Fractional counting: a method to fractionally weight and count integrated administrative data for population statistics

Table of Contents

Key Messages of Paper

Purpose

- This paper presents research that assesses whether fractional counting could produce less biased multivariate admin-based population counts than alternative integer counting methods (both model- and rule-based). Where integer counting seeks to resolve attribute conflict errors in admin data via discrete classification, fractional counting weights and counts contradictory values thereby capturing uncertainty underlying the relative likelihood of alternative values being correct.

Recommendation

- Research should continue with the following goals:
    - Reproduce the analyses in this paper when Census 2021 microdata and contemporary admin data become available.
    - Assess the robustness of using fractional counting to produce multivariate admin-based population counts that are constrained to population totals produced from integer counting (to be fed into an estimation process to be determined to produce admin-based population estimates).
    - Alternatively, assess robustness of using model-based methods for including individuals recorded in admin data in statistical population datasets when compared to the current rules-based methods (a rules-based inclusion method is not available for the 2011 data used in this paper).
    - Demonstrate the efficacy of replacing existing rules-based methods to resolve admin data attribute conflicts over statistical geography and ethnicity with model-based methods where appropriate (either via fractional weighting or discrete classification).
    - Demonstrate that model-based methods for fractional weighting or discrete classification can be kept performant over time e.g. via retraining.
    - Identification of appropriate data to supply true negative target population inclusion labels for retraining the population inclusion model in non-Census years.
    - Exploring more complex modelling approaches e.g. including relationships between individuals, and predicting joint probabilities across attributes.
    - The extension of our experimental fractional counting to include other population characteristic attributes.
    - Potential to use fractional counting to estimate/adjust over-coverage when producing admin-based population estimates.

Key Asks of MARP

- What does the panel consider the highest priority research goals described in our recommendation?
- Is there any work we have not proposed that the panel believes needs to be done before they would endorse a recommendation to use fractional counting in the production of admin-based population counts?
- To what degree does the panel believe the intepretability of machine learning models used in fractional counting is important to understand and communicate to users? By intepretability we mean the ability to explain the basis by which the models make predictions e.g. through feature importance.
- Does the panel have any suggestions for other aspects of population statistics where a fractional counting approach might be advantageous?

Executive Summary

- Fractional counting is a method to weight target population membership and contradictory attribute information in admin data (for example where different sources record different ethnicities for the same individual) so that alternative versions of individuals are counted in proportion to the probability that they accurately describe a real person in our target population.
- This is in contrast to integer counting, where the goal is to edit and filter admin data with discrete categorisation so that it contains a single coherent description of each member of our target population.
- Fractional and integer counting can both be implemented in a model-based approach or a rules-based approach.
- In model-based approaches a statistical model of some sort is fit to predict weights or act as a classifier for administrative data.
- In rules-based approaches these weights or classification decisions are made according to deterministic rules that are usually manually developed and curated by researchers.
- We have developed model-based fractional and integer counters using supervised machine learning algorithms to predict target population membership, geographic placement, and ethnicity from available admin data.
- Using Census and admin data from 2011, we have compared these two counting methods against one another, and against existing rule-based integer methods that have been developed by ONS transformation teams to select statistical geography and ethnicity in record-level admin data during the production of admin-based population counts.
- Fractional counting appears to produce less biased admin-based population counts than integer counting where the statistics are disaggregated into multiple dimensions.
- Model-based integer counting appears to produce less biased admin-based population totals.
- We would expect the performance of fractional counting to improve with the number and quality of data sources used to produce admin-based population counts.
- Fractional counting and model-based integer counting both seem to produce less biased admin-based population counts than comparable rules-based integer counting.
- Our findings should be confirmed with more contemporary data, but they suggest that model-based approaches to constructing and counting statistical population datasets (whether with fractional or integer counting) are more accurate than rules-based approaches and should be preferred in future statistical production.
- Additionally, a key requirement for supervised machine learning models in production is the ability to remain performant over time. We have conducted research to investigate this that will be included in a subsequent paper. This research will also need to be assured and endorsed before we recommended the use of fractional counting in production.

Fractional counting

Introduction

Fractional counting is a model-based approach to fractionally weight record-level integrated administrative data (linked record-level data constructed from multiple separate administrative sources) for the purposes of producing admin-based population counts (ABPCs) (1). A fractional counting model attempts to predict the probability that an integrated administrative record accurately describes a real individual in our target population (usual residents of England & Wales). Specifically, the model should account for the probability that an administrative record describes a person in our target population, and where desirable the probability that the record accurately records the individual's geographic location and key demographic characteristics.

When constructing an integrated administrative dataset for population statistics (a statistical population dataset; SPD), the typical aim is to produce a single view of the integrated data that is as accurate as possible and then count the observed population or estimate the true population. This means the SPD should include a single coherent and correct entry for every member of the target population recorded in administrative data as far as is feasible (Fig 1). In practice it is unlikely to be possible to create accurate ABPCs from administrative data alone, and estimation methods involving survey data may be necessary to account for under- and over-coverage and produce admin-based population estimates (APBEs) (Fig 2). While fractional counting may be an appropriate method to estimate over-coverage, population estimation is beyond the scope of this paper and we will instead focus on fractional counting as a method to improve the quality of ABPCs before estimation.

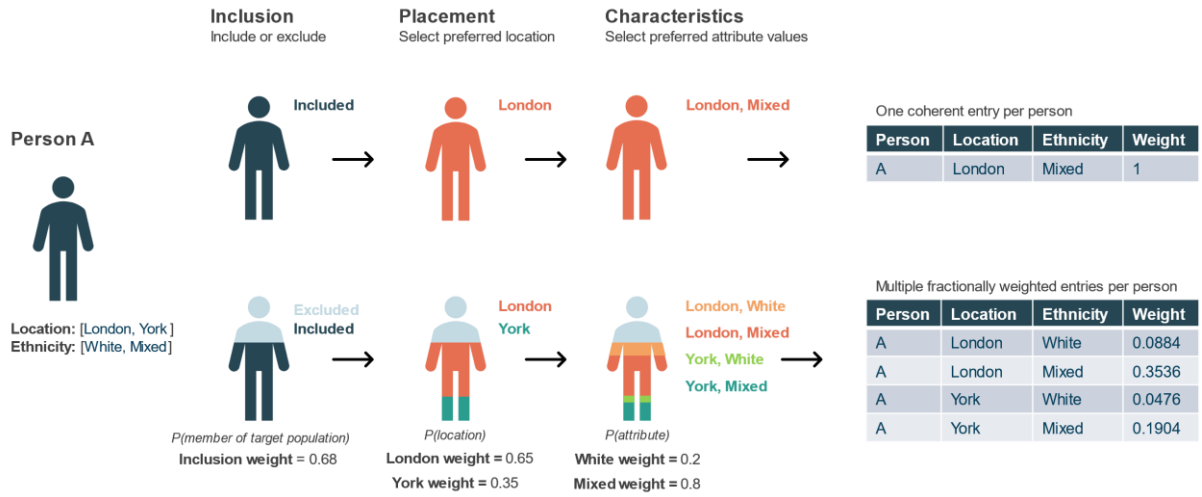

Figure 1 Diagrammatical representation of integer counting with an SPD (top row) and fractional counting with an EPD (bottom row) across target population inclusion, geographic placement, and ethnicity.

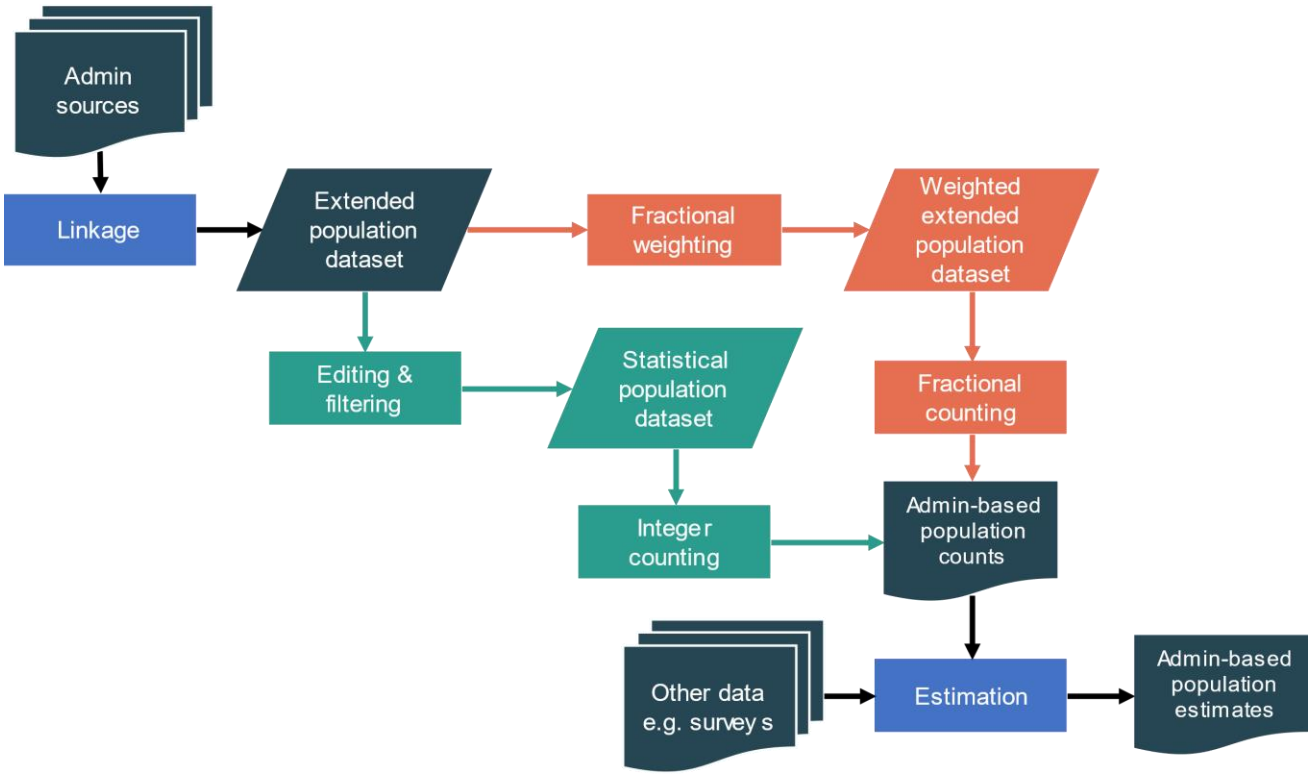

Figure 2 General process for producing admin-based population estimates via integer counting (green path) or fractional counting (orange path).

The population statistics transformation programme at ONS currently implements curated rules to define population inclusion, geographic placement, and demographic characteristics for administrative records e.g. activity- or presence-based inclusion, and hierarchies of trust where sources conflict (2; 3). We can define this approach to constructing and producing population counts from an SPD as rules-based classification (where we are classifying administrative records into particular categories according to deterministic rules e.g. included or excluded from our population estimates based on some activity threshold) followed by integer counting (where the outcome of each classification decision is for a record to be counted wholly in one category). These rules could be replaced by predictive models to produce a model-based integer approach.

When counting fractionally, we are instead interested in predicting the probability that each possible version of a person described in administrative data exists in our target population, i.e. we want to retain alternative attribute values for individuals as well as individuals outside of the target population in an extended population dataset (EPD), and then count the predicted weights (Fig 1). We can conceptually make a distinction between ‘real individuals’ and ‘administrative individuals’, where administrative individuals are possible versions of real individuals as recorded in administrative data (potentially with multiple administrative individuals relating to each real individual). Where we predict these probabilities using a statistical model we can define this approach as model-based fractional weighting followed by fractional counting.

The motivation for fractional counting is to address potential bias that integer counting might introduce when producing multivariate admin-based population statistics. The hypothesis is that fractional counting could more accurately capture uncertainty in the underlying administrative data and its linkage. Consider a situation where two administrative sources record an individual living in two different locations, and both are equally likely to be correct. An integer counting approach would allocate the individual fully to one location. A fractional counter, however, would allocate the individual fractionally to both locations in proportion to the predicted probability of each being correct. Counting fractionally could therefore reduce the bias of ABPEs. This is particularly advantageous when aggregating population counts across multiple characteristic attributes as in regular Census population characteristics outputs. It may also be the case that integer counting would benefit from model-based classification where manually-defined rules also introduce bias. Employing a model-based approach also allows the relationships between admin data and predicted outcomes to be updated in a more automated way as they evolve, rather than relying on manual reviewing and updating of defined rules.

In this paper we outline our research into the feasibility of using supervised machine learning models to build a 'fractional counter' for ABPCs, and compare their use against rules-based and model-based integer approaches. The rules-based integer approaches are those developed by teams in Population and Migration Statistics Transformation (PMST) and Social Statistics Admin First (SSAF) as part of the admin-based population and characteristics transformations programmes. We have also developed MVP (Minimum Viable Product) model-based integer approaches for comparison. While the motivation for these integer models was a lack of rules-based integer methods in some of our test scenarios, we also include them in some comparisons with rules-based methods for consistency. We have used three alternative algorithms (logistic regression, random forest, and gradient-boosted trees) to train three stages of predictive models that assign fractional weights to administrative individuals described in EPD records. These three stages are:

Stage 1. Population inclusion
Stage 2. Geographic placement (Census 2011 Output Area; OA)
Stage 3. Characteristic attributes (ethnicity)

We have used ethnicity as a case study for a population characteristic attribute according to SSAF prioritisation. While it has one of the highest coverage rates of any characteristic in administrative data, it is still far from what is required for high quality ABPCs. While our fractional counting models are designed to assign weights to real ethnicity attribute values as recorded in admin data, it may be possible to use them to impute fractional weights for possible ethnicity values where ethnicity is missing. This may be brought into future work but we will not discuss the potential for imputation in this paper.

We have also conducted an investigation into detecting decay in the performance of fractional counting models over time (a phenomenon called model drift) and the need to regularly retrain them with new data. That work will be included in a future paper.

This paper will describe the data and methods used to build our fractional counters, compare their performance against equivalent integer counting methods, and where possible make recommendations on future research and the design of the transformed statistical system. We are seeking input from MaRAG regarding the strength of our evidence and recommendations, and suggestions for where the work should go next.

Data

Extended population dataset

To construct the spine of our EPD we use record-level data from the Patient Register (PR), Higher Education Statistical Agency (HESA), English and Welsh School Census' (ESC, WSC), and ONS birth registrations.

We use the linkage table from the Demographic Index (DI; v1.2) for linkage, and the actual construction process is implemented in the SPD v3.1.2 build pipeline developed by PMST (with the SPD spine extracted before any editing or filtering takes place).

These data are the source of most core person-level attributes including sex, single year of age, postcode, and ethnicity for students.

Additional ethnicity attributes are sourced from Hospital Episode Statistics (HES), Emergency Care (EC), and Improving Access to Psychological Therapies (IAPT).

All source data are from 2011 and our EPD is constructed with a reference date of June 30th 2011. The spine of our EPD covers all Local Authorities (LAs) in England & Wales, but some ethnicity data is only available for England.

Training labels

Training supervised learning models requires data containing predictive features (independent variables) and labels (dependent variable). In our case, this means administrative records that describe potential members of our target population, the statistical geography of their usual residence, and their ethnicity, where the correct value for each of these attributes is known. True labels (sometimes also referred to as 'gold standard' labels, or 'ground truth' labels) are taken from Census 2011 as linked to our 2011 EPD.

Comparator data

Census 2011 microdata are used to produce our benchmark population counts.

Methods

Concepts and definitions

Target population

We define members of our target population for 2011 as those observed in the Census 2011 record-level microdata. Applying this definition to admin data carries

assumptions that every member of our target population responded to the 2011 Census, and no members outside of our target population are in the Census data. We know that approximately 6% of the 2011 population were not included in the Census (1), and any of these people would be flagged incorrectly as not members of our target population where they appear on admin data in 2011, but we consider the coverage high enough for research purposes. The over-coverage rate of Census 2011 was estimated to be approximately 0.5% (2). Any part of this over-coverage that represented real people who weren't usual residents of England & Wales in 2011 would be flagged incorrectly as non-members of our target population where they appear on admin data in 2011, but we also consider this to be low enough for research purposes.

