Methods for producing multivariate population statistics using administrative and survey sources

MQD Small Area Estimation and Imputation teams

November 2022

Purpose

This paper provides an outline for the programme of methodological work to produce multivariate population outputs which are primarily based on administrative data but use survey and other data sources to provide robust outputs that account for missingness and other data problems. To provide multivariate outputs for detailed geographies a combination of methods will be required to overcome the different data challenges; the outline summary in Appendix 1 provides an overview of the research work that is currently underway to address this and demonstrates the overall strategy. This paper provides more narrative around this and a basis for discussion. This work sits alongside the programme of work to produce administrative based statistics on population characteristics and will inform the 2023 recommendation to Government on the future of population statistics.

The different stages of the strategy at a high level are to:

1. Improve the quality of the linked (integrated) administrative dataset at unit level
2. Explore estimation methods that make use of independent survey data to reduce the bias in the administrative (admin)-based outputs. Outputs from these models would provide aggregate level outputs with measures of uncertainty
3. Blend unit level and aggregate level data to allow estimation of more detailed output variables.

Questions:

- Does the panel support the overall direction of this methodological strategy?
- Are there any other methods that we should build into this?
- Does the panel have any advice about extending the imputation and small area estimation methods for multivariate outputs?
- We welcome your thoughts about the objective to create an improved unit level data set (that accounts for missingness) as opposed to other approaches, and what alternative approaches might look like.

1. Introduction

ONS has made considerable progress in developing new methods to create population estimates from administrative data to reduce reliance on the decennial Census. Estimates for the size of the population in England and Wales, for 2011-2016[1] and 2016-2020[2] have been published as research outputs with improved versions designed to better reflect the coverage of the target population. Research has also combined administrative data to produce census-type information about the characteristics of the population on a range of topics but focusing particularly on income, ethnicity, and housing.

Characteristic outputs to date have mostly focused on the coverage of the combined sources, for example the number of people on the linked administrative based population dataset who have information on the target characteristic. Although coverage is promising the data are not complete and thus cannot currently be used alone to obtain robust counts.

Research in ONS has been focused on investigating the properties of the datasets and understanding the data journey.[3] There is still more to do but this helps to inform the development of methods to account for the missingness and to explain differences between the outputs from the administrative datasets and census or survey outputs.

The user requirement is for detailed multivariate characteristic outputs by small geographies, for example individual personal income by detailed ethnic group (16 categories), by age and sex at Lower Layer Outputs Area level. ONS has considered two case studies for its research: ethnic group by income and ethnic group by a housing characteristic. No one method will achieve these outputs but a strategy using a combination of methods may help to resolve some of the data challenges.

This paper sets out a methodological strategy to improve outputs from the integrated administrative dataset by using survey data. Although the intention is to discuss methods that could potentially be used for different multivariate characteristic outputs, it specifically considers the case study on Ethnicity by Income to demonstrate the data issues encountered (primarily missingness) and to give context for the methods that are suggested to overcome these. It considers first, the feasibility of imputation methods to improve the unit level data set, and secondly how small area estimation methods that draw strength from survey data can be used to provide aggregate level outputs including how they may be extended for the multivariate outputs. Finally, it suggests an approach for blending the unit level data set with higher level estimates to achieve the detailed outputs to meet user needs.

Although the overall approach is to achieve a unit level data set it is not intended that this would be an output, but a means of obtaining consistent aggregate level outputs. It is envisaged this might be for key multivariate outputs rather than a full census type dataset. This methodological work strand aims to investigate what can be achieved with the different methods for different topic areas in the administrative data context and to better understand their limitations.

Section 2 provides a high-level description of the integrated administrative dataset for the case study on ethnic group by income and highlights the extent and nature of the missing data. This provides some background information about the challenges that we seek to resolve.

2. **The integrated administrative datasets for income and ethnicity**

Data have been integrated with the ambition of producing admin-based income by ethnicity statistics (ABIES) for the tax year ending 2016. The dataset combines the admin-based income statistics (ABIS)[4] dataset with version 2 of the admin-based ethnicity statistics (ABES)[5] dataset. An initial assessment of the combined multivariate dataset has been published: Feasibility research on developing subnational multivariate income by ethnicity statistics from administrative data for England)[6].

The ABIS and ABES datasets are both derived from multiple administrative data sources, which have been linked together to produce statistics about the single topic areas, ethnicity, and income. Both datasets use the Statistical Population Dataset (SPD) v3.0 (previously known as the ABPE's) as a population spine and are joined together based on a unique identifier to create the admin-based income by ethnicity statistics (ABIES). The ABIES integrated dataset includes individuals who are resident within England (only) and if there were no missingness or error would allow (anonymised) person level analysis of individual incomes by ethnicity at low level geographies by age and sex.

The publication referenced above explores the coverage of the dataset by age, sex, region, and Local Authority level (e.g., proportion of people with ethnicity and income data).

Of the 43.9 million individuals aged 16 years and over in the SPD v3.0

- 77.0% have income information and a stated ethnicity
- 15.6% have income information but no stated ethnicity
- 5.6% have a stated ethnicity but no income information
- 1.8% have neither income information nor a stated ethnicity

The proportion of individuals with both income and ethnicity data differs by region and age.

