updated since the introduction of AddressBase. | Data Architecture/PMST/SST |
| Propose (and test?) methods for measuring coverage of CEs on ABPE (IPACS submodule or separate collection) | Options for assessing the coverage of CE populations on the ABPE need to be outlined to users, or possibly tested to demonstrate the potential for producing statistics on CE populations from the new system. | PMST/SST |
| Review of methods developed for ABPE household statistics, including: number of occupied addresses, household size and household composition.  | Research outputs on admin-based household statistics have been de-prioritised over the last few years to concentrate on population. A review of methods is needed to determine whether progress at this stage is satisfactory to be used as our evidence in support of the recommendation in 2023.  | Methods |

1. **Population Characteristics**

**8.1 Estimating Key Characteristics in the Future Population and Social Statistics System**

2021 Census outputs will include the use of administrative data to produce integrated outputs for certain topics. Examples include income, number of rooms, and data on armed forces personnel to support outputs on veterans. Methods to deliver these outputs are at an advanced stage of development so are not discussed in this paper.

Beyond 2021, the future design of population characteristics requires a framework for producing modelled estimates from combined data sources. The future system will be required to continue delivering traditional Census topics, as well as expand the range of topics to meet wider user needs.

One of the key challenges for SDR is ascertaining the extent to which we will be able to meet existing user requirements for small area, multivariate statistics. The census is unique for providing granular and detailed statistics, however they are only provided once every ten years. Some key topics are produced from the LFS and APS collections in between Censuses and multivariate analysis is carried out by some users of microdata. However, due to sample sizes ONS published statistics are largely univariate and lack the required levels of precision below regional level.

**8.1.1 Developing the ABPE for Population Characteristics**

The new system for social statistics will aim to exploit data from multiple sources by integrating data via the Demographic Index. The availability of population characteristics on admin data can broadly be categorised into three types:

1. Topics with high coverage of population, with potential to use directly in official statistics. An example is ONS experimental Admin Based Income Statistics (ABIS)
2. Topics with partial coverage of population or definitional inconsistency with the target variable of interest. The majority of key characteristics currently fall into this category, including ethnicity, nationality, education, employment and health.
3. Topics with minimal or zero coverage on admin data, and where information needs to be collected primarily from a survey (IPACS). Examples include carers, religion, language, and hours worked.

At this stage, the range of topics from which direct estimates could be produced from admin data are very few. Research outputs have been produced on income, number of rooms, students in employment, and educational attainment of younger adults. These publications have all drawn attention to limitations with the admin data estimates, for example the unavailability of certain components to fully meet definitions, or limited coverage of certain age groups.

**8.1.2 Methods for producing population characteristics in the future system**

One of the recommendations of the Skinner Review of Beyond 2011 Methods was that admin data be explored as auxiliary information to support estimation with an ongoing characteristics survey. The IPACS master wave has been designed to sample approximately 2% of address annually and will include key socio-demographic topics currently collected from the Census[[8]](#footnote-8). A significant design challenge for SSAF and MDR is to develop a statistical framework for producing estimates on population characteristics from combined sources. A number of options are being considered, which are not mutually exclusive:

* Modelled estimates based on aggregated statistics from admin and survey sources. Examples include structure preserving estimators (SPREE) and small area estimation (SAE)
* Record level imputation of characteristics onto the ABPE. Examples include donor imputation or machine learning from linked survey data
* Linking characteristics from Censuses (2011 and 2021) to the ABPE and longitudinally rolling forward. Similar approaches to those described above could potentially be used to produce modelled characteristics when combined with a survey

The datasets generated in the production of population characteristics need to be stored in the ONS data architecture to support wider analysis projects undertaken ONS. It has not yet been established whether a fully imputed ABPE (similar to the adjusted Census dataset) is viable or even necessary to meet user requirements for population and social statistics. A dataset with this level of utility would provide maximum flexibility for small area multivariate tabulations, however the complexity of statistical adjustments required to produce such a dataset would require long term development beyond the 2023 recommendation.

For the 2023 recommendation, ONS will need to demonstrate that methods are in place to calibrate population characteristics to estimates of population size. Any adjustments for missingness or definitional inconsistency of topics on the ABPE will also need to take into account adjustments that have being made for coverage error (under-coverage and over-coverage).

**Table 5: Requirements for future population characteristics**

|  |  |  |
| --- | --- | --- |
| **Requirement** | **Purpose** | **Owner(s)** |
| Agree key topics to produce univariate estimates for (ref year 2022).  | Supporting evidence for the 2023 recommendation, proposed as part of success criteria for the future pop stats system | SSAF/PPA |
| Agree key tabulations to produce multivariate estimates for (ref year 2023).  | Supporting evidence for the 2023 recommendation, proposed as part of success criteria for the future pop stats system | SSAF/PPA |
| Review content of IPACS (wave 1 and longitudinal) and identify new requirements for additional questions or adaptations  | Ensure that key topics are included in the survey and determine whether they should be captured in wave 1 or longitudinal waves.  | SSAF/SST/PPA |
| Collect key characteristics data from IPACS (Start collection in July 2021, 60K per quarter throughout 2022 and 125K per quarter in 2023) | Provides data to detect change in population characteristics since 2021 Census so that estimates for population characteristics can be updated. | SSAF/Methodology |
| Develop research methods for producing univariate estimates for agreed topics using a combination of survey, admin and census data  | Need to show that we have methodological framework for combining survey, Census and admin data to produce estimates of characteristics for key topics | Methodology |
| Establish levels of precision that can be produced for different levels of geography when characteristics are derived from combined IPACS/ABPE data  | Quality targets and an indication of our ability to meet these quality targets for population characteristics need to be provided for users | Methodology |
| Outline future options for producing multivariate / fully imputed ABPEs, supported with simulation studies. | Yet to determine whether a fully imputed ‘utility’ dataset is viable for an admin data Census. But need to demonstrate that there is a framework for producing multivariate outputs. | Methodology |