Ethnicity
We use a 5-category definition of ethnicity: White, Black, Asian, Mixed, and Other.

Classification vs fractional counting
Classification is a task in which observations are placed discretely into one or more categorical classes. This is an appropriate approach to use when integer counting, and we might call models built for this purpose integer counters, or classifiers.

When fractional counting, we generally use the same kind of classification models, but we are instead interested in the predicted probabilities underlying each class. It is these fractional weights that are counted to produce ABPCs.

Address vs. geography
Throughout this paper we refer to an individual's address and statistical geography interchangeably. By 'address' we mean the statistical geography in which their usual residence exists, rather than a postal address or other property-specific identifier.

Constructing the extended population dataset
We link together all records from our core source data using the linkage table of the DI. This initially produces an EPD spine with one row per DI record containing the information from all the source records (with a maximum of one source record per source dataset for each DI record), retaining any duplicate records arising from erroneous linkage or deduplication in the DI construction process. To this EPD spine we join attributes of interest from the core source data (e.g. single year of age, sex, and address), and additional ethnicity attributes from various healthcare data (also via the DI). While sex and age could theoretically be counted fractionally, for simplicity, and to focus on placement and ethnicity, we treat them in a classification approach and assign the single most recent value for both where there are conflicts.

This person-based EPD is subsequently restructured to produce a table for each modelling stage in our fractional counter as follows:

- Population EPD: One row per DI record
- Geography EPD: One row per DI record / OA combination
- Ethnicity EPD: One row per DI record / ethnicity combination

As a toy example:

If the population EPD is as follows:

ID	*OA_source_1*	*OA_source_2*	*Ethnicity_source_1*	*Ethnicity_source_2*
1234	A	B	Black	White

The geography EPD would be structured as:

ID	*OA*	*Source*
1234	A	1
1234	B	2

And the ethnicity EPD would be structured as:

ID	*Ethnicity*	*Source*
1234	Black	1
1234	White	2

Additional attributes are derived for each table for use as features (independent variables or predictors) in our models. Some describe record metadata e.g. time since the record was created or updated. Some describe individuals e.g. flags to indicate whether an individual has been observed in an education dataset in the previous 5 years. Some describe geographies e.g. the number of DI records that are associated with an OA over a reference period.

While the geographical coverage of the EPD is England & Wales, we restrict the EPD to England when analysing ethnicity population counts due to poor coverage of ethnicity in Wales in the 2011 data. The coverage of ethnicity in the EPD is ~35% for 2011, however this has significantly improved in the subsequent years to ~87% in 2016-2018.

Generating training and testing sets

Training and testing datasets are sampled from the EPDs using stratified random sampling to ensure relevant sub-populations are represented. We stratify the data by age, sex, and LA (where individuals have more than one recorded address we select one randomly to be their sampling address) and randomly select a proportion of individuals from each stratum equal to the proportion of desired sample size to target population size. Sampling weights are used when training the models to account for slight differences in selection probabilities between strata.

To train the 2011 models we select a 0.4% sample from the person EPD (n=239,775) and the geography EPD (n=250,941), and a 1% sample from the ethnicity EPD (n=202,971). Each model was subsequently validated on a separate distinct holdout dataset to assess record-level model accuracy, though these results are not reported in this paper (0.8% from the person EPD (n=479,975) and geography EPD (n=501,378), and 2% (n=327,009) from the ethnicity EPD).

Model training

Training

We have used three candidate algorithms to build fractional counters: logistic regression (LR), random forest (RF), and gradient-boosted trees (GBT). We used implementations from sklearn (LR, RF), XGBoost (GBT) or Cloudera's PySpark-compatible MLLib (LR, RF, GBT) depending on the size of the data used for training.

Each algorithm was trained as follows:

1. Optimise hyperparameters (parameters whose values specify how the algorithms are trained e.g. maximum depth of a decision tree) for each algorithm using 3-fold cross-validation on the training set using log-loss as the performance metric
2. Apply optimised models to holdout set and select best candidate method based on holdout metrics
3. Retrain selected model with optimal hyperparameters on full training set
4. Apply fitted model to EPD to produce weights

Note that our training data is a subset of the data we use to produce our experimental ABPCs. This is not strictly appropriate as the models will be tailored to perform well on this subset (a form of data leakage), but given the relatively small size of the training sets we do not believe this meaningfully impacts our conclusions.

Population inclusion (stage 1; S1) is treated as a binary classification problem (included or excluded). Geographic placement (stage 2; S2) and ethnicity (stage 3; S3) are treated as one vs. rest classification problems. This is an approach where a multi-class classification problem is operationalised as a binary classification problem; for each prediction the model predicts whether the target is a focal value or any other value e.g. if ethnicity is Asian or not-Asian. Additional features describing the alternative values are given to the model so that the predictions are not wholly independent, for example the number of other sources and the proportion that agree with the focal value. This allows us handle variation in the number of available classes for each observation (e.g. different numbers and combinations of recorded ethnicities across source data) with one general model rather than training multiple models for different combinations of possible outputs. These models thus make separate predictions for each possible class, and the positive prediction weights across all possible classes are constrained to sum to 1. The placement model predicts at the Output Area (OA) level (OAs are aggregated to LAs for analysis), and the ethnicity model value predicts ethnicity from the 5-category framework used in other ONS social statistics transformation research (Asian, Black, Mixed, White, Other).

True labels for population inclusion, geographic placement, and ethnicity used to train and test our models are taken from Census 2011 linked to the EPD via the DI.

We have also trained equivalent integer counter models in the same manner except that hyperparameters were optimised by maximising AUC (Area Under the ROC Curve) as the performance metric. AUC is more appropriate for classification tasks as it captures a model's ability to discriminate classes across classification thresholds. In comparison, log-loss compares predicted probabilities against the true

values which makes it more appropriate for optimising models used for fractional counting.

We also optimised the classification threshold of the integer model from its default of 0.5 by treating it as an additional hyperparameter.

Counting

We use the trained models across the three stages to apply weights to our person, geography, and ethnicity EPDs. These separate EPDs are then integrated to produce a single full EPD table. The product of the stage weights produces the final weight for each 'administrative individual' i.e. each possible version of a person. The sum of the final weights for an individual is equal to their inclusion weight.

A toy example of an integrated full EPD with weights:

ID	Age	Sex	OA	Ethnicity	Inclusion weight	Placement weight	Ethnicity weight	Final weight
123	31	M	A	Mixed	0.8	0.2	1	0.16
123	31	M	B	Mixed	0.8	0.7	1	0.56
123	31	M	C	Mixed	0.8	0.1	1	0.08
234	45	F	B	Asian	0.9	1.0	0.6	0.54
234	45	F	B	White	0.9	1.0	0.4	0.36

These final weights can be produced and counted flexibly according to the desired output. For example, to produce national population disaggregated by ethnicity the final weights would be the product of the inclusion and ethnicity weights, and they would be then totalled by ethnicity group.

For practical purposes we use the integer methods to assign integer 'weights' of 1 or 0 to the EPDs in the same manner so that they can interact with our data processing pipeline.

Assessment

We are primarily concerned with the accuracy of aggregate population counts produced using a fractional counter rather than the record-level classification performance of the trained models. We therefore report comparisons of the fractionally counted population (and alternative counting methods) compared to the equivalent population counted from our Census 2011 benchmark. Our standard benchmark is unadjusted population counts from the Census 2011 microdata linked to the EPD i.e. records in the union between Census 2011 microdata and the EPD. However, we also include some adjusted final Census estimates in some comparisons.

While the Census 2011 microdata benchmark counts do not equal the adjusted Census estimates they more accurately reflect the 'truth' of the data that the fractional and integer counting methods are trying to reproduce. Thus, using the unadjusted Census microdata allows us to better compare the methods relative to their true benchmark target, or best possible performance, rather than also having to contend with coverage and other error in the microdata.

To compare our aggregate population counts we calculate absolute and percentage differences between our predicted population (either the total fractional weights or integer counts) versus the total benchmark population (normalised and non-normalised) When comparing disaggregated population counts we also calculate Root Mean Square Error (RSME) and Mean Absolute Error (MAE) across the units in the strata of interest. Where possible, RSME and MAE were normalised by the mean of the estimated counts to allow standardised comparison across strata in multivariate results.

For each stage of our fractional counter, we have alternative possible sources of aggregate population counts to produce national and sub-national population counts, in total or disaggregated by ethnicity:

Source of population count	**Source of inclusion weight***	**Source of placement weight**	**Source of ethnicity weight**
Model-based fractional counter	LR, RF, or GBT	LR, RF, or GBT	LR, RF, or GBT
Model-based integer counter	LR, RF, or GBT	LR, RF, or GBT	LR, RF, or GBT
Rules-based integer counter	Not available for 2011	PMST selection rules	SSAF selection rules
Benchmark	Census 2011 microdata	Census 2011 microdata	Census 2011 microdata

***** where a 2011 inclusion weight was not available, we used Census 2011 microdata linkage to define a standardised target population for all methods to compare their respective placement and ethnicity stages.

Note that while we refer to ‘weights’ from integer counters this is only for practical purposes, and they are integer weights of 0 or 1.

We also present some results disaggregated by age and sex, but as these are not within the scope of our assessment we simply select individuals’ most recent value across admin data sources.

The PMST placement rules come from the SPD v3 address selection rules based on a hierarchy of trust for sources. The SSAF ethnicity selection rules operate based on choosing the most recent valid ethnicity value while honouring refusals (rules used here are those as of February 2022).

In our results we report the performance of the most accurate model-based counters using the most performant algorithm for each stage. For the model-based integer counter and fractional counter this was GBT across all stages.

EPD subsets

For some results we subset the EPD into nested categories (note that for each subset we report the number of individuals and the number of alternative values for geography or ethnicity):

- Target population (TP) – only individuals in the EPD who are linked to Census 2011 with at least one geography or ethnicity value in the relevant EPD

(geography EPD $n_{individuals}$=50,366,632, $n_{geography}$=51,499,023; ethnicity EPD $n_{individuals}$=15,853,632, $n_{ethnicity}$=15,919,280)

- True address (TA) – individuals in the TP subset of the geography EPD whose admin data contain their 'true' geographic location found on Census ($n_{individuals}$=46,827,503, $n_{geography}$=47,821,921)
- True ethnicity (TE) - individuals in the TP subset of the ethnicity EPD whose admin data contain their 'true' ethnicity value found on Census ($n_{individuals}$=15,369,560, $n_{ethnicity}$=15,435,207)
- True address with conflicts (TAC) – individuals in the TA subset of the geography EPD who also have more than one unique geographic location recorded in admin data ($n_{individuals}$= 994,377, $n_{geography}$=1,988,795)
- True ethnicity with conflicts (TEC) - individuals in the TE subset of the ethnicity EPD who also have more than one unique ethnicity recorded in admin data ($n_{individuals}$=65,595, $n_{ethnicities}$=131,242)

While we are interested in the accuracy of statistics that contain the entire TP set, analysing these additional subsets highlights the contribution of fractional and integer counters to the overall accuracy of the ABPCs. Restricting analyses to the TA/TE subsets removes cases where no method could ever select a correct value because one does not exist in the admin data. Restricting analyses to the TAC/TEC subsets allows us to assess only those cases where it is possible for a counter/classifier to weight/select both correct and incorrect values; we see this as the most meaningful subset as it removes all cases where any method would either always be right (if there is no conflict) or always be wrong (if there is no correct admin record).

For results disaggregated by ethnicity, we filter the TP subset (and subsequent derived subsets) to only include individuals who have at least one recorded ethnicity on their source admin data records. We are interested in whether a fractional counter can provide better accuracy than integer methods, and so taking into account the degree of error due to attribute under-coverage is not within the scope of our research. Additional estimation or imputation methods would be required to account for under-coverage in the case of either fractional or integer counting. However, due to the low coverage of ethnicity in our 2011 EPD, we urge caution when interpreting our more granular ethnicity results.

Results

National population

Here we compare a model-based fractional counter against a model-based integer counter with population counts from Census 2011 microdata as a benchmark. Additional results are reported in Annex A.

National total

The target population size as defined by the number of individuals in the EPD linked to Census was 50,366,632. The fractional counter's total weight was 50,130,595, whilst the integer counter summed to 50,326,861. Thus, whilst both counters slightly underestimated the total population size, the integer counter was closer than the fractional, with 99.92% and 99.53% of the target population weight respectively.

National total by age and sex

The population counts were further disaggregated by both sex and 5-year age group, where the fractional counter scored had lower normalised and non-normalised error metrics than the integer counter (Table 1).

Table 1. Error metrics for the integer and fractional counter for national population disaggregated by single year of age and sex.

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0072	0.0060	10078	8408
Integer	0.1087	0.0878	142127	122872

Figure 3 shows how each method compares to the target population across single-year ages for each sex. Whereas the fractional counter consistently matches the target population, the integer counter was less accurate for almost all ages. This is especially true for working-age population totals, where females were largely overestimated, and males largely underestimated.

Figure 3 National admin-based population counts by sex and age for an integer counter (blue) and fractional counter (green), Census 2011 microdata population shown as a benchmark (red).

National total by age, sex, and ethnicity

Further disaggregating the estimates by sex, single year of age, and ethnicity we can

see a difference in the model performance, with the fractional counter scoring better than the integer counter across all the relevant metrics (Table 2)

Table 2. Error metrics for the integer and fractional counters for national population disaggregated by single year of age, sex, and ethnicity.

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0785	0.5092	1251	811
Integer	0.2237	0.1111	3565	1771

Figure 4 also shows the difference in performance (shown as the absolute difference in counts compared to the Census target) across single year of age, by sex, for each of the five ethnicity categories. The difference in the age/sex breakdown for integer counted individuals (Annex A) is composed of an over-estimation of female individuals within this range combined with an under-estimation of male individuals. The fractional counter on the other hand does not appear to suffer from this discrepancy in sexes, with an under-estimation across the whole age range for both sexes but with a slightly smaller under-estimation for males aged 15-40.

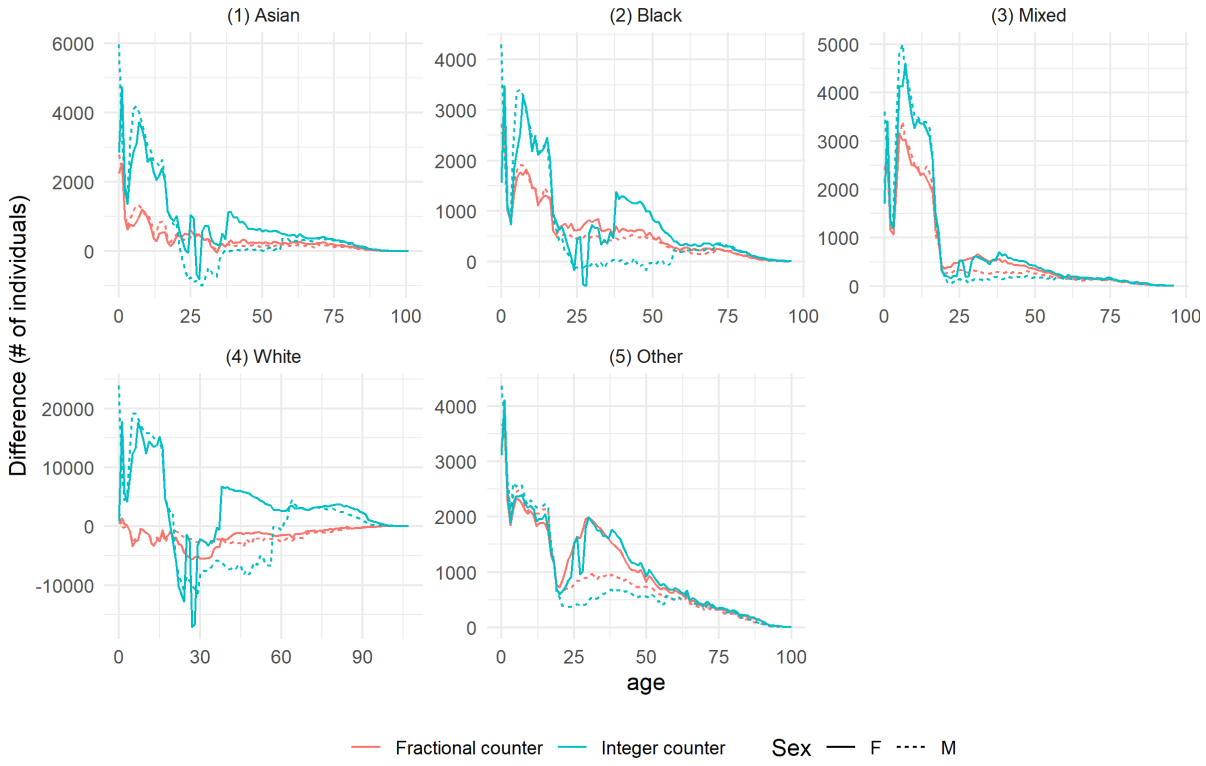

Figure 4 Absolute differences between national admin-based population counts and the 2011 Census microdata benchmark by single year of age, sex, and ethnicity when using the integer counter (blue) and fractional counter (red).