Ethnicity Missingness in SPD

Of the 43.9m records aged 16+ in SPD 3.1, 17.4% have no stated ethnicity data. Coverage varies by age, sex, and LA (e.g., 17-26 and 28-49 males are underrepresented). Based on comparisons with 2011 Census, Asian, Black, and Mixed ethnic groups aged 20-34 were underrepresented. Also, the proportion of people in the Other ethnic group was higher than expected compared to 2011 Census.

Sources of missingness consist of:

- Missed links between the ethnicity data set and the SPD
- Cases where a record exists within the ethnicity data, but ethnic group is missing or refused
- The record does not exist within the ethnicity dataset so can't join to SPD e.g., people who infrequently interact with health/education services

Income Missingness in SPD

Of the 43.9m records aged 16+ in SPD 3.0, 7.4% have no income data. Coverage varies by age – it is lower for young people aged 16 to 20 years, but from age 25 years, over 90% of individuals, for both sexes, have some income information in ABIS.

Sources of missingness:

- Missed links between income data and SPD. It may be reasonable to assume some of these individuals have zero income though as they do not have a National Insurance Number. They may be in education, not working or potentially migrants not working etc. However, there may be some missing for other reasons too.
- The record is included in the income dataset, but some (or all) income information is missing.
- Record is included in the income dataset, but their income is on data sources not yet used in construction of ABIS (see Appendix 2 for details). This can lead to underestimation of income statistics as well as bias due to missingness.
- Record not in the income dataset so can't join to SPD. Examples include students who have left education but not working, income from benefit components not in ABIS, informal payments, low incomes below PAYE threshold.

Coverage of the population spine (SPD v3)

The summary above represents missingness of characteristic data for individuals who are represented in the SPD, there is also under coverage of the SPD, however. The publication Admin-based population estimates and statistical uncertainty: July 2020[7] provides an overview of the quality including coverage of the population target. The SPD v3.0 for England in 2016 contains around 43.9 million individuals aged 16 years and over. Previous analysis of the coverage of the population base[8] has shown that even after accounting for possible linkage error, there are still some gaps in SPD coverage, particularly for age groups where we do not yet have access to sources of activity data, for example, the self-employed or the working-age population that is neither working nor receiving a benefit. These coverage issues will also have an impact on the coverage of the ABIES data.

Appendix 3 discusses the linkage of datasets for the SPD and mechanisms by which linkage error results in under coverage of the target population. The linkage errors and also errors in applying “activity-based” inclusion rules to “trim” the dataset (to better replicate the usually resident population) means that the initial SPD has various coverage issues before linkage to other sources (such as income and ethnicity data). The linkage errors between the administrative sources are unlikely to occur randomly across population sub-groups, meaning that there will be linkage bias in the SPD.

Implication of missingness for multivariate outputs

A consideration in producing multivariate outputs is the association between the single variables in the admin dataset and the extent to which this represents the association in the target population. To obtain robust multivariate outputs, the methods used to account for missingness will need to account for the impact of missingness on this relationship as well as the distributions of missingness across the univariate categories. This issue may be exacerbated when combining different input sources and the variance in the missingness between areas (or the categories in the second variable) is inconsistent.

3. Methods to resolve missingness and conflicting records

The approach taken is to first consider how we might improve the integrated admin-based dataset at the unit level. If complete and accurate, this provides the most detailed resource to provide multivariate outputs on population characteristics.

3.1 Imputation methods

Imputation has been used in ONS for both Census and survey outputs to account for item missingness where there is a record for the individual in the dataset, but characteristic or variable information is missing. There are sizeable challenges for the administrative data sets, however. This section demonstrates some work undertaken to investigate the potential of imputation for the admin-based outputs on income and ethnicity and to understand the limitations. It also considers what would be required to obtain imputed values that could be used for multivariate outputs.

The imputation methods aim to substitute missing cases based on calculations or estimations. Methods range from simply substituting the mean, median or mode value of the observed records, to model-based imputation where the value of the missing variable is regressed on other observed variables and based on observed coefficients. Where the missingness is non-ignorable the imputation could adjust the data for that non-random missingness mechanism through the covariate data. The more complex methods provide more reliable imputed values but require covariate information related to the missingness in the target variable i.e., data are missing at random once covariate information is considered. Single imputation methods, such as mean imputation, ignore uncertainty and almost always underestimate the variance. Multiple imputation approaches can overcome this problem, by considering both within-imputation uncertainty and between-imputation uncertainty.

For the administrative datasets, missingness occurs for several specific reasons as outlined in Section 2, and the nature of the missingness and availability of covariate information will be different accordingly. The potential for development of imputation methods would likely need to be specific to the data challenges for that component of missingness.