**8.2 Analytical products and wider analysis**

A population and social statistics system based on administrative data will enhance capability for wider research and analysis. Although it isn’t yet clear whether a fully adjusted ABPE for multivariate analysis (similar to the adjusted Census dataset) can be produced annually, we currently envisage there being two types of analytical datasets to be available for research.

1. Statistically adjusted datasets designed for analysis of published estimates. These datasets will include coverage weights from IPACS and possibly item level imputations to ensure that aggregations on topics correspond with estimates of population size and international migration. In the short term we are likely to produce a series of smaller bespoke datasets that allow tabulation across key topics.
2. Unadjusted datasets based on longitudinally linked data. These datasets are intended for wider analytical purposes, when it is preferable to work with data that has not been statistically adjusted. Examples include cross sectional linkage studies and inter-generational analysis, where there are often requirements to integrate data from multiple sources. The Demographic Index and the assignment of ONS statistical identifiers will support construction of these datasets, which can be undertaken efficiently and flexibly, and when necessary temporarily.
3. **Privacy and Security**

The datasets described in 7.2 (both adjusted and unadjusted) will only include de-identified records. The RDMF will enable linkage to be undertaken flexibly with ONS statistical identifiers, allowing person identifying information (names, dates of birth, addresses) to be removed. In practice, the statistical design has considerable scope for integrating a wide range of data sources, so the risk of secondary disclosure, i.e. individuals being identifiable by a unique combination of characteristics needs to be managed.

A system for managing secondary disclosure needs to be in place to support the future statistical design. Approaches adopted in other NSIs based on segmentation of data attributes should be considered. These methods allow multiple datasets to be held together temporarily for statistical processing, before disassembling attributes across the data architecture. Researchers can then request access to necessary variables of interest to undertake research. Depending on the level of sensitivity in the data, decisions can be made regarding the combination of variables and geographic granularity to be made available.

Designing the system for holding analytical variables in the ONS data architecture is outside the scope of SDR, but we will need to work with ONS Data Architecture to ensure that there is an appropriate balance between statistical utility and privacy. Methods for Statistical Disclosure Control (SDC) will need to adapt for a system based on multiple sources, possibly bringing new risks that are not yet fully understood (for example reconstruction attacks). Our research on SDC will evolve as we progress towards research outputs on multivariate statistics. We will also continue to take advice from the National Statistician’s Data Ethics Committee.

1. **Governance of the statistical design**

This paper highlights many of the key dependencies and methodological requirements needed to support the design of the future of population and social statistics system. Given the range of transformation projects in progress across ONS there is a need to make decisions regarding the best use of resource in the time available between now and the recommendation in 2023.

This paper will be discussed at the PPI Board, PPP Design Authority, and CDCTP Board to ensure coherence in future statistical design across the office. The methodological aspects will also be reviewed by the Census External Assurance Panel.

We will also establish a clearer view of the statistical design at UK level, working with Devolved Administrations through the UK Population and Censuses Strategic Group. While we acknowledge that there may be differences in the methods and timeline for implementing admin-based approaches, we need to ensure that statistics produced across UK countries remain coherent. Part of this includes establishing the ongoing funding model, particularly the rollout of the IPACS collection from July 2021 onwards.

1. Quality standards for the 2021 Census: 95% confidence intervals, +/- 3% for local authorities. 95% confidence intervals, +/- 0.2% for the national estimate. [↑](#footnote-ref-1)
2. Proxy versions in RDMF v0.1 were based on some of the key sources, but not the complete set of sources needed to construct each index. [↑](#footnote-ref-2)
3. The SDR Quality and Requirements Team have detailed documentation on specific data sources and when they are needed. [↑](#footnote-ref-3)
4. Current quality targets for linkage are for false positives (incorrect links) to be estimable below 0.5%, and false negatives (missed links) to be estimable below 2%. [↑](#footnote-ref-4)
5. SDR are in the process of specifying data linkage requirements to support research in PMST, SSAF and CSD. [↑](#footnote-ref-5)
6. Dependent interviewing has been used in countries that have successfully transitioned to admin-led censuses despite the prevalence of over-coverage on population registers (e.g. Italy and Israel) [↑](#footnote-ref-6)
7. The definition of ‘address’ requires further clarification. Unique Property Reference Numbers (UPRNs) have been used to identify individual addresses, however there is limited granularity for multi-occupancy dwellings. [↑](#footnote-ref-7)
8. The IPACS Wave 1 that has been developed for testing includes age, sex, country of birth, ethnicity, religion, marital status, highest education level, employment status, and disability. [↑](#footnote-ref-8)