Sub-national population

LA totals

Counting total population by LA, the integer counter produced more accurate weight totals (Table 3).

Table 3. Error metrics for the integer and fractional counter for population disaggregated by LA.

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0624	0.0389	9026	5628
Integer	0.0327	0.0198	4726	2867

LA totals by age and sex

Disaggregating LA weights by both sex and 5-year age group produces fractional counts with a greater level of accuracy compared to the integer counts (Table 4).

Table 4. Error metrics for the integer and fractional counter for population disaggregated by LA, single year of age, and sex.

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0861	0.0420	346	169
Integer	0.1586	0.0921	638	370

The distribution of percent differences from target populations at this high level of disaggregation is shown below in Figure 5. The majority of fractional counts closely match their target populations' sizes, whereas the integer counter overestimates most counts (67.6%) and has a long tail of underestimates.

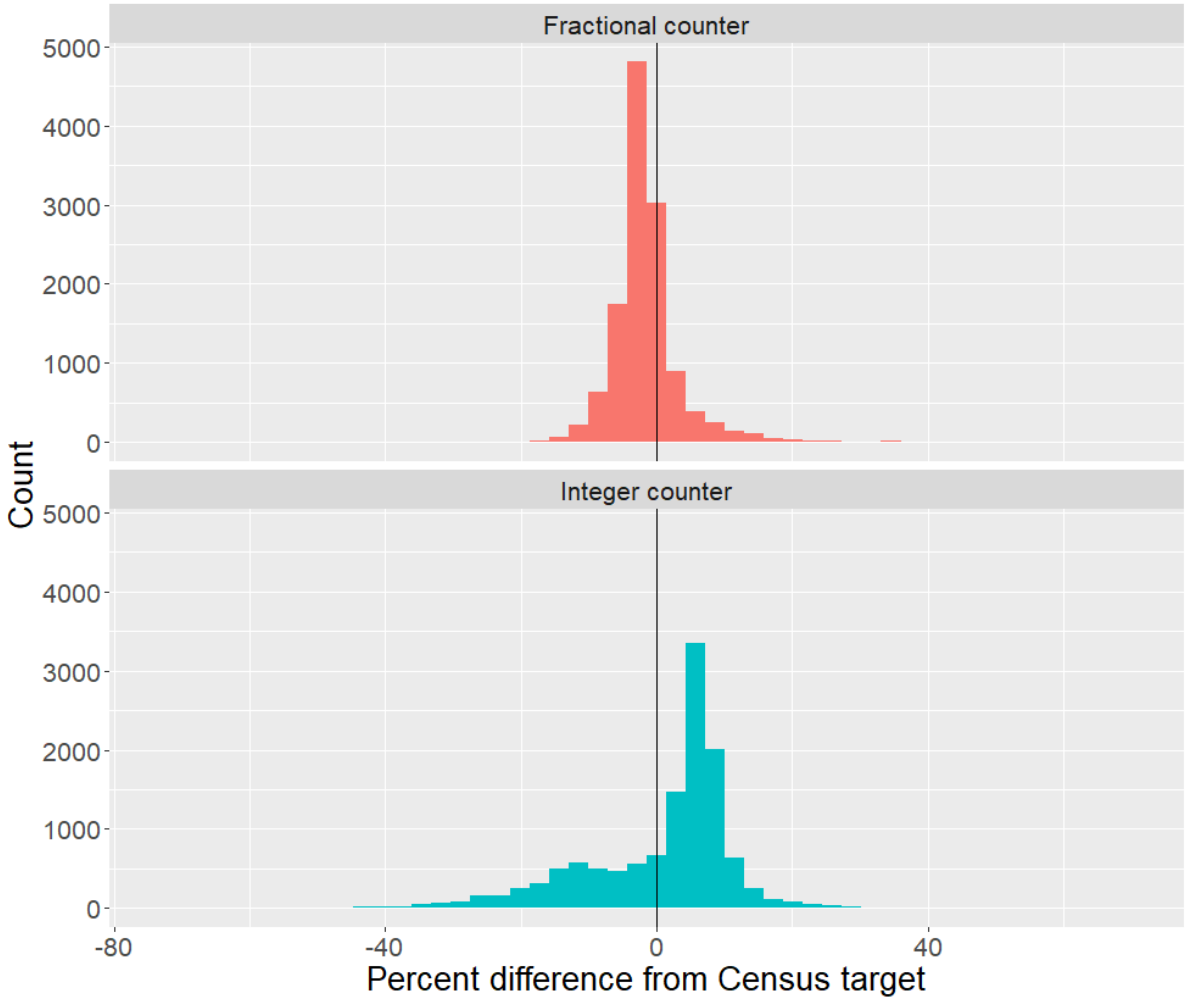

Figure 5 Frequency histogram for the percentage difference between LA admin-based population counts (disaggregated by age and sex) and the Census 2011 microdata benchmark for the integer counter (blue) and fractional counter (red).

Separately calculating error metrics for each age by sex group shows that the integer method is less accurate for every group considered (Fig 6). The gap between fractional and integer counting was especially pronounced for young adult males.

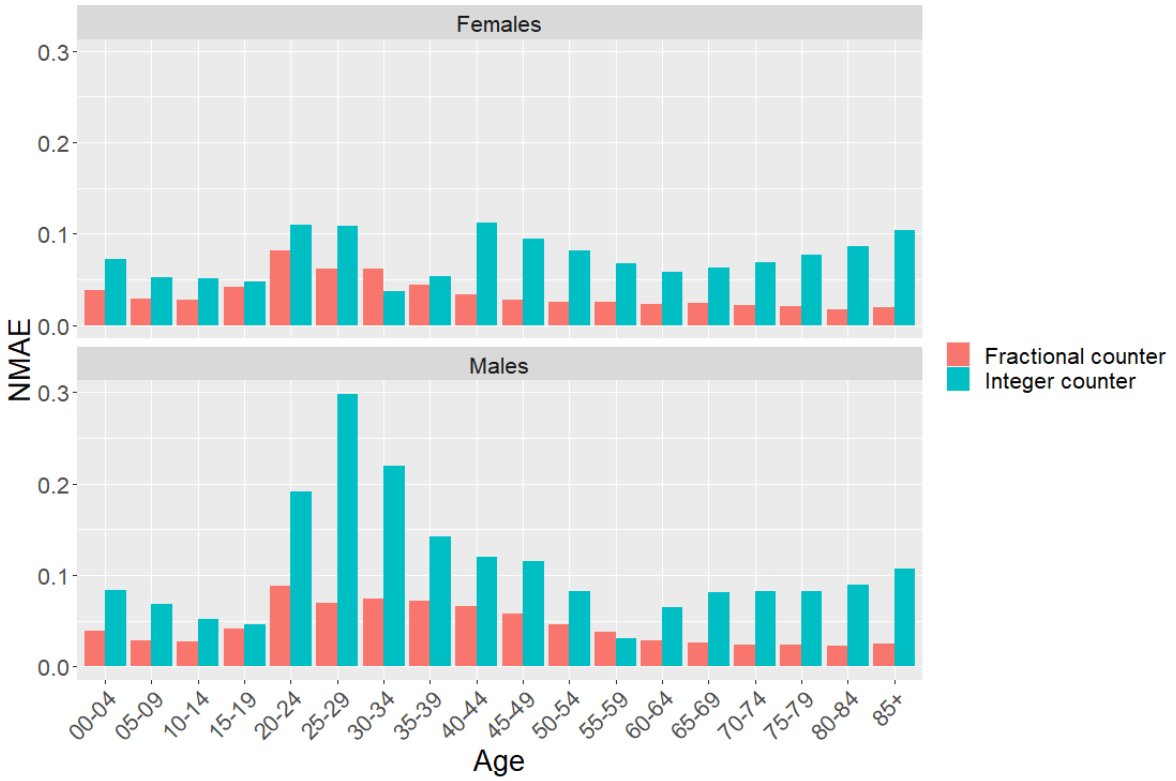

Figure 6 Bar plot showing the normalised mean absolute error (NMAE) between LA admin-based population counts (disaggregated by age and sex) and Census 2011 microdata benchmark for the integer counter (blue) and fractional counter (red).

LA totals by ethnicity

Combining all the stage weights we can look at the geographic (LA) performance of the fractional and integer counters when including ethnicity. Across the applied metrics the fractional counter performs far better, with reduced errors compared to the integer counts (Table 5).

Table 5. Error metrics for the integer and fractional counter for population disaggregated by LA and ethnicity.

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.1397	0.0611	1293	566
Integer	0.2247	0.1068	2079	988

Figure 7 shows the percent difference between the fractional (red) and integer (blue) counter combined weights and the true target population weights. We can see that the fractional counter produces estimates at an LA level which are skewed towards under-estimating the population. The integer counter produces more LA estimates that over-estimate the population size, with this shift seen most clearly for those Mixed category individuals where the fractional counter is more likely to under-estimate whereas the integer counter will over-estimate. The greatest qualitative difference is seen for those Other category individuals, where the integer counter under-estimates almost all LAs, whereas the fractional counter difference distribution peaks at a zero difference, but with a wider spread of differences than the integer counter.

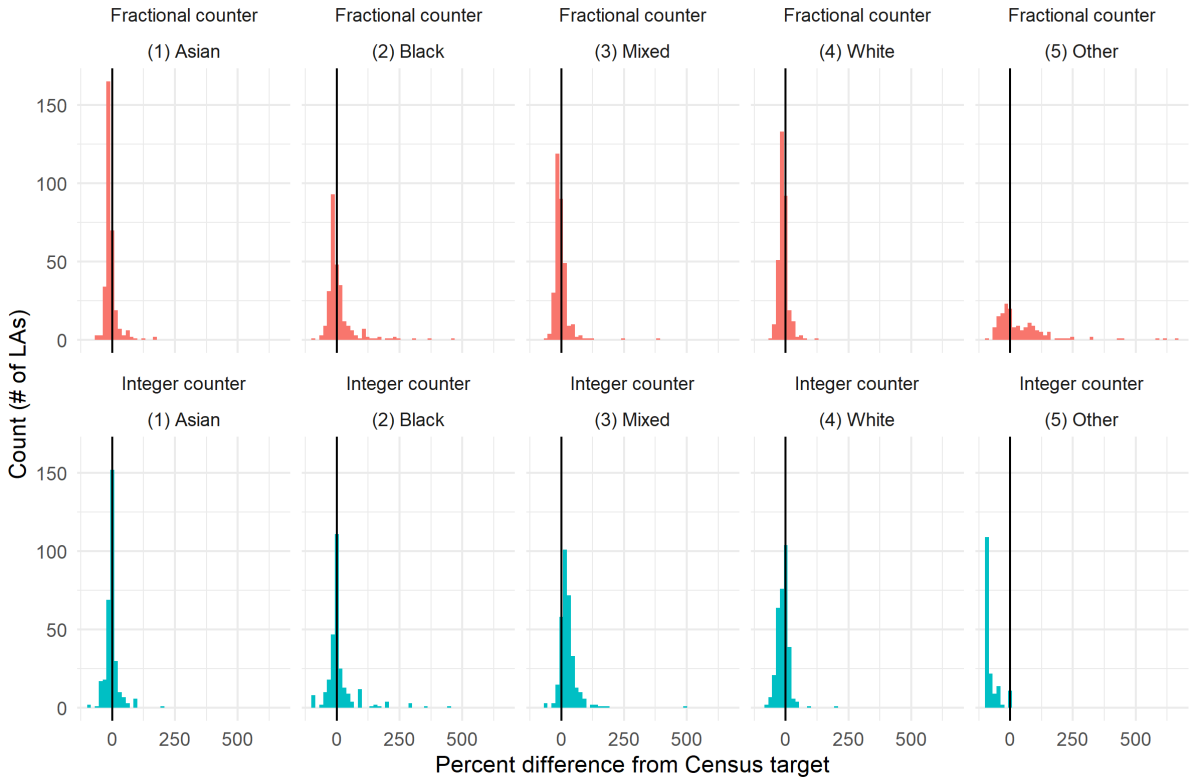

Figure 7 Frequency histogram of the percentage difference between LA admin-based population counts by ethnicity and the Census 2011 microdata benchmark for the integer counter (blue) and fractional counter (red).

LA totals by age, sex, and ethnicity

Further disaggregating the combined stage weights down to LA totals by age, sex and ethnicity level the fractional counter again performs better than the integer counter (Table AA6). We do not report normalised error scores for this breakdown in the same way as previous results because of small count constraints at this level, previous normalisation requires the predicted values for each cell.

Table 6. Error metrics for the integer and fractional counter for population disaggregated by LA, single year of age, sex, and ethnicity.

Counter	RMSE	MAE
Fractional	1.749	0.0416
Integer	2.824	0.3865

Delving further into the disaggregated LA totals by age, sex and ethnicity we normalise the absolute differences across the disaggregated levels by the total estimated size of each LA. Figure 8 shows the distribution of the LA size-normalised differences between each counter (fractional/integer) predicted LA*age*sex*ethnicity weight and the Census target as boxes and whiskers. The outliers have been removed for visual clarity, the number of outliers identified for each of the categories (Asian, Black, and Other) were similar with between ~680 outliers, category (Mixed) had ~450 outliers and category (White) had ~250 outliers. However, the outliers for category (White) were all significantly larger than those for the other categories.

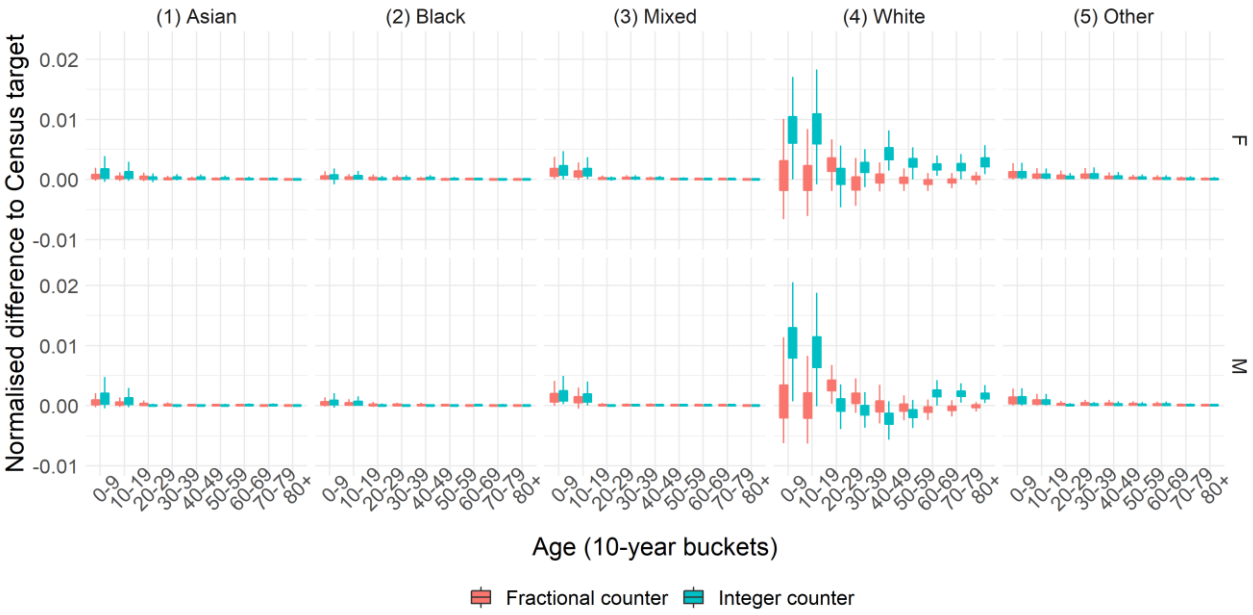

Figure 8 Box and whisker plot showing the distribution of normalised differences to the 2011 Census microdata benchmark from the fractional (red) and integer (green) counters across 10-year age buckets for males (top panels) and females (bottom panels) and ethnicity. Outliers have been removed for clarity. Similar numbers of outliers found in each category, with the outliers for (4) White M/F being significantly larger than the other categories.