This programme of work has so far considered the potential of imputation methods for the administrative income data ABIS. In this example, one significant challenge has been to identify which values are missing/incorrect in ABIS. That is, records in ABIS are allocated a £0 where that record does not have an income value noted for a specific data source. However, we do not know which of these values are genuine £0 or are values that require imputation. Instead, we have focused on imputing missing values that have been identified via implausible relationships between variables. For example, there are a group of records in ABIS who receive tax credits, either working, child, or both. For individuals who receive working tax credits, a requirement for individuals to receive working tax credits is that they should work a minimum number of hours per week. The assumption, therefore, would be that these individuals will receive an income (most likely from PAYE or self-assessment) for this work. Investigating the ABIS suggests that of the 5,990,217 records with tax credits (both working and/or child components), a maximum of 1,061,301 records (approx. 18% of tax credit recipients, approx. 2% of records aged 16+) do not have either PAYE or self-assessment income. The true count of missing incomes will be lower than this amount because we cannot currently separate records who claim working versus child tax credit component. The assumption we make on identifying missing incomes applies only to those receiving working tax credits, therefore anyone receiving just child tax credits should be left with a £0 income value. Work is currently in progress to investigate other characteristic information (that we have data for) that might help to inform the imputations.

The main challenges for this application have been:

- How much of the missingness in the income data set described in Section 2 above can be addressed with imputation?
- A lack of clarity about what values are missing within the datasets and the process by which this has occurred. This has demonstrated the need to work with teams to explore how missingness has been identified and treated throughout the data journeys of the individual administrative data sources to better identify values that are believed to be missing or incorrect/implausible through metadata that contains flags.
- Lack of covariate information in the dataset to inform the imputed values. We will investigate the potential to increase the scope of variables in the administrative dataset via access to more income components, or by linking other administrative data characteristics from ethnicity, qualifications, housing characteristics and so on.

Table 1 below provides a broader summary of some of the challenges and considerations of using imputation methods for administrative data sets.

Table 1. Imputation for Administrative based income: challenges and considerations

	What do we know?	Challenges / considerations
Missing At Random (MAR) assumption	Widely used methods rely on the MAR assumption (where the probability of any data-item being missing depends on variables in the data set where there is complete information). This is not going to be the case in many admin data scenarios. For example, where missing data occur from: • Missing links • Refusals based on outcome of interest	Not able to determine the missingness mechanism and prove MAR or Missing Not at Random (MNAR - Where the probability of any data-item being missing depends on the values in the variable itself or something not captured in the dataset). What methods can we use if the data are MNAR?

Linkage of admin data	For missing data caused by missed links it is harder to justify the MAR assumption Bias may be introduced as a result of linkage error. Coverage bias may occur in one of the linked datasets (e.g., systematic exclusions, which are harder to identify in admin data sources) Erroneous links (i.e., false positives)	Each linked data set has unknown linkage accuracy and coverage • Need generic principles for scenarios
Sufficient covariates	As with any non-simple imputation method, are there the right covariates that predict the missing data in the same admin data source or integrated data set. We can link additional data sources to increase the availability of covariates but does this cause linkage bias issues?	Will depend on the admin data source. May need to make sure there are generic principles for the methods when advising the sourcing of admin data.
High levels of missingness	High levels of missingness will result in more difficult imputation and higher levels of uncertainty in results	We do not know what missingness looks like in all admin data sources One option may be to simulate missingness in order to test methods?
Analytical aims	The 'Best' imputation model is dependent on the analytical aims It is likely that we will not be able to assume a one-size-fits-all approach	Do not know what all analytical aims of admin data might be – imputation approach for an admin data source could be different if interested in fitting prediction models for identifying associations of interest versus attempting to estimate total population size. Is it appropriate to use different methods on the same data for different analytical aims, or should we aim to deal with missingness in one approach (like Census) such that all analysts apply analysis to same data? The risk then involves the data being seen as the 'truth'.
Imputation variance and error	May be introduced into data when applying imputation	How to measure it depends on imputation method applied and this may be complicated where there are various stages of imputation. Do we need to report to users? How do we do this?

Maintenance of multivariate relationships	Need to maintain multivariate relationships	Can be difficult if edit constraints also applied so depends on method applied Imputation may occur in different stages
Stage of edit and imputation (E&I)	It is possible to conduct E&I pre- or post-integration of data sources	What are the impacts of doing each way? Will the impact always be the same?
Properties of the imputed dataset are more difficult to understand.	Imputed values are not 'the true, correct values' or equivalent to the observed values. The impact on variance is not fully understood	How do we communicate the implications of E&I in the admin data? We will need to produce methods papers with each publication

Another project, the Admin-Based Housing Stock (ABHS) brings together admin data sources for housing. It includes data from the Valuation Office Agency (VOA) property attribute dataset, geographical variables from the Census Address Frame and information on household size from the Admin-Based Household Estimates (ABHEs). The VOA dataset includes many household characteristics, such as property type, number of rooms, and number of bedrooms. Despite being an overall good source of information on the household attributes, the VOA variables in the ABHS dataset contain missing values. Where these values are missing, typically all the variables in the VOA data are missing for a unit, which causes significant issues when attempting to use imputation methods that require matching variables. There are different reasons for the missingness: missing cases can arise due to linkage failure, insufficient coverage of data sources to capture the target population, and missing attributes for individual records in the linked dataset. Work is on-going to explore whether data from Energy Performance Certificate's (EPC's) can provide data that resolves missingness on the ABHS data.