Comparing the normalised differences in Figure 8, the fractional counter performs better than the integer counter, with a greater number of LAs with lower normalised differences to the Census target, with the greatest difference being the prediction of category (White) individuals. Whereas the distribution of LA normalised differences are more closely centred on 0 for the fractional counter, the integer counter overestimates counts in all LAs for individuals <20 years old, both male and female, whilst also broadly overestimating individuals >40 years old (male) and >60 years (female).

Standardised performance of the placement and ethnicity models

We have also conducted investigations into the standardised performance of our geographic placement and ethnicity models compared to rules-based integer methods. For brevity we have placed these investigations into Annex B.

Discussion

We have demonstrated that a fractional counter can be constructed to produce admin-based counts of populations believed to be resident at a given level of geography and with specific characteristics. If we could assume no under-coverage in our admin data then the outputs of a fractional counter could be considered unbiased estimates, but we know this is not the case. In this paper we have produced experimental ABPCs rather than ABPEs. Fractional counting should be combined with estimation methods to adjust for under-coverage if it were to be used to produce ABPEs.

Throughout our initial investigation using 2011 data, we are comparing the total counts output by models to the totals that would be obtained by a perfect model. This is defined as the model that predicts the 2011 Census microdata value for each person on the EPD (our extended version of the SPD) in terms of inclusion in the target population, placement within a single statistical geography, and ethnicity. Some responses to the 2011 Census are not linked to the SPD population spine, therefore fractional counting will not assign a weight to these individuals. Likewise, some records on the 2011 SPD population spine that are not linked to 2011 Census will actually be present in the population. Such records are labelled as not in the population here due to Census non-response or missed links. By comparing to the "correct" answer as defined by the linkage via the DI to Census, some people are excluded from our investigation of the performance of counting methods. Primarily, they are those who are not linked to Census but would ordinarily be estimated via DSE based on the Census coverage survey and adjusted for. In the case that these missing individuals are an unbiased subset of the population as a whole, a supervised learning model trained on a representative sample of the population may be general enough to produce accurate predictions, but if there is any structural bias in the membership of this missing group then this is less likely. There are likely to be differences between the methods in their ability to make accurate predictions for these individuals, but these are more difficult to measure and beyond the scope of our current work.

We compared the method of fractional counting, which uses the probability output by classification models, to "integer counting", which assigns to each individual or attribute a zero or one, either according to a threshold or by comparison to the other options available. When training the integer counter, we included tuning of the classification threshold to obtain total population estimates with as little bias as possible. Likewise, training of the classifier models to produce probabilities for fractional counting includes calibration of those probabilities, so we would expect overall totals to match well. The accuracy of the breakdown of counts to Local Authorities (LAs), age and sex tells us how effectively the methods are extracting information from the feature variables to include and place individuals correctly, and whether bias is introduced. Model-based integer counting exceeded our expectations by producing more accurate counts at LA level than fractional counting, but fractional counting was much better after breaking down by age and sex. We would expect fractional counting to perform better here as there is no cumulative bias from summing lots of individuals more likely to have the same outcome.

We also compare to Rules-Based Integer (RBI) counting, which are the methods already developed by colleagues in the Coherent Integrated Population & Social Statistics (CIPSS) team (Annex B). The model-based counting methods perform better than rules-based methods, which is what we would expect given that they can take account of more information and subtlety in relationships between features of admin records. However, it is important to note that we have optimised the model-based counting methods for this specific task of predicting the correct census outcome for those individuals on the EPD. SPD inclusion rules were designed with the aim to reduce net under- and over-coverage as much as possible when SPDs

were compared to mid-year estimates and final census estimates. It is possible that rules-based methods would perform better on our test if the rules were manually tuned with this dataset specifically to optimise performance.

The model-based methods also have a slight advantage that we tested them on the whole SPD population spine, but a small fraction of this was used as training data to fit and optimise the models. Usually, we would make every effort to exclude training data from tests of performance, but in this case for simplicity we used the entire SPD to allow comparison with other experimental methods using the entire SPD, knowing that the slight performance benefit would only apply to a small fraction of the predictions. We validated this by reproducing the analyses in this paper after excluding the training data from the SPD, and found it made little difference to our results and no difference to our conclusions. This work is excluded for brevity, but we include record-level performance metrics on a holdout set of data in Annex B that show the geographic placement and ethnicity models (fractional and integer) out-perform the rules-based methods on data the models have not seen before. It is also worth noting that the rules-based methods that we have compared against were developed and tuned on the full sets of SPD and admin data, which theoretically should give them an advantage over the model-based methods where we exclude the training data.

It is clear that the problem of conflict resolution (i.e. multiple possible geographies or ethnicities for an individual) is small compared to the total size of the EPD. For placement and ethnicity, this is the area where differences between methods result in differences in counts, because where there is only one possibility, that receives a weight of one. Inclusion (Stage 1) modelling has a much greater impact on the LA estimates than placement (Stage 2) modelling. Inclusion modelling is more challenging and less accurate, but we would like to simplify the problem where possible. For example, by using death registrations to remove the deceased from the EPD We could restrict the problem of over-coverage to duplicate records, short-term immigrants, and emigrants.

The ABPCs produced from a fractional counter carry an implicit assumption that the EPD does not contain under-coverage. The fractional inclusion weights then aim to down-weight elements of over-coverage. Levels of under-coverage in the DI are currently being investigated, and parallel research is investigating estimation methods to account for both under- and over-coverage in ABPEs. Fractional counting is one method under consideration for estimating over-coverage in ABPEs, but in this paper we only present it as a method to reduce the bias in ABPCs before they are fed into an estimation method.

Our aim was to produce a minimum viable fractional counting method, so we have not spent a large amount of time exploring additional explanatory variables. The range of features we use is limited to values found on admin records and metadata about those records e.g. their age and how many alternatives were seen on the admin data. We know that with AUC (Area under ROC curve) of around 0.7-0.8, the model is not able to consistently predict higher probabilities for those more likely to be in the population. We will therefore attach more weight to the wrong individuals,

and even if aggregate level counts are reasonable for LA, age, and sex, we may find that estimates of characteristics downstream are biased if those characteristics were not included in the original fractional weighting model. Admin data coverage is now better than it was in 2011, so there is more information to feed into models. Additional explanatory variables may also improve this, for example accounting for geographical differences, or introducing other relevant admin data e.g. country of birth for ethnicity.

During development, we considered alternatives to the one-vs-rest method for ethnicity, as there are a small number of possible values this variable could take. We tested a multinomial model that outputted a weight for every possible value of ethnicity for every person, including those that did not appear on that person's admin data and people with no admin record of ethnicity. These results are not presented in this paper, but this multinomial model performed similarly to one-vs-rest on those with admin records, and was also able to make ethnicity predictions for all those on the SPD by taking into account other features and their relationship with ethnicity as learned from the 2011 Census data. In practice, this process is carrying out fractional imputation for those individuals rather than fractional weighting of existing records.

In the modelling methods, we make two assumptions of independence. Firstly, individuals living at the same address are considered to be completely independent of one another in terms of their probability of inclusion, placement and ethnicity. In reality, they would be likely to form a household and move as one unit, and are also more likely to share the same ethnicity. We have not considered how this dependence could be captured in modelling. We also assume that the stages of modelling are independent of each other, which allows us to multiply weights together from the stages that follow on from one another. A combined model that could predict a joint probability would be much more complex, but could potentially model the problem that these stages are dependent on each other to some extent. Currently, we mitigate for this by including the features used in each earlier stage in every following stage also.

Due to data availability at the time this work was conducted we have used 2011 as our reference date. The availability and quality of data in 2011 is much more limited compared to what is currently available (particularly coverage of attributes like ethnicity). It would be instructive to repeat the analyses reported here with Census 2021 microdata and equivalent modern admin data.

Conclusion

- Integer counting produces more accurate ABPC totals at the national and sub-national level when using model-based methods.
- Fractional counting produces more accurate disaggregated ABPCs when population totals are broken down by age and sex.
- Fractional counting produces more accurate ABPCs for most age-sex-ethnicity groups when population totals are further disaggregated by ethnicity.

- Model-based methods produce more accurate ABPCs than rules-based methods when resolving conflicts over geographic placement and ethnicity in admin data (regardless of whether the models produce fractional or integer counts) (Annex B).
- Model-based fractional counts produce more accurate ABPCs than model-based integer counts in cases where there are conflicting admin values for placement and ethnicity, and where one of the conflicting values is correct (Annex B).
- While the detailed breakdowns of our results that include ethnicity should be interpreted cautiously due to the low coverage of ethnicity in 2011 admin data, they are broadly consistent with our findings for national and sub-national population by age and sex.
- We recommend that the analyses in this paper are re-run using Census 2021 microdata and contemporary admin data when these are available in order to make a definitive recommendation on replacing rules-based methods with the model-based fractional or integer methods described in this paper.
- We have also conducted research into retraning these machine learning models over time to avoid any decay in performance. This research will be included in a subsequent paper and should also be assured and endorsed before any definitive recommendation is made.

Future Steps

- Seek assurance on our investigation of model drift and retraining
- Assess the potential to use similar models in a fractional counter for fractional imputation
- Re-run these investigations when Census 2021 data is available to make use of higher ethnicity coverage in recent data.
- Investigate options to re-train population inclusion models without a Census where a survey does not provide reliable true negative labels.
- Address whether it might be possible to predict the likelihood of displacement (where an individual's true location is not found on admin data) and equivalent concepts for other attributes
- Assess a hybrid approach where fractional counting is only used for subsets of the population where admin data is in conflict or has changed.

References

1. *On provision of UK neighbourhood population statistics beyond 2021.* **Zhang, Li-Chun.** s.l. : arXiv, 2021.

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

3. —. Admin-based ethnicity statistics for England, feasibility research: 2016. [Online] 2021. https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/adminbasedethnicitystatisticsforenglandfeasibilityresearch/2016.

4. —. 2011 Census Statistics for England and Wales: March 2011 QMI. [Online] https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/2011censusstatisticsforenglandandwalesmarch2011qmi.

5. *Overcoverage in the 2011 Census.* **Abbott, Owen and Brown, James.** 2007.

Annex A: Supplementary national and sub-national results

National total by sex

The fractional counter accurately matched the population of males and females, but the integer counter underestimated the male population and overestimated the female population by around 1 million each (Table 7).

Table 7. National population totals by age from the integer and fractional counters.

	Males	Females
Target population	24310131	26056501
Fractional counter	24237424	25893170
% Difference (fractional)	0.3	0.6
Integer counter	23234936	27091925
% Difference (integer)	4.5	3.9

National total by age

The fractional counter also more accurately predicted the population sizes of 5-year age groups (Fig 9). In contrast, the integer model only accurately predicted age groups between 40 and 54; young adults were heavily underestimated, whilst children and age groups over 50 were slightly overestimated.

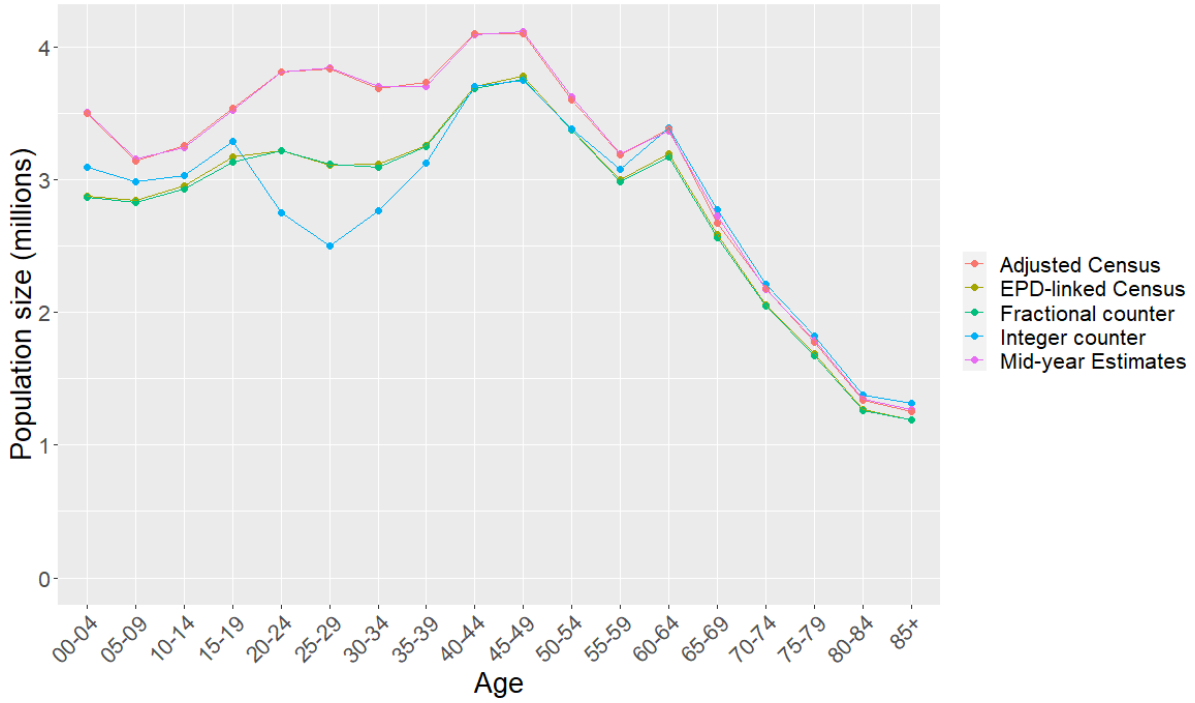

Figure 9 National admin-based population counts by age across different counting methods, including the Census 2011 microdata benchmark (EPD-linked Census) and the adjusted final Census 2011 estimates.

National total by ethnicity

The coverage of ethnicity in the admin data in 2011 is ~31.5%. As such, the size of the target population subset with available ethnicity is 15,853,632. The fractional counter's predicted total ethnicity subpopulation weight was 16,018,845, whilst the integer predicted total weight was 16,845,704. Both the fractional and integer counters return subpopulations that are larger than the target subpopulation. This is most likely due to a coverage difference between the target and non-target populations admin ethnicity data coverage, with more individuals not within the target population being added to the subpopulation (Table 8). Whilst both models predict a larger subpopulation, the integer counter adds more individuals (992,072, +6.26%), whereas the fractional counter is closer to the true subpopulation with fewer added (165,213.5, +1.04%).

Table 8. National population totals by ethnicity for the integer and fractional counters.

	Asian (1)	Black (2)	Mixed (3)	White (4)	Other (5)
Target Population	1191081	553094	253038	13835157	21262
Fractional Counter	1258352	660164	373706	13518938	207685
Integer Counter	1328416	677719	400495	14235885	203189

Alternatively, we can look at the percentage distribution of ethnicity across the 5 categories as opposed to the total counts (Table 9).

Table 9. Percentage distribution of national population total across ethnicity categories for the integer and fractional counters.

	Asian (1)	Black (2)	Mixed (3)	White (4)	Other (5)
Target Population	7.51	3.49	1.60	87.27	0.13
Fractional Counter	7.82	4.11	2.33	84.51	1.35
% Difference (fractional)	0.31	0.62	0.74	-2.76	1.21
Integer Counter	7.44	3.90	2.22	80.36	1.28
% Difference (integer)	-0.07	0.42	0.62	-6.91	1.15

Looking at the five-category ethnicity distributions predicted by the fractional and integer counters, both predict a smaller proportion of White individuals, however, the integer counter prediction is further from the true proportion (-6.911% vs -2.763%). Both the fractional and integer counters predict greater proportions of Black, Mixed and Other individuals, however, the integer counter does predict proportions slightly closer to the true distribution, although they are at most 0.2% closer than the fractional predictions. Looking at the Asian category, the fractional counter predicts a greater proportion compared to the true target population (+0.311%) whereas the integer counter predicts a slightly smaller proportion (-0.073%). Overall, the fractional counter produces a 5-category ethnicity distribution that is closer to the target population than the integer counter, with the greatest contributor to this closer prediction being the White category.

National total by ethnicity and age

We can disaggregate the population counts by age as well as ethnicity (Fig 10). Both counters qualitatively match the age-ethnicity distributions of the target population, however, the integer counter overestimates the count of White category individuals aged 0-18 and underestimates those ages 18-40. The fractional counter broadly underestimates the count of White category individuals across the whole age range whilst slightly overestimating the count of all other category individuals across the age range.

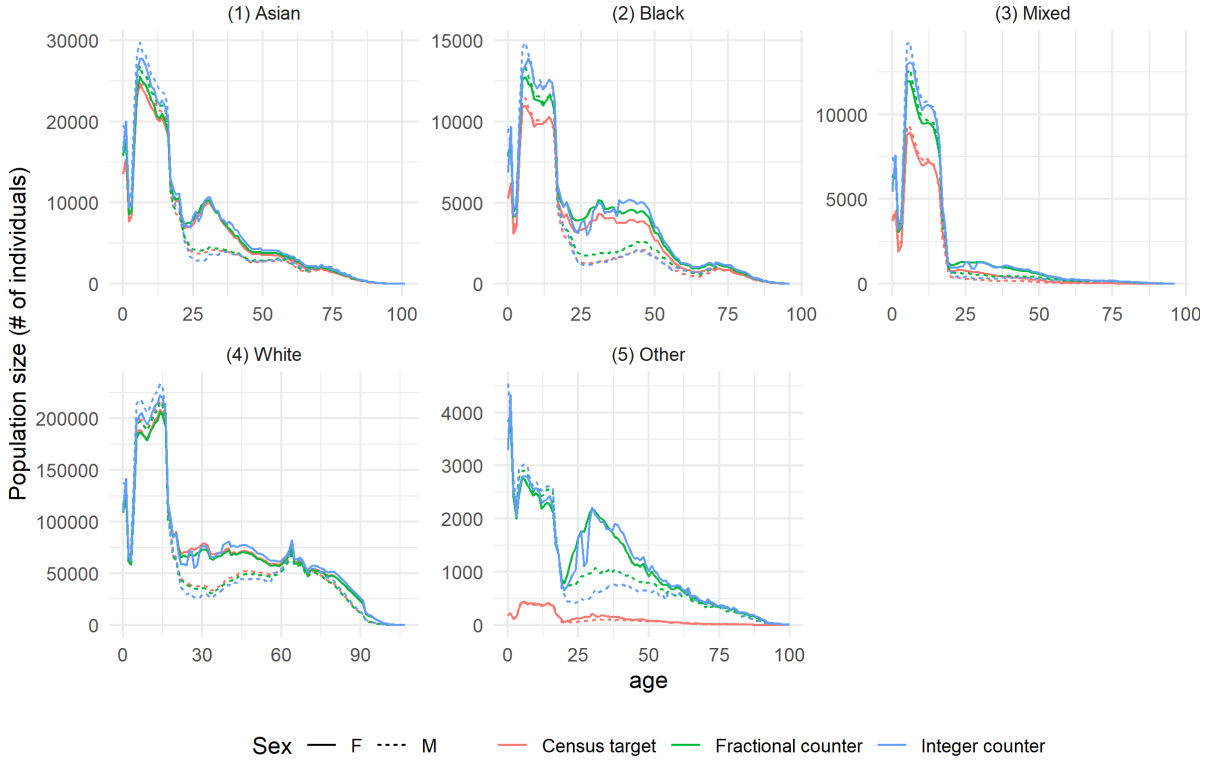

Figure 10 Population size estimates for the 5-category ethnicity (Asian (1), Black (2), Mixed (3), White (4) and Other (5)) predicted using the fractional (red-line) and integer (green-line) counters and the target population (blue-line).

LA totals

Plotting the percentage difference between predicted and actual LA populations shows that the fractional counter slightly underestimated most LAs (77.3%), whilst substantially overestimating a handful of them (Fig 11). Meanwhile, most weight totals predicted by the integer counter are close to the target population size, with no asymmetrical trend towards underweighting or overweighting.

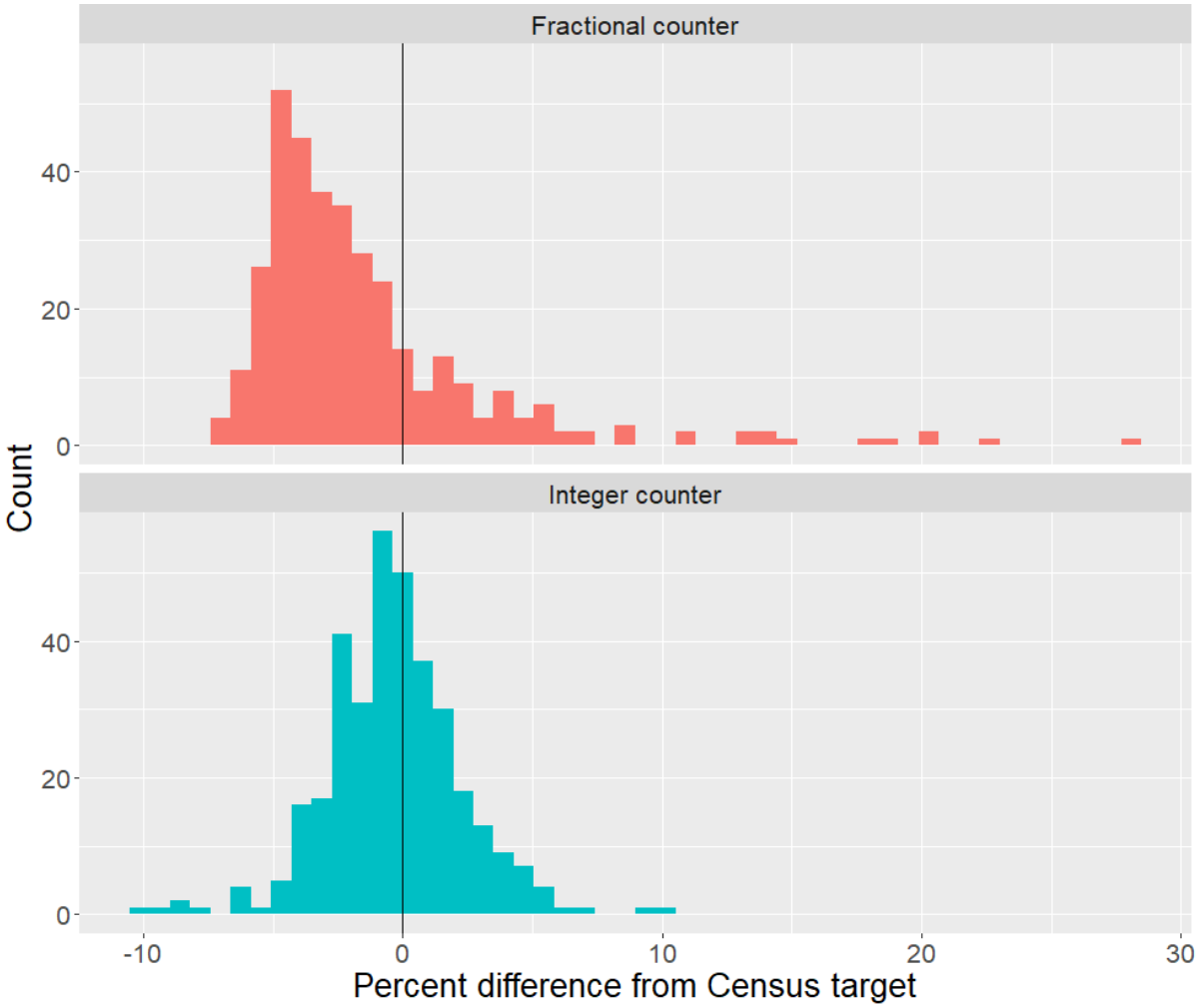

Figure 11 Frequency histogram for the percentage difference from the Census 2011 microdata benchmark for an integer counter (blue) and fractional counter (red) for LA population.

Mapping the percentage differences on a choropleth chart shows how the fractional counting method was especially inaccurate for certain London LAs (Fig 12). The most severely overweighted LA was Kensington and Chelsea: whilst the target population was 101,533, the fractional weight total was 130,177: a 28.2% overestimate. Meanwhile, the integer counter produced a comparatively modest overestimate of 9.5%, with a weight total of 111,153.

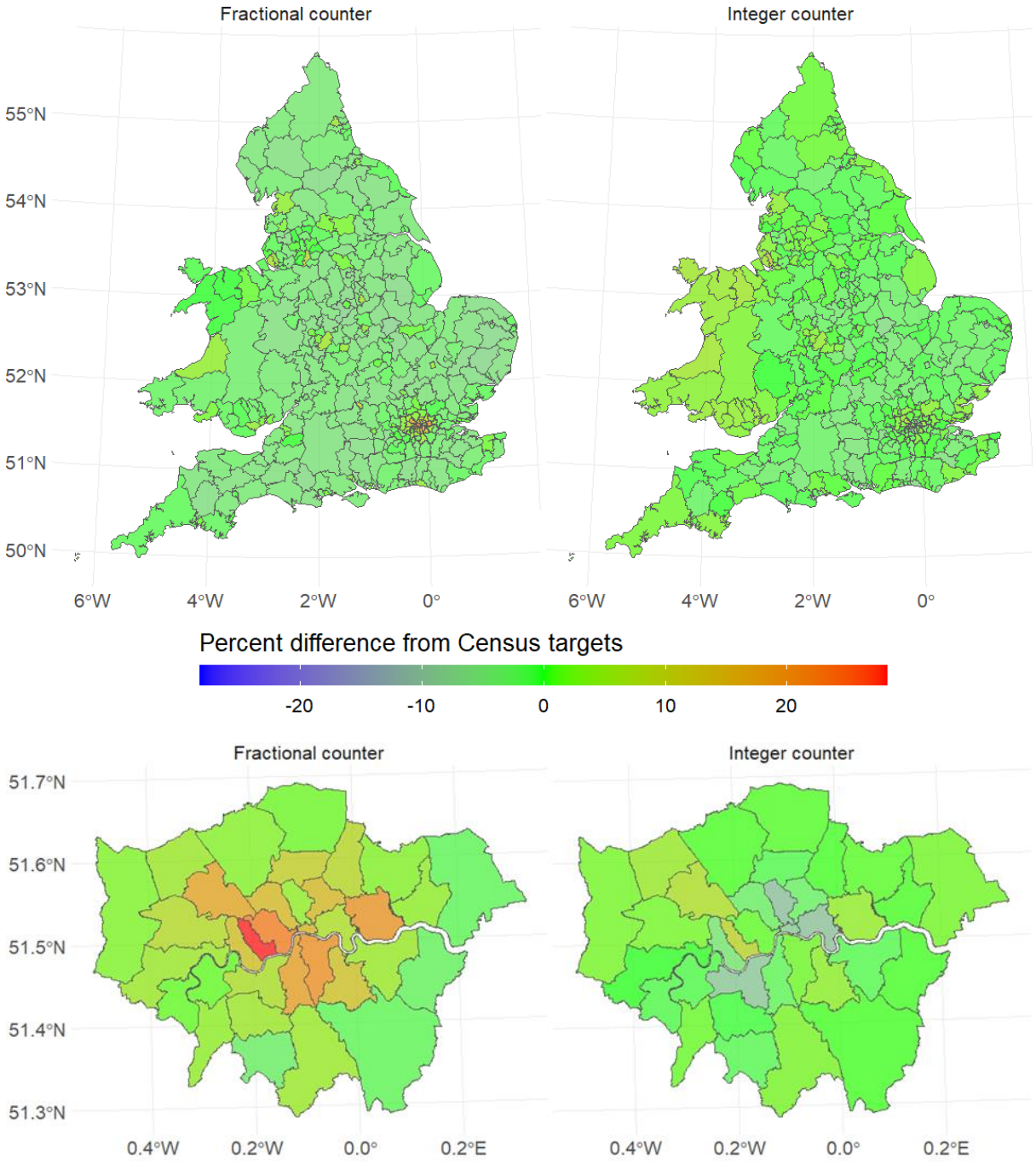

Figure 12 Choropleth map showing percentage difference from Census 2011 microdata benchmark for the integer and fractional counters for LA populations. Lower panels depict London LAs.

Both counting methods produced broadly similar results for females, with overestimates in many London LAs. However, the two methods behaved very differently for males, with the fractional counter largely overweighting and the integer counter largely underweighting. The overweighting of females and underweighting of males by the integer counter cancel each other out, leading to overall LA weights closer to the target population than those produced by the fractional counter.

LA totals by age

Disaggregating LA-level weights by 5-year age group leads to a larger gap between the fractional and integer counters than disaggregating by sex (Table 10).

Table 10. Error metrics for the integer and fractional counter for population disaggregated by LA and 5-year age groups.