Beyond the imputation methods being considered in the ABIS and ABHS projects (mean, regression, nearest neighbour donor imputation, MICE, machine learning), the next couple of sections outline some newer methods that we would also like to explore, which don't exhaust the list of potential options, but gives an idea of the breadth of potential research that is required in this space.

Multiple imputation and latent class analysis (MILC)

When using administrative data for research, we must be aware that they are collected for administrative purposes so they may not align conceptually with the target definitions and so as well as missing data, administrative data can contain classification errors. These may be due to mistakes made when entering the data, delays in adding data to the register or differences between the variables being measured in the register and the variable of interest. This means that both administrative and survey data may contain classification errors, although originating from different types of sources.

A method that is of interest in this context is the recently proposed (Boeschoten et al. 2017)[9] that combines multiple imputation and latent class analysis (MILC) to correct for misclassification in combined data sets. Latent class analysis (LCA) identifies a categorical latent variable using categorical observed variables. LCA can be used to evaluate measurement errors for categorical response variables when different sources that measure the same phenomenon are available. A multiply imputed data set is generated which can be used to estimate different statistics of interest in a straightforward manner and can ensure that uncertainty due to misclassification is incorporated in the estimate of the total variance.

Compared to other methods, the MILC method takes both visibly and invisibly present errors into account simultaneously by combining Multiple Imputation (MI) and LCA. The name 'invisibly present errors' is given because these errors could not have been seen in a single data set. They can be dealt with by estimating a new value using a latent variable model. Some errors can then be observed already when an impossible combination between a score on the attribute and a covariate is detected, which we define as a visibly present error. The name 'visibly present errors' is given here because (some of) these errors are visible in a single data set.

Assigning values to the latent variable can be beneficial for several reasons. First, imputations can also be created for individuals having missing values on either one of the observed variables. Second, as imputations are created for the entire population, it becomes straightforward to produce consistent small-area estimates or to create cross-tables with different covariates. Third, because multiple imputation is used, all results can be supplemented with appropriate variance estimates.

Fractional hot deck imputation (FHDI)

There are also multiple imputation frameworks for imputation to be considered. One such method is the fractional hot deck imputation (FHDI) method[10], which is suitable for mass imputation in which there is likely to be greater imputation variance. It is a hot-deck donor imputation method that provides a single imputation based on the conditional probabilities of obtaining the value. It is suitable for categorical variables.

Multivariate imputation

A key decision in the imputation of linked administrative datasets is whether to use a univariate or multivariate imputation approach. That is, whether to impute the target variables separately one-by-one, or to impute them simultaneously.

If there is correlation between target variables, then we will want to use a multivariate method to preserve the relationship between them. It is however not always statistically optimal or even possible to impute all variables in a complex dataset simultaneously, particularly when missingness in just a single covariate variable prevents the use of the record for imputation of the target unit and thereby reduces the size of donor pool and consequently the quality of the imputation. If there are lots of imputable variables, it may also be too computationally costly to determine the optimum imputation actions for all variables at the same time, plus if there are any edit rules these may need to be applied first. A solution used by the Census imputation method, was to modularise the data to separate the imputable variables into subsets (modules) and impute just the variables in each module simultaneously. This enables a level of multivariate imputation but reduces the complexity of the task. Modularisation also makes it possible to include only the variables that help to predict each other in each module, therefore potentially ensuring better preservation of higher order distributions compared with a fully multivariate method. Priority would be given to the higher priority variables or the analytical task. As mentioned previously, for the administrative data outputs we consider the use of imputation in the context of smaller datasets which include data for key multivariate outputs only rather than a large Census type dataset with lots of characteristic variables.

We will continue to explore parts of the missingness that we can address with imputation, but it is unlikely that we will be able to impute values for all the item missingness in the ABEIS sufficiently to address the resulting bias. This is due to the specific nature of missing data (that is related to target variables) and the lack of covariate information from which to make predictions. In other words, there will still be considerable proportions of missingness that is not at random. We can consider the imputed data set as improved as far as is currently possible and explore estimation methods that make use of independent survey data to reduce the bias of the outputs. Outputs from these models would provide aggregate level outputs with measures of uncertainty. If successful they could potentially form outputs in themselves or be used as robust benchmarks in a wider framework. We consider two main methods Generalised Preserving Estimation (GSPREE) and the more traditional design-based Small Area Estimation models used to obtain more detailed but robust survey outputs.

3.2 Generalised Structure Preserving Estimation (GSPREE)

The GSPREE method uses small area estimation techniques to combine and draw strength from a number of different data sources. Most traditional small area estimation methods use covariate data that can act as a predictor to improve survey estimates. It is sometimes the case, however, that correlating auxiliary data for the separate categories are not readily identifiable. In such instances the GSPREE method can take proxy data source(s) that contain information for the same set of areas and categories but have issues such as they are outdated or have a slightly different definition. These proxy data source(s) are then supplemented by a social survey to generate more reliable and complete estimates than it would be possible to generate from each source individually. GSPREE establishes a relationship between the proxy data source(s) and the survey data for the target year, with this relationship being used to adjust the proxy distributions towards the survey distributions.