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0825	0.0413	663	332
Integer	0.1260	0.0700	1013	563

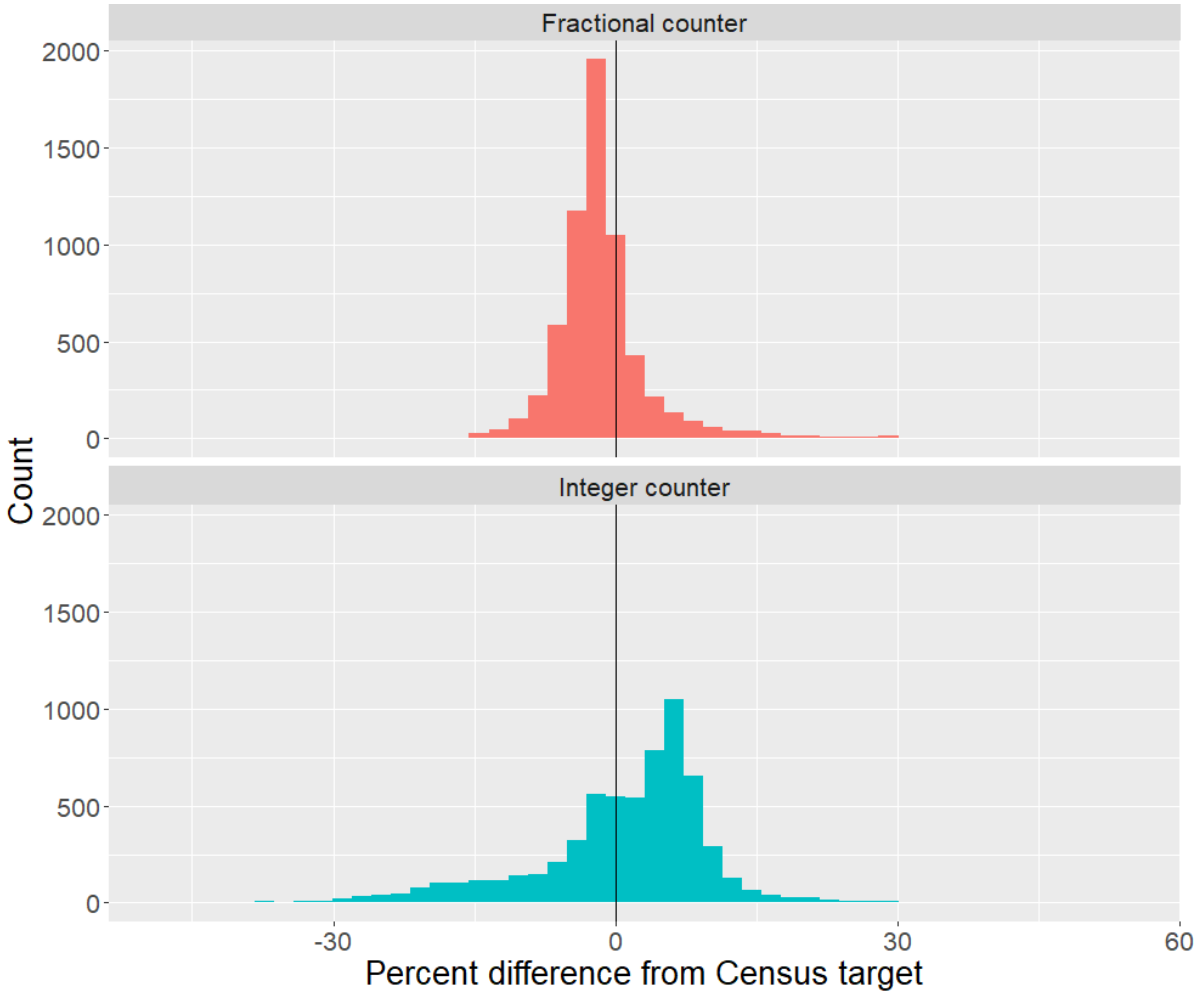

Figure 13 Frequency histogram of percentage differences between Census 2011 microdata benchmark and LA population (disaggregated by five year age groups) for the integer counter (blue) and fractional counter (red)

LA totals by age and sex

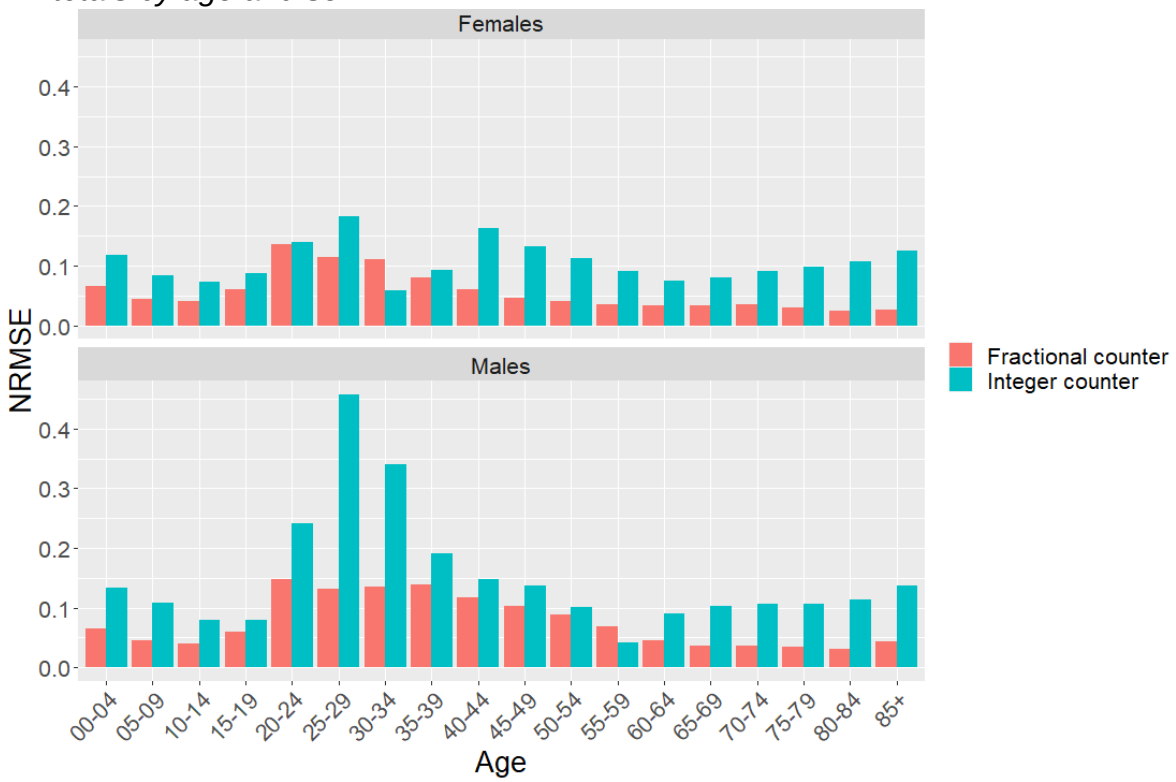

Figure 14 Bar plot showing normalised root mean square error (NRMSE) for LA population (disaggregated by five-year age group and sex) between the Census 2011 microdata benchmark and the integer counter (blue) and fractional counter (red).

LA totals by age, sex and ethnicity

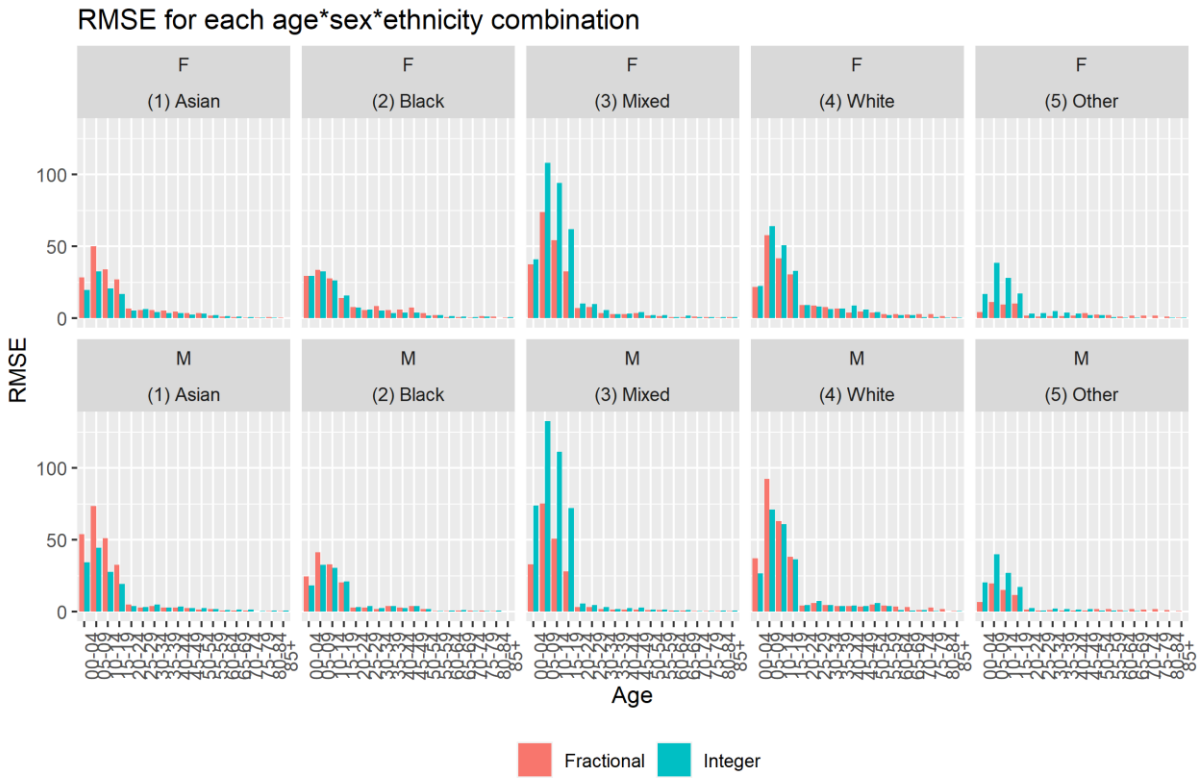

Figure 15 Bar plot showing root mean square error (RSME) for LA population (disaggregated by five-year age groups, sex, and ethnicity) between the Census 2011 microdata benchmark and the integer counter (blue) and fractional counter (red).

Annex B: Analysis of the geographic placement and ethnicity models

The results so far have explored how closely the overall fractional and integer counters match target population sizes at various levels of disaggregation. For the overall counters, the weights for stages 2 (S2; placement) and 3 (S3; ethnicity) are multiplied by the stage 1 (S1) inclusion weights. However, the different sets of inclusion weights for the integer and fractional counters make it difficult to directly compare the performance of S2 and S3.

This section of the results will instead assess S2 and S3 in isolation, by taking a standardised starting population and comparing how it is distributed. Alongside the fractional and integer models for S2 and S3, rules-based integer (RBI) methods are also assessed. The S2 RBI method was developed by PMST and uses a hierarchy of sources to assign individuals to the location associated with the highest-ranking source. The S3 RBI method was developed by SSAF and assigns weight to the most recently recorded valid ethnicity while honouring refusals.

Standardised placement models

To compare the performances of the integer, fractional and rules-based placement methods, each method was applied to individuals labelled by Census 2011 microdata as belonging to the target population. In effect, the S1 true labels were used as standardised population inclusion weights. The 'target population' (TP) EPD subset consists of 51.5 million location-level records for 50.4 million individuals.

Target population subset

LA totals (TP)

The placement weights for the TP subset were aggregated by LA. Error metrics for LA-level weight totals were then calculated, using the counts of census microdata in the target population as truth values (Table 11). The modelled integer method had the best performance, with slightly better metrics than the modelled fractional method. The RBI method was worse than either of the modelled approaches.

Table 11. Error metrics for the integer, fractional, and rules-based counters for LA population totals ('target population' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0114	0.0066	1656	960
Integer	0.0099	0.0059	1426	851
Rules-based integer	0.0155	0.0086	2240	1245

Plotting histograms of the percentage difference between the weights allocated by each method to LAs and the target weights shows that all three methods slightly underestimated most LAs whilst heavily overestimating a small handful. This trend

was exaggerated for the rules-based integer counter, with 73.2% of LA populations underestimated.

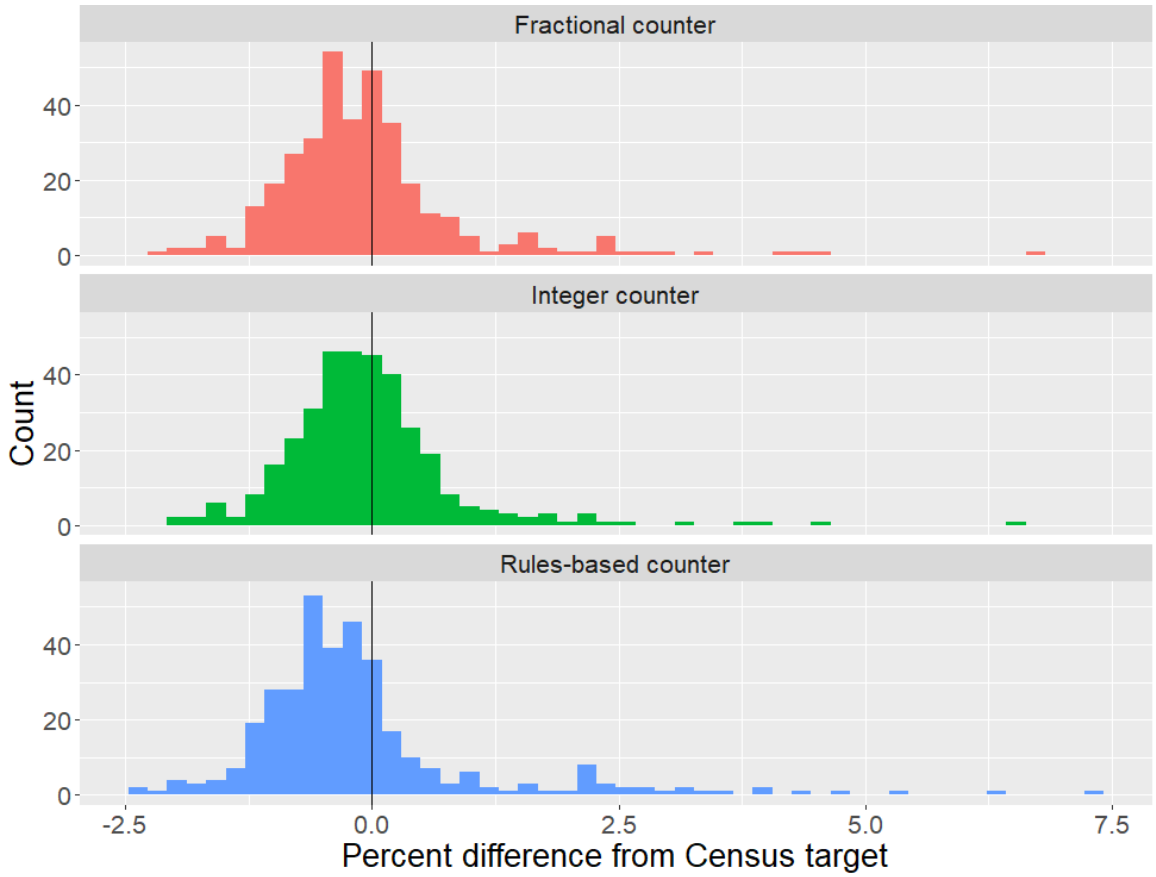

Figure 16 Frequency histogram of percentage difference from Census 2011 microdata benchmark for LA population totals for the model-based integer counter (green), rules-based integer counter (blue), and fractional counter (red) ('target population' subset only).

LA totals by age and sex (TP)

The integer S2 model was also the most accurate when disaggregating by sex and 5-year age group (Table 12; Fig 17). This is in contrast with the combined S1*S2 weights trend in the main results, where the fractional counter outperformed the integer counter for disaggregated populations.

Table 12. Error metrics for the integer, fractional, and rules-based counters for LA population disaggregated by age and sex ('target population' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0341	0.0101	137	40
Integer	0.0289	0.0090	116	36
Rules-based integer	0.0464	0.0125	187	50

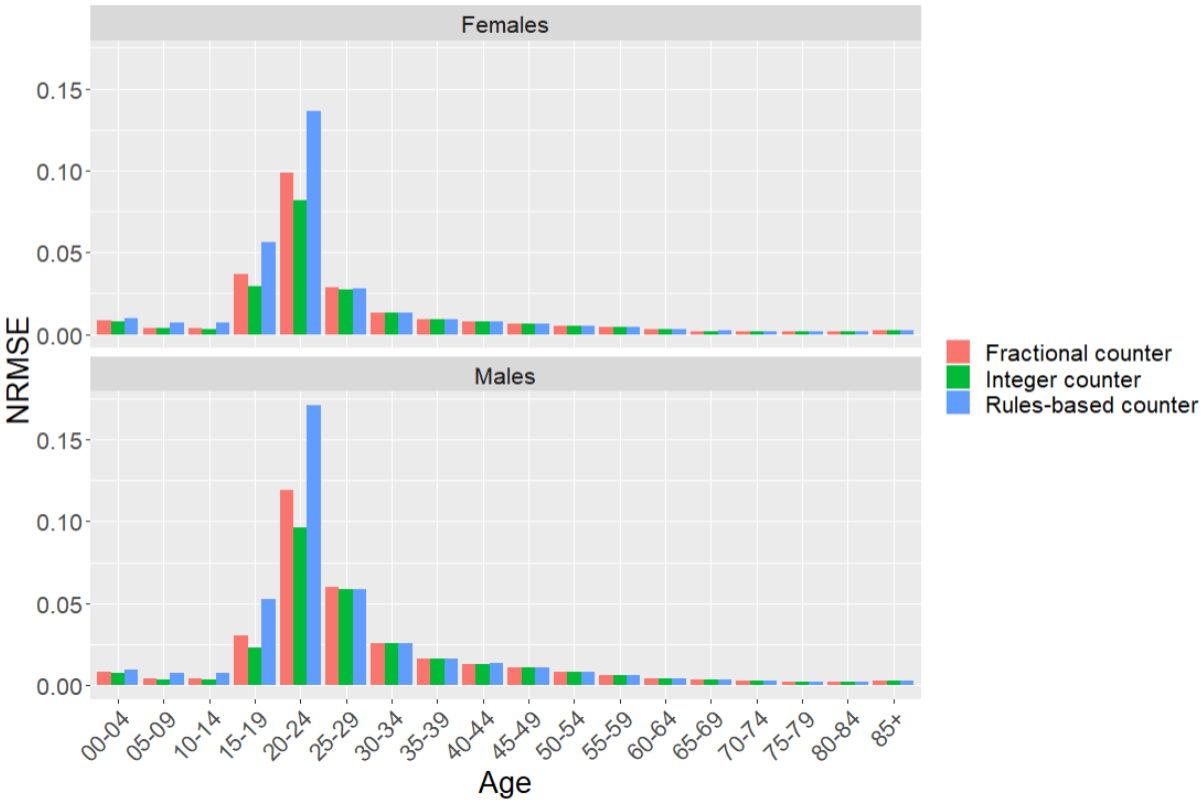

*Figure 17 Bar plot showing normalised mean absolute error (NMAE) between Census targets and predicted placement weight totals for each age*sex group across LAs. Weight totals derived from individuals in 'target population' subset of EPD.*

True address subset

LA totals (TA)

Although considering the TP subset allows a standardised comparison of placement models, this assessment has a major limitation: over 3.5 million individuals in this subset do not have a true address recorded in the EPD. Thus, at the OA level, the placement methods are unable to assign weight to the correct address. Allocation of weight to the correct LA is therefore a largely random process that does not reflect the relative abilities of each method to select addresses correctly.