For the structural approach, which underlies the GSPREE method, the more detailed sources (usually administrative or census data) are considered “proxy” data for the required cross tabulation. The term proxy is used here in the customary sense of proxy variable as defined in Upton and Cook (2008): “A measured variable that is used in the place of a variable that cannot be measured”. These sources have detailed information available in the same structure as the table of interest (i.e., target cross tabulation), but do not have the exact definition required. They may refer to a previous time point, not cover the correct population, or have categories that are defined slightly differently to the variable of interest (i.e., target variable).

On the other hand, the survey data is designed to represent the target population and measure the variables of interest. It typically provides aggregate estimates and therefore has some information about the target cross-tabulation but is not robust across all cells in the table. The GSPREE method obtains the “best" possible estimates for cells in a target cross tabulation by combining the proxy data with the current survey data through a statistical model. The approach assumes that the survey data provide “correct”, unbiased, and precise estimates for the target variables at the national level. Having obtained estimates of the cross tabulation by combining the proxy and survey data the GSPREE method then benchmarks this table to estimates of the row and column totals (i.e., margins) that meet the required quality standards. Variance of estimates are calculated by a semi-parametric bootstrap re-sampling of the observed survey and proxy sources to produce mean squared error measures, from which confidence intervals or coefficients of variation (CV) estimates can be derived. For full details on the GPSREE methodology, please see Correa-Onel, Whitworth and Piller (2016)[11].

ONS has previously explored implementation of the GSPREE methods for ethnicity by local authority and published research papers (Correa-Onel, Whitworth & Piller, 2016). More recently this research has been extended as more data sources on ethnicity have been obtained and integrated via the ABES[12].

Advantages and limitations

The method provides a strong theoretical basis for estimating crosstabulations; it provides aggregate estimates with measures of quality and is robust to missingness in the data sources. It also provides the flexibility that sources can easily be included if they provide useful information about the distributions in the target variable even if they are incomplete or definitionally incorrect. The closer the distributions to the true distributions (as measured by the survey data) the higher the weighting they will receive in the modelling process. However, there are specific considerations as more sources are incorporated, for example the variable categories must be consistent, the best approach for integrating data sources for the proxy tabulation must be established, and the validity of the association structure when derived from integrated sources.

Theoretically the methods can easily be extended to multivariate outputs by extending the number of categories in the table. This would require the survey data to include all the component variables however, and the detail of the tabulations would depend on the sample counts for the cross-tabulated cells of the table. As with all modelling approaches for estimating at small area or domain level the detail of robust outputs is likely to be limited by the survey sample sizes. If a large number of the cells of the table have zeros or very low counts this can lead to poor quality outputs and artificially low estimates of variance. The problem is compounded when there is a very unbalanced distribution of population across variable categories. We plan to undertake work to establish a clearer understanding of the sample requirements for robust (but detailed) GSPREE outputs.

The method estimates a cross-tabulation and is used for categorical data. For the multivariate case study ethnic group by income, the income data would need to be aggregated into income bands to use GSPREE. This will obviously involve a further trade off in terms of the granularity of the income outputs.

3.3 Regression based SAE methods

Another alternative would be to consider synthetic based small area estimation. This type of modelling uses the strength of survey and auxiliary data sources where survey data alone are insufficient to produce reliable direct estimates for small areas. Its performance relies on covariate information on an area basis and for all areas in the target population. ONS has previously developed and published outputs for household income at a small area level (MSOA) using small area estimation models that regress individual survey responses (household income) on area-level covariate information (MSOA level)[13]. Even though these models are fitted to household level responses from sampled areas only, it is assumed that this relationship also holds for the out of sample data therefore modelled estimates and associated uncertainty measures apply to all areas nationally. So far, these methods have only been implemented for the single income variable and not multivariate outputs. This section describes early thinking on how this modelling approach can be extended for producing income by ethnicity estimates at small area level.

Currently income is modelled as a continuous variable in multilevel regression to produce average weekly household income. If modelling is to be extended to a multivariate scenario (income by ethnicity) different approaches should be considered where income would be treated either as a continuous variable or categorical.

Multilevel modelling of income with poststratification could be used to preserve the continuous nature of the income variable. This method is applied in two steps, first a multilevel model is fitted to individual/household level response data and then the results are weighted according to information in poststratification tables. The nature of the method means it can produce estimates not just for the variable of interest but also for its breakdown by categorical variables (strata), provided these variables are included as predictors in the multilevel model. The poststratification table usually comes from either a large-scale survey or census and finding a data source of a sufficient quality for population counts with a similar level of detail might prove challenging.

Multinomial logistic regression is another approach to consider, where a unit level model is fitted with income by ethnic group as the dependant variable and independent variables are included at area level. However, for this type of model income will have to be categorised into income bands, which would result in information loss. Also, there is currently no consensus on how best to categorise income so further work would be required to establish this. Assuming income is categorised into 3 income bands and ethnicity to 5 ethnic groups, then the income by ethnicity variable will have 15 categories. The number of categories would be quite large for interpretation of the results to be practical.