To account for this limitation, a second analysis was conducted considering only individuals with a census-matched true address in the EPD. By comparing how each method distributes weights on this 'true address' (TA) subset, the ability of each method to accurately place individuals can be more directly assessed. The TA subset consists of 47.8 million address-level records for 46.8 million individuals.

Aggregating TA subset placement weights by LA and calculating error metrics revealed that the fractional modelled S2 assigned weights more accurately than the integer modelled approach (Table 12). This contrasts with the superior performance of the integer model found by analysing the TP subset. Again, the RBI method was far less accurate.

Table 12. Error metrics for the integer, fractional, and rules-based counters for LA population totals ('true address' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0013	0.0007	172	88
Integer	0.0016	0.0008	213	113
Rules-based integer	0.0064	0.0035	868	469

Unsurprisingly, the LA population predictions were much closer to the census counts for the TA subset than the TP subset for all three methods, as the latter contains an additional 3.5 million individuals who cannot be correctly placed.

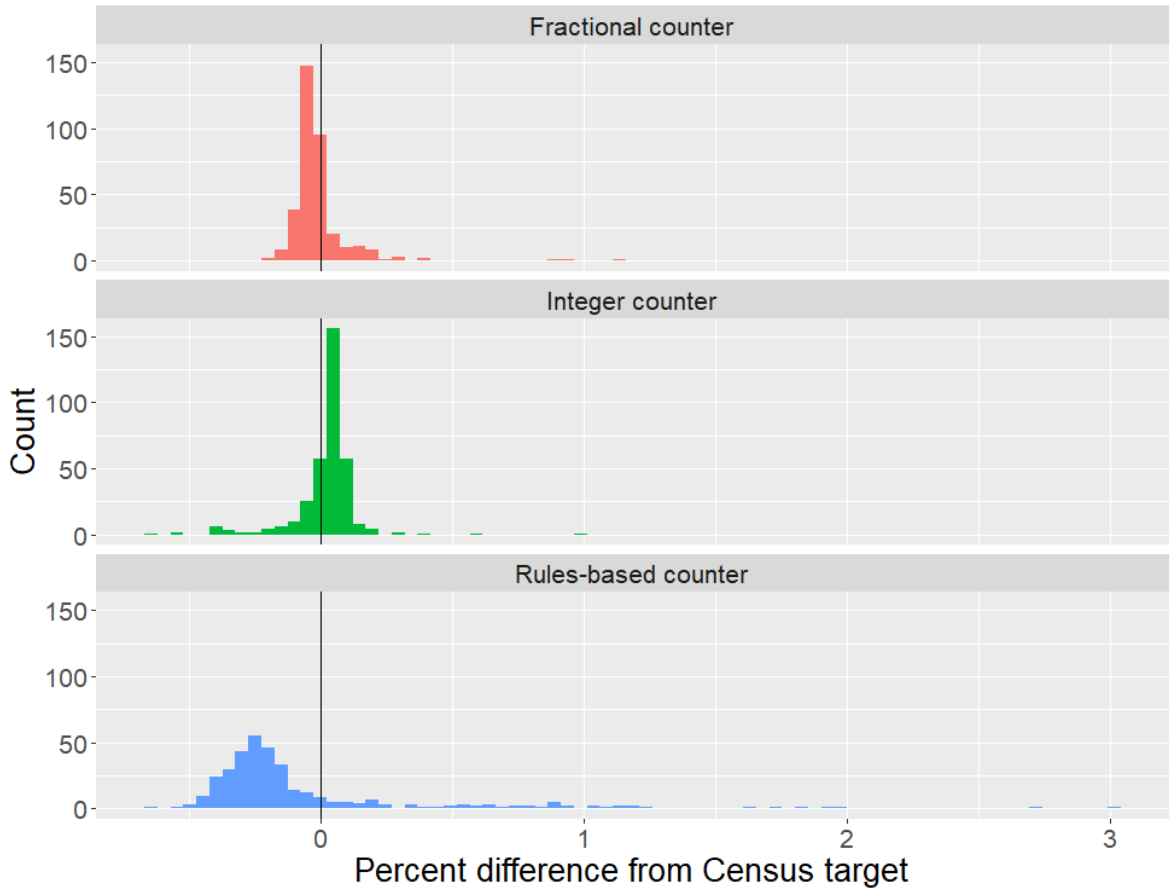

Figure 1818 Frequency histogram of percentage difference from Census 2011 microdata benchmark for LA population totals for the model-based integer counter (green), rules-based integer counter (blue), and fractional counter (red) ('true address' subset only).

LA totals by age and sex (TA)

Further disaggregation by 5-year age group and sex on the TA subset produced similar results: the fractional model slightly outperformed the integer model, whilst the RBI method was substantially less accurate (Table 13).

Table 13. Error metrics for the integer, fractional, and rules-based counters for LA population disaggregated by age and sex ('true address' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0039	0.0010	15	4

Integer	0.0056	0.0013	21	5
Rules-based integer	0.0194	0.0039	73	15

Again, error metrics were calculated separately for each sex*5-year age group in the TA subset and plotted. One striking difference from the equivalent plot for the TP subset (Fig 17) is the far wider performance gap between the two modelling methods and the RBI method (Fig 19).

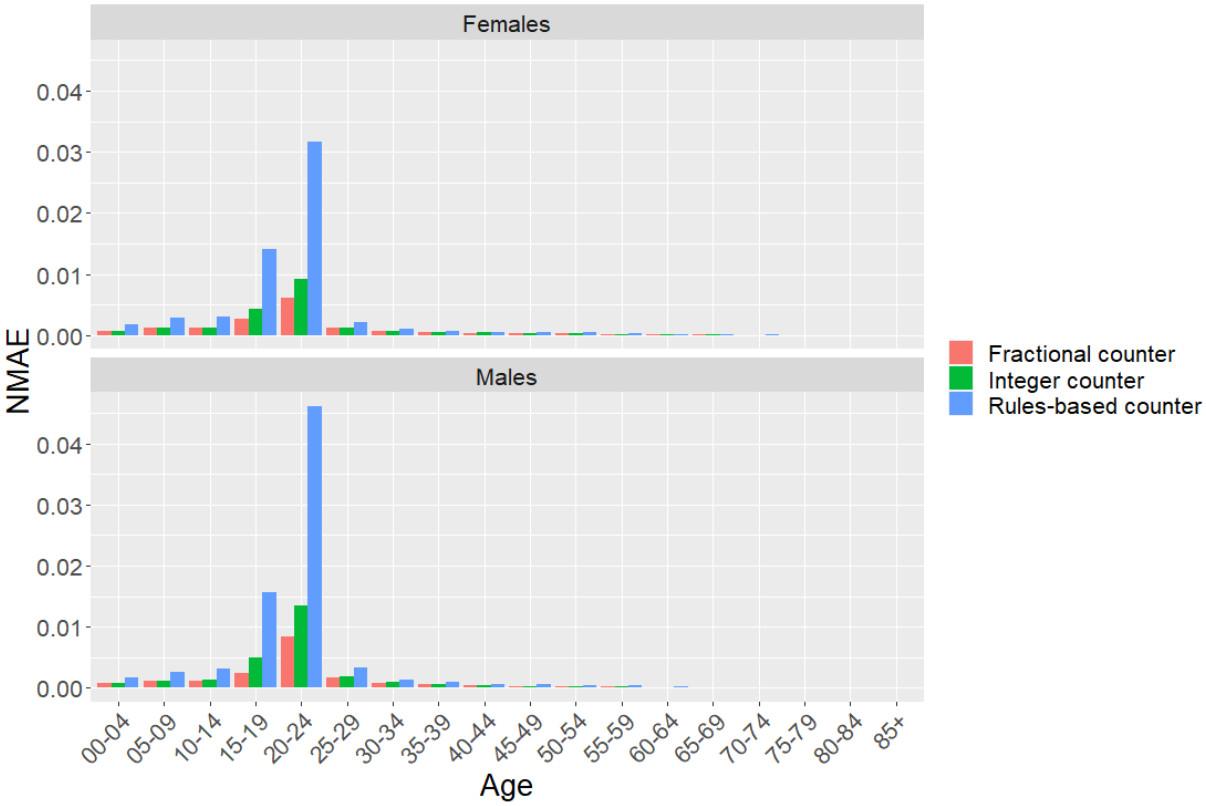

*Figure 19: Bar plot showing normalised mean absolute error (NMAE) between Census targets and predicted placement weight totals for each age*sex group across LAs. Weight totals derived from individuals in 'true address' subset of EPD.*

The standardised analysis of placement methods for the TA subset demonstrates that when a correct address is available, the two modelled approaches vastly outperform the rules-based hierarchy approach. However, for 45.8 million of the 46.8 million individuals in the TA subset, the only address recorded is the census-matched true address. As all three placement methods work by distributing weight across the addresses available, addresses with no conflict receive the full weight of the associated individual. Thus, for 97.8% of individuals in the TA subset, weights are assigned perfectly by default. All the differences between methods shown so far reflect differences in weighting a small minority of individuals with address conflicts.

True address with conflicts subset

LA totals (TAC)

An additional analysis was performed on individuals who had both a true address

and an address conflict. This ‘true address + conflict’ (TAC) subset is unique in that all individuals have both correct and incorrect addresses recorded, between which each method must discriminate.

As expected, the normalised error metrics calculated on the TAC subset are far higher than those calculated on the TA subset (Table 14). This is due to the normalisation process, whereby scores are divided by the mean of the true values.

Table 14. Error metrics for the integer, fractional, and rules-based counters for LA population totals (‘true address with conflicts’ subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0601	0.0307	171	87
Integer	0.0744	0.0395	212	112
Rules-based integer	0.3037	0.1640	867	468

Although the percentage differences between predicted and actual population counts have a far wider range for the TAC subset than the TA subset, the overall shape of the distributions is broadly similar for all three methods.

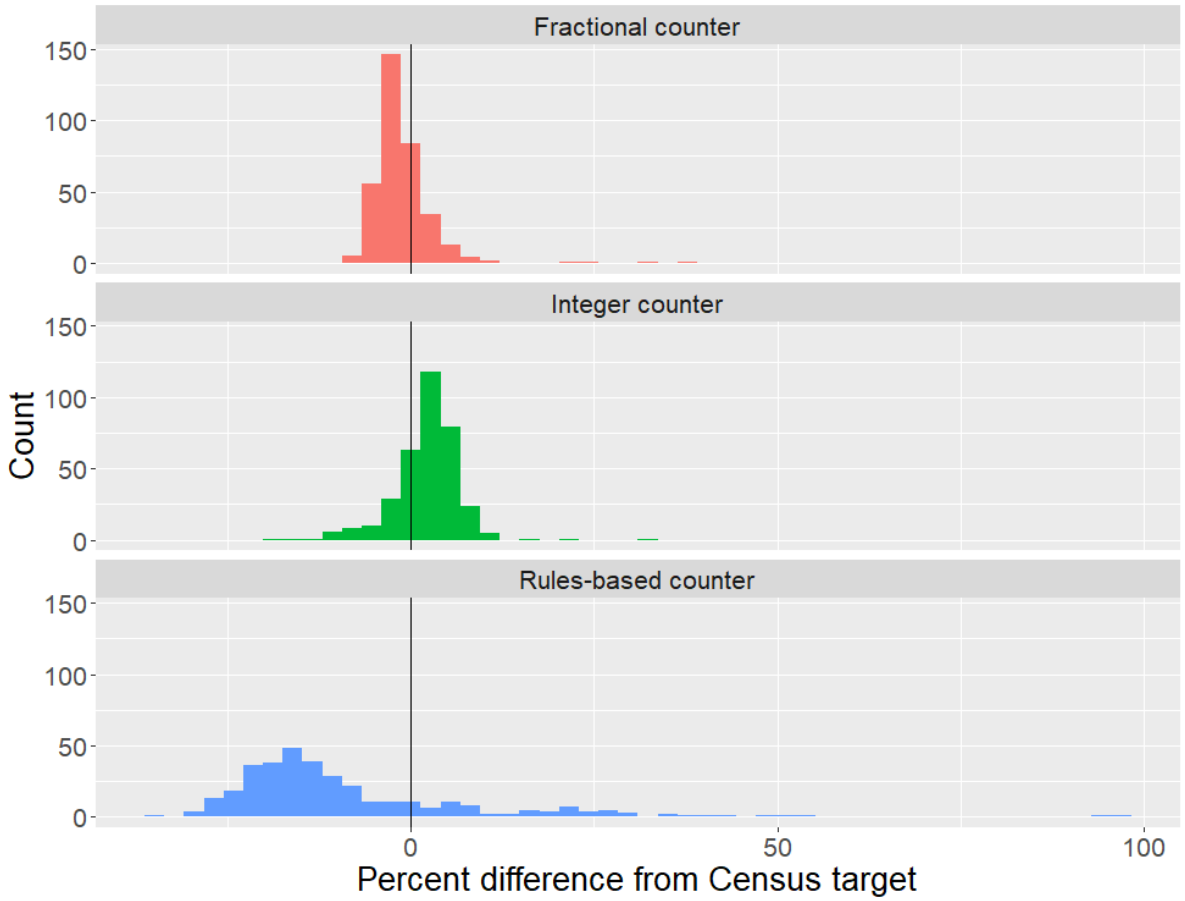

Figure 20: Frequency histogram for % difference from census targets for modelled fractional, modelled integer and rules-based integer placement weights for LA-level totals. Weight totals derived from individuals in ‘true address + conflict’ subset of EPD.

LA totals by age and sex (TAC)

Again, results were calculated with the data further disaggregated by LA, age and sex. The fractional model remained the highest performing at this level of granularity, with the rules-based method performing substantially worse than the two modelling approaches across all sub-groups (Table 15; Fig 21).