Overall, performance of regression-based methods would be subject to good quality covariate data and sufficiently detailed survey data. It might prove challenging to find data at the required granular breakdowns. The Family Resource Survey (FRS) from Department for Work and Pensions (DWP) is currently used for official model-based estimates of the mean household weekly income and this survey could potentially be used for the multivariate modelling since it collects information on both income and ethnicity. From preliminary analysis of FRS data, it appears that the prevalence of Local Authorities (LAs) with zero sample counts for the rarer ethnic groups is likely to impact on the performance of the models, and even more so if estimation at smaller geographical areas like MSOA is considered.

4. Under coverage of the SPD: blending the unit level and aggregate level data

We are considering options for blending the improved unit level dataset (where some missing values have been imputed) with the higher-level estimates obtained from the GSPREE and SAE models that meet the required quality standards and are consistent with the marginal totals obtained from the Dynamic Population Model (DPM)[14]. This could potentially provide a mechanism to ensure outputs are adjusted for under and over coverage of the SPD. Initially we consider the approach used for the 2021 Census, albeit the coverage error is much larger in the administrative data.

The census coverage adjustment used for the 2021 Census amended the unit level Census database so that it accounted for net under-coverage in the Census returns for both households and people, and robust estimates could then be obtained for lower-level geographies. The Census coverage survey is used with the Census data in a model-based approach to provide estimates for a variety of key demographic characteristics in local authority areas that account for patterns of under and over coverage that are typically concentrated within geographical areas and “hard to count” population subgroups.

The adjustment uses a combinatorial optimisation (CO) imputation approach for each LA separately using estimates of the population derived from the coverage estimation process as benchmarks[15]. The CO procedure starts with a random selection of households (and people within them) and then substitutes households to obtain an optimal solution for person and household level benchmarks. The aim is to obtain representative aggregate level population totals rather than an accurate unit level database.

We will initially consider the feasibility of this type of approach for the administrative based population characteristics within the case study on ethnic group by income. Key points that we will need to address are:

- What benchmarks we will need to take account of a) for the single variables, b) for the multivariate relationships?
- Can we derive these estimates to the required level of detail and quality? As for the Census we would need coverage adjusted estimates for the total population based on the SPD and robust estimates of the key characteristic variables at LA level.
- Understanding and exploring methods for blending the unit level dataset with these higher-level (benchmarks) estimates, including a) weighting methods, b) combinatorial estimation, c) machine learning predictions
- Quality assessment of the lower-level outputs derived from the adjusted dataset

If successful, the approach would address missingness in the characteristic information due to under coverage in the SPD but would not necessarily address missingness for people present on the SPD.

5. Expected outputs using these methods and considerations

We do not yet know what all the analytical aims of the output data might be. The methods we have talked about in this paper must be considered in the context of the user need. A challenge will be determining whether it is appropriate to use a different set of methods on the same data to suit different analytical aims, or whether we should be aiming to deal with missingness with a one-size fits all approach (like Census) such that all analysts apply their analysis methods to the same data. There are the following considerations:

The unit level dataset would have less missingness following imputation, but we anticipate there will be some missingness left that we cannot successfully impute, particularly where we do not know the nature of the missingness and where there is a lack of covariate information. We will continue to look for ways to improve the imputation, for example by looking at additional auxiliary data sources available, and how to account for it within small area estimation approaches. Additionally, we will have a range of possible values for the imputed cells in the table due to variance in the imputation procedures. We would need to consider how to select a point estimate and represent the uncertainty in this.

Estimates at the aggregate level, depending on the sample counts, are likely to be at LA level or higher and at broader characteristic categories in order to meet quality standards, depending on the sample counts. There will inevitably be a trade-off between granularity in the demographic detail and the level of geography. They also will have uncertainty, and this will need to be reflected in quality measures for the cross-tabulated table.

The adjustment or weighting of the unit level dataset to blend with aggregate estimates (benchmarks), if successful, would account for the coverage error in the SPD but would not account for missing characteristic information for individuals who are present on the SPD and where we cannot successfully impute values. We would need to consider the impact of this on statistical outputs and methods to account for the remaining missing data.

It is likely that we would have to estimate key multivariate outputs separately to tune methods according to the properties of the source data. We would then need to consider the coherence of the different admin based multivariate outputs.

6. Future Work

The research work going forward will include the following:

- More work to better understand the missingness in the integrated datasets including deep dives into the data journey.
- Understanding the representativeness of the datasets, possibly with use of a quantitative summary measure derived by linking administrative data with the 2021 census or survey data, for example R-indicators[16]
- Continuing investigation of imputation and SAE approaches for the case study on ethnicity by housing characteristic as well as by income.
- Developing options for an attributes survey: to consider what can be achieved with different levels of granularity and regularity of the survey; sample design requirements e.g., target sampling for hard to capture and small groups; coherence with the coverage survey for the total population by age and sex[17]
- Assessing whether quality targets for outputs are met in the approaches outlined above
- Exploring methods for updating estimates, rolling estimates forward and estimating change.
- Establishing the requirements and options for periodic benchmarking or auditing.

References

[1]Tinsley, B (2019). Developing our approach for producing admin-based population estimates, England and Wales 2011 and 2016. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/developingourapproachforproducingadminbasedpopulationestimatesenglandandwales2011and2016/2019-06-21

[2]Blake, A (2021) Developing admin–based population estimates 2016 – 2020. Office for National Statistics.
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/developingadminbasedpopulationestimatesenglandandwales/2016to2020

[3]HM Government (2021) The Government Data Quality Hub: The Government Data Quality Framework. The Government Data Quality Framework - GOV.UK (www.gov.uk)

[4]Pendleton, S (2021) Admin-based income, England and Wales: tax year ending 2016 revised results. Office for National Statistics. Admin-based income, England and Wales - Office for National Statistics (ons.gov.uk)

[5]Morgan, A (2022) Developing admin-based ethnicity statistics for England: 2016. Office for National Statistics.
https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/developingadminbasedethnicitystatisticsforengland/2016

[6]Harkrader, J & Bellham, M (2022) Developing subnational multivariate income by ethnicity statistics from administrative data, England: tax year ending 2016. Office for National Statistics. Developing subnational multivariate income by ethnicity statistics from administrative data, England - Office for National Statistics (ons.gov.uk)

[7]Blackwell, L (2020) Admin-based population estimates and statistical uncertainty: July 2020. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/adminbasedpopulationestimatesandstatisticaluncertainty/july2020

[8]Blake, A (2020) Population and migration statistics system transformation – recent updates: evaluating coverage and quality in the admin-based population estimates. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmi

gration/articles/populationandmigrationstatisticssystemtransformationrecentupdates/evaluatingcoverageandqualityintheadminbasedpopulationestimates

[9] Boeschoten, L, Oberski, D & de Waal, T. (2017). Estimating Classification Errors Under Edit Restrictions in Composite Survey-Register Data Using Multiple Imputation Latent Class Modelling (MILC). https://dspace.library.uu.nl/handle/1874/363295

[10] Kim, J. K & Fuller W. (2004) Fractional hot deck imputation. https://academic.oup.com/biomet/article-abstract/91/3/559/230380?redirectedFrom=PDF

[11] Correa-Onel S, Whitworth A & Piller K (2016). Assessing the Generalised Structure Preserving Estimator (GSPREE) for Local Authority Population Estimates by Ethnic Group in England. https://www.ons.gov.uk/file?uri=/methodology/methodologicalpublications/generalmethodology/currentmethodologyarticles/assessingthegeneralisedstructurepreservingestimatorgspreeforlocalauthoritypopulationestimatesbyethnicgroupinengland.pdf

[12] Sargent, Z & Morgan, A (2023) Ethnicity Statistics using GSPREE, EAP181, forthcoming UK Statistics Authority publication.

[13] ONS (2022), Income Estimates for Small Areas, England and Wales Statistical Bulletins. https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/bulletins/smallareamodelbasedincomeestimates/previousReleases

[14] Blackwell, L (2022) Dynamic population model for England and Wales: July 2022.
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/dynamicpopulationmodelforenglandandwales/2022-07-14

[15] Whitworth, A, Sexton, C & North, R (2018) The 2021 Census Coverage Adjustment Strategy. https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2020/07/EAP106-The-2021-Census-Coverage-Adjustment-Strategy.docx

[16] R Indicators.pdf (manchester.ac.uk)

[17] Law, E & O'Farrell, K (2023) SPD Estimation Options,). EAP184, forthcoming UK Statistics Authority publication.

Appendix 1. Outline summary of research to support multivariate admin-based outputs

1) Introduction
   - Transformation work on characteristics to date. ONS has:
     - Constructed integrated admin datasets for the total population by selected key characteristics
     - Reported on the coverage and quality of the datasets
     - Investigated the properties for the datasets including the data journey
     - Explored methods to account for the missingness and to estimate from the integrated administrative datasets.
   - The aim is to produce outputs by low level geographies and detailed multivariate characteristic groups. Methods will need to account for under and over coverage and missingness in the underlying data sources to ensure outputs are robust
   - This project sets out a methodological strategy to improve outputs from the integrated administrative dataset using survey data. It specifically considers how estimation methods that draw strength from survey data can be extended for multivariate outputs and explores a framework for achieving the detailed outputs to meet user needs. The research is based primarily on the case-study income by ethnic group but aims to explore methods that are relevant across other characteristic outputs.

2) The integrated multivariate admin dataset for income by ethnic group. Description:
   - It is constructed from multiple data sources at the person level and includes information on ethnic group, income, age group, gender, place of residence.
   - The Statistical Population Dataset (SPD) forms the population spine

- An overview of integrated dataset is provided in the ONS publication: Feasibility research on developing subnational multivariate income by ethnicity statistics from administrative data for England
- The publication explores the proportion of the SPD population with characteristic information at National, regional and LA levels.

3. Research questions concerning quality of dataset, focusing on coverage and missingness.
   - Missingness from Ethnicity (difficult to count groups)
   - Missingness from income (missing components of income, difficult to count groups)
   - Missed links in construction of integrated datasets for ethnic group and income.
   - Missingness from SPD
     - SPD linkage error (SPD produced from the Demographic Index (DI))
     - False positives/false negatives
     - Fully automated linkage has likely caused linkage bias
   - Item versus unit missingness
   - What do we know about the nature of the missingness? (i.e., missing at random, not at random)
   - Is the association of ethnicity by income correct for individuals and representative of target population?

4. Methods research at the unit level
   Imputation for item missingness (i.e., there is a record for the individual but characteristic information is missing). To include:
   - Method options, how they use the data attributes, requirements
   - Imputation in the admin data setting; challenges, considerations
   - Example of imputation for income and for housing type
   - Multiple Imputation using latent class - Example for ethnicity using MILC (to account for measurement error)
   - Method specification to obtain imputed values for multivariate outputs
   - Data requirements, challenges etc.
   - Advantages and limitations *(e.g., we may only impute for a small part of the missingness due to limited covariate information)*

5. Methods research for aggregate level outputs
   GSPREE (Generalise Structure Preserving Estimation)
   - Methods, how they use the data attributes, requirements
   - GSPREE in the admin data setting, requirements and what this means for income outputs
   - Example implementation for ethnicity by local authority
   - Extension of methods specification for more detailed multivariate outputs
   - Data requirements, challenges etc *(e.g., Need survey data with both variables, need to explore what sample sizes we would need for robust GSPREE estimates)*
   - Advantages and limitations of GSPREE *(e.g., Well documented theory for method, provides cross-tabulated, aggregate estimates with measures of quality, but more geographic and characteristic detail might be limited by survey sample sizes)*

   Other SAE model-based approaches
   - Methods, how they use the data attributes, requirements
   - Use in that admin data setting, requirements
   - Example implementation for housing tenure
   - Methods specification for multivariate outputs
   - Data requirements, challenges etc
   - Advantages and limitations of model-based approach

6. Research on adjustment of admin characteristic dataset to account for under coverage (early thinking)

- Concept is a) improve admin data set using imputation, b) obtain aggregate level estimates to required quality standards using Small Area Estimation approaches. These form benchmarks for key outputs, c) adjust the admin dataset to meet the benchmarks
- Need to determine what benchmarks are required
- Method options: a) weighting, b) combinatorial optimisation, c) machine learning predictions
- Advantages, limitations and challenges

7. The wider demographic system
   - Population size, characteristics, components of change and demographic rates (how the methods for population characteristics fit within the wider demographic system, e.g. Demographic Population Model)
   - Options for a coverage and attributes survey
   - How to measure change over time
   - Statistical Disclosure Control - considerations

Appendix 2. Missing components of income

The ABIS does not include all components of gross income included in the Canberra handbook definition. Components not included are:

- income (earnings and pensions) from an employer not paid through PAYE
- investment income, including income from property, dividends and interest from Individual Savings Accounts and other saving accounts, bonds, stocks, and shares
- some state support benefits including Universal Credit and Personal Independence Payment
- current transfers, for example, parental contributions, child maintenance payments and educational grants

Appendix 3. The impact of linkage error for the Demographic Index and SPD Coverage

The Statistical Population Datasets (SPD), previously known as the Administrative Based Population Estimates (ABPE), are produced from the ONS Demographic Index (DI), which is an anonymised database of longitudinally linked administrative data. The DI is subject to both over and under coverage. Firstly, under coverage can occur in the DI when incorrect (false positive) links are made between any of the administrative sources that make up the DI. Under coverage can also occur if a person is not present in any of the administrative sources used to create the DI. Over coverage can arise when links are missed (false negatives) between DI data sources, or when duplicates are not resolved within individual DI data sources. The 2021 Census has been linked to the DI to help understand the quality of the DI linkage. Analysis is ongoing to analyse the unmatched (residual) census records from the census and DI linkage, as these will help us to further understand DI under coverage.

To produce the SPD from the DI, “activity-based” inclusion rules are applied to replicate the usually resident population. This could result in additional coverage issues as some records may be incorrectly retained or removed by these rules. For example, 2011 Census to 2011 ABPE v3.0 linkage identified some 16-59 year olds that are in the census but not the ABPE (under coverage). Where these individuals appear on both the Patient Register and CIS and therefore would be expected to receive some income, we might assume they should be in the ABPE, however no recent activity is seen in the Benefits and Income data. Over coverage in ABPEs has also been identified through this linkage (possibly emigrants and short-term residents).

All the above means that the initial ABPE have various coverage issues before linkage to other sources (such as income and ethnicity data) takes place. The linkage errors between the administrative sources are unlikely to occur randomly across groups, meaning that there will be linkage bias in the ABPE. DI linkage is fully automatic, and we have seen evidence of linkage bias in other automatic linkage projects. For example, linkage bias in the 2021 Census to Census Coverage Survey linkage could only be fully removed using a combination of automatic and clerical matching. Matching the data using only automated techniques resulted in linkage bias, by hard to count groups and ethnicity.