Table 15. Error metrics for the integer, fractional, and rules-based counters for LA population disaggregated by age and sex ('true address with conflicts' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.1611	0.0462	16	4
Integer	0.2286	0.0635	23	6
Rules-based integer	0.7977	0.1856	83	19

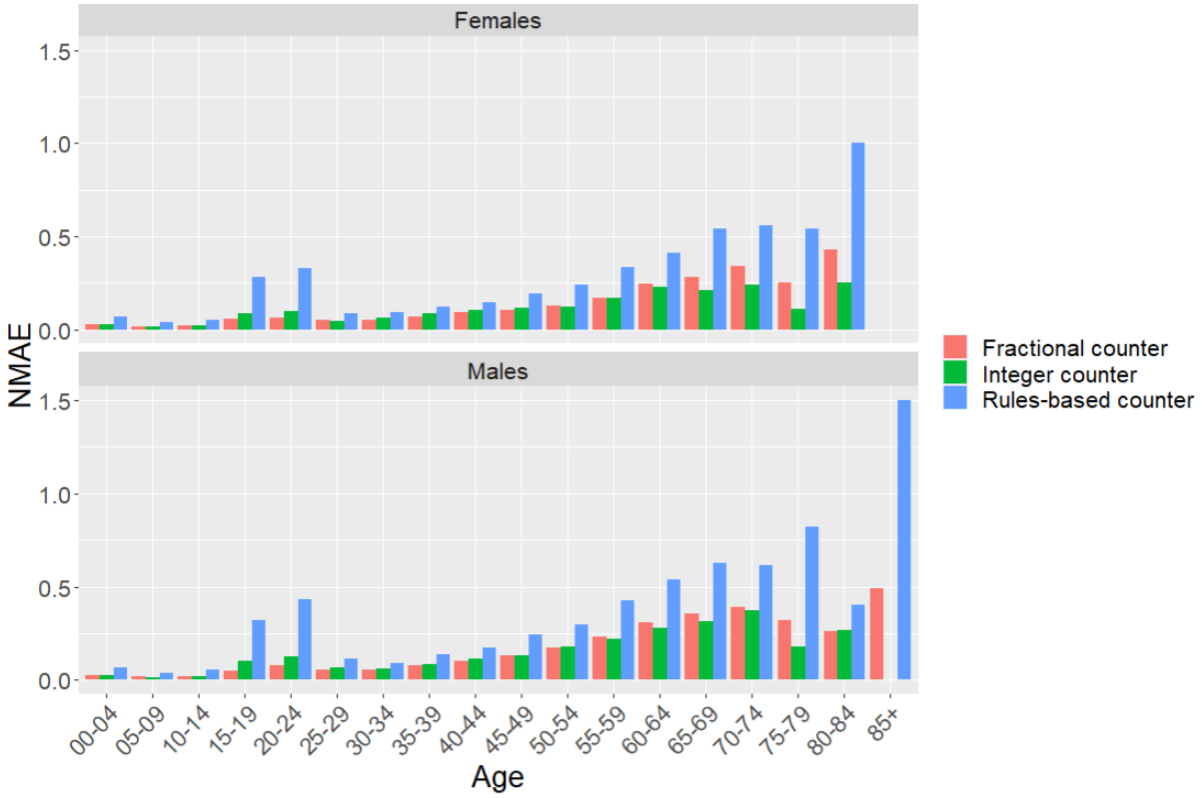

*Figure 21 Bar plot showing normalised mean absolute error (NMAE) between Census targets and predicted placement weight totals for each age*sex group across LAs. Weight totals derived from individuals in 'true address + conflict' subset of EPD.*

Summary of standardised placement analysis

The most consistent result from the analysis of geographic placement methods is the superior performance of the two modelling methods compared to the rules-based integer method. There are also key differences between the S2 integer counter and fractional counter. Although the integer counter performed better when considering all individuals in the target population, this included a large portion of individuals for whom no correct address is available for selection by the models. For such

individuals, accurate placement weights are largely coincidental and do not reflect the ability of the supervised learning methods used here. Meanwhile, when only considering individuals with a correct address recorded, the fractional counter outperforms the integer counter. This was true at both LA level and LA*age*sex level, with a slightly greater performance gap at higher levels of disaggregation.

Standardised ethnicity models

As with the standardised placement investigation, to compare the performances of the integer, fractional and rules-based ethnicity methods, each method was applied to individuals labelled by Census 2011 microdata as belonging to the target population. In effect, the S1 true labels were used as standardised population inclusion weights. To standardise S2 weights when breaking down the results by LA we used a 'true address' flag sourced from Census 2011 microdata to place individuals consistently across the range of ethnicity methods.

We first present the results at the national level across the three EPD subsets (TP, TE, TEC; see Methods), and then at the LA level.

National results

Target population subset (national)

National ethnicity (TP)

Looking at the target population (TP) subset of the standardised ethnicity modelling results, there appears to be little difference between the performance of the fractional and integer counters (Table BB11). The rules-based integer on the other hand appears to perform slightly better than both methods at the aggregate total ethnicity level.

Table 16. Error metrics for the integer, fractional, and rules-based counters for national population by five-category ethnicity ('target population' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.03784	0.03435	119971	108929
Integer	0.03775	0.03435	119703	108929
Rules-based integer	0.03125	0.02543	99090	80621

The differences in performances of the standardised methods are also seen at the 5-category national level (Table 17). There is little to separate the fractional and integer counters, with the fractional counter estimates closer for 3 of the categories and the integer counter closer for 2. Again, at this level the rules-based integer model produces closer estimates for all but the Mixed category. However, the differences at this aggregation appear small.

Table 17. Percentage distribution of five-category ethnicity for the integer, fractional, and rules-based counters for national population ('target population' subset)

	Asian (1)	Black (2)	Mixed (3)	White (4)	Other (5)
Target Population	7.51	3.49	1.60	87.27	0.13
Fractional Counter	7.60	3.68	2.21	85.33	1.18
% Difference (fractional)	0.08	0.19	0.61	-1.94	1.05
Integer Counter	7.60	3.68	2.23	85.32	1.17
% Difference (integer)	0.07	0.19	0.63	-1.97	1.07
Rules-based integer	7.54	3.65	2.13	85.39	1.29
% Difference (r-b integer)	0.03	0.16	0.54	-1.88	1.15

National ethnicity by sex (TP)

Looking at the target population (TP) subset of the standardised ethnicity modelling results at the ethnicity by sex level, there again appears to be little difference between the performance of the fractional and integer counters. The rules-based integer performs slightly better than both methods at the aggregate total ethnicity by sex level (Table 18).

Table 18. Error metrics for the integer, fractional, and rules-based counters for national population by five-category ethnicity and sex ('target population' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0380	0.0344	60234	54464
Integer	0.0379	0.0344	60099	54464
Rules-based integer	0.0314	0.0254	49783	40311

National ethnicity by age and sex (TP)

When disaggregating the target population (TP) subset of the standardised ethnicity modelling results further by age, there is little difference between the performance of the fractional and integer counters. The fractional counter errors are slightly smaller, with the rules-based integer performing slightly better than both modelled methods at the aggregate total ethnicity by sex level (Table 19).

Table 19. Error metrics for the integer, fractional, and rules-based counters for national population by five-category ethnicity, age, and sex ('target population' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0532	0.0344	847	547
Integer	0.0533	0.0344	849	547
Rules-based integer	0.0484	0.0294	771	468

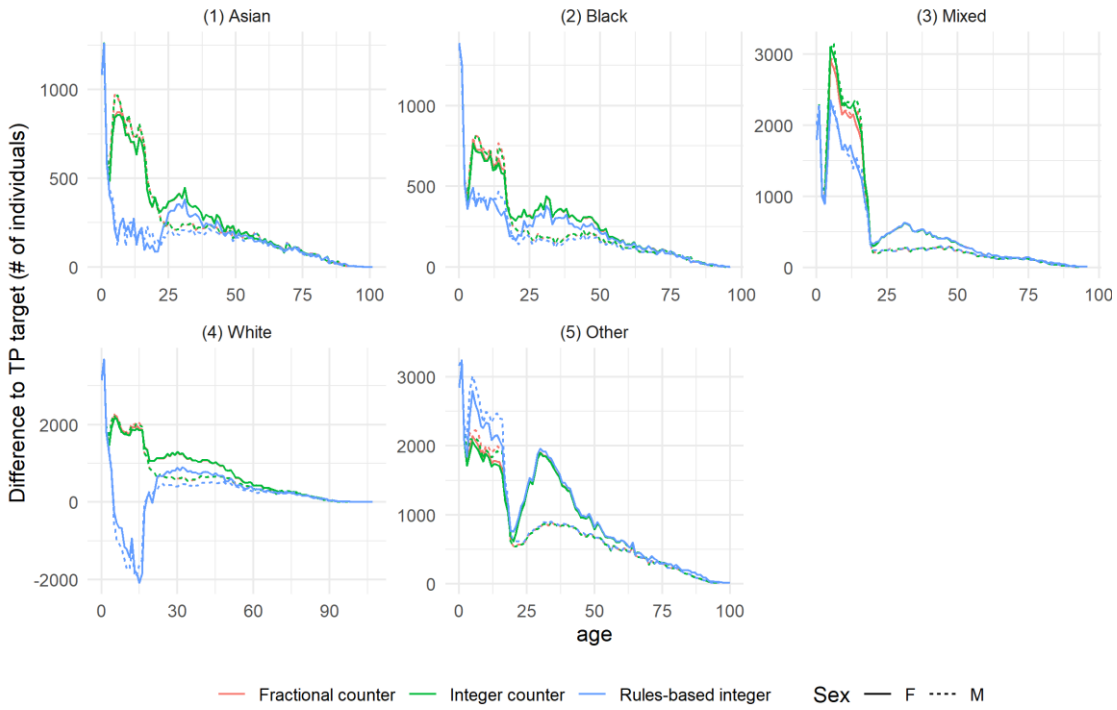

Figure 22 Line plot showing the absolute differences between the target population (TP) subset and the fractional counter (red-line), integer counter (green-line) and rules-based integer (blue-line) approaches for both male (dashed-line) and female (solid-line) for the 5 categories of ethnicity.

The differences between the standardised fractional/integer/rules-based integer ethnicity predictions, by age and sex, and the TP subpopulation are shown in Figure 22. There is little difference between the fractional and integer weight estimates by age and sex. The rules-based integer counter produces smaller differences to the TP target for categories 1-3 (Asian, Black, and mixed), however, for category 4 (White) the rules-based approach underestimates for individuals 5-18 years old where the fractional and integer approaches broadly overestimate.

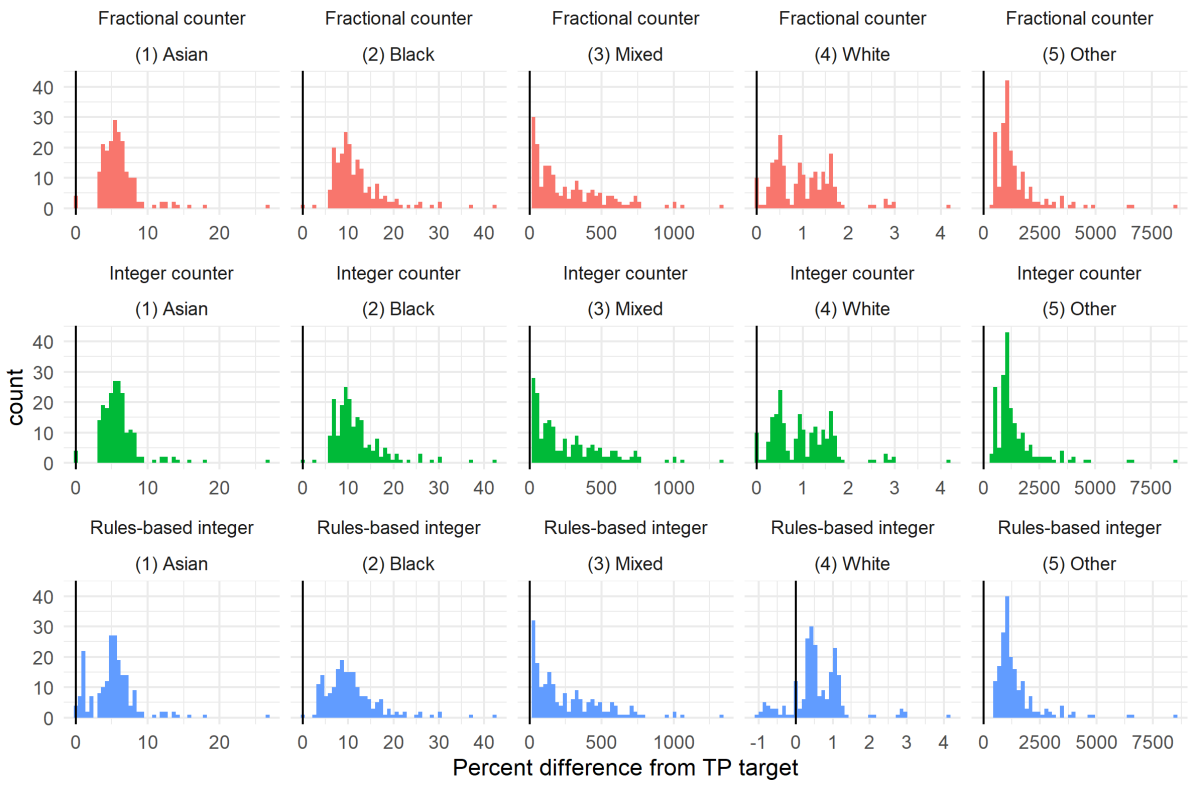

*Figure 23 Frequency histogram of the percentage difference between admin-based population counts by ethnicity*age*sex and the target population (TP) benchmark for the integer counter (green), fractional counter (red) and rules-based integer (blue).*

Looking at the distributions of differences from the TP target for the different approaches in Figure 23, there is again little difference between the fractional and integer approaches. The rules-based integer underestimation for category 4 (White) individuals is also visible.

True ethnicity subset (national)

National ethnicity (TE)

As with the standardised placement analysis, we consider the subpopulation of individuals who have a census-matched true ethnicity (TE) recorded in the EPD. Considering these individuals, the fractional counter performs far better than the integer counter and both perform better than the rules-based integer approach (Table 20).

Table 20. Error metrics for the integer, fractional, and rules-based counters for national population by five-category ethnicity ('true ethnicity' subset).

Ethnicity weight totals, TE subset				
	NRMSE	**NMAE**	**RMSE**	**MAE**
Fractional	0.00016	0.000147	508	466
Integer	0.00058	0.000459	1846	1457
Rules-based integer	0.016	0.0106	50662	33590

The improvement in performance of the fractional counter versus the modelled and rules-based integer approaches can also be seen when looking at the national 5-category aggregated ethnicity distributions. The differences between the fractional distribution and the target population distribution are smaller than the differences for both integer approaches (Table 21).

Table 21. Percentage distribution of five-category ethnicity for the integer, fractional, and rules-based counters for national population ('true ethnicity' subset)

	Asian (1)	Black (2)	Mixed (3)	White (4)	Other (5)
Target Population	7.51	3.49	1.60	87.27	0.13
Fractional Counter	7.51	3.49	1.60	87.27	0.14
% Difference (fractional)	0.00	0.00	0.00	0.00	0.00
Integer Counter	7.51	3.48	1.62	87.26	0.12
% Difference (integer)	0.00	0.00	0.02	-0.01	-0.01
Rules-based integer	7.45	3.45	1.53	87.33	0.24
% Difference (r-b integer)	-0.06	-0.04	-0.07	0.06	0.10

National ethnicity by sex (TE)

Disaggregating the TE subset of the ethnicity totals by sex, we again see an improvement in the fractional counter metrics versus both integer approaches. The modelled integer counter performs worse than the fractional counter but significantly better than the rules-based integer approach (Table 21).

Table 21. Error metrics for the integer, fractional, and rules-based counters for national population by five-category ethnicity and sex ('true ethnicity' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.000201	0.000176	318	279
Integer	0.000586	0.000459	928	728
Rules-based integer	0.016	0.0106	25349	16795

National ethnicity age and sex (TE)

Further disaggregating the TE subset by age, the fractional counter performs better than the model integer counter and rules-based integer approach across all the available metrics (Table 22). Figures 24 and 25 show the differences between the predicted ethnicity x sex x ages weights and the TE target. The fractional counter clearly performs better, with smaller differences across all ages/categories than either alternative method. The integer counter is broadly similar, but with increased underestimation for category 2 (Black) and 5 (Other) individuals and an overestimation for category 3 (Mixed) individuals. However, the rules-based integer approach underestimates across most of the ethnicity categories, and overestimates across the age range for the final category.

Table 22. Error metrics for the integer, fractional, and rules-based counters for national population by five-category ethnicity, age, and sex ('true ethnicity' subset).

Counter	NRMSE	NMAE	RMSE	MAE

Fractional	0.00056	0.000249	9	4
Integer	0.00163	0.000575	26	9
Rules-based integer	0.0342	0.0107	551	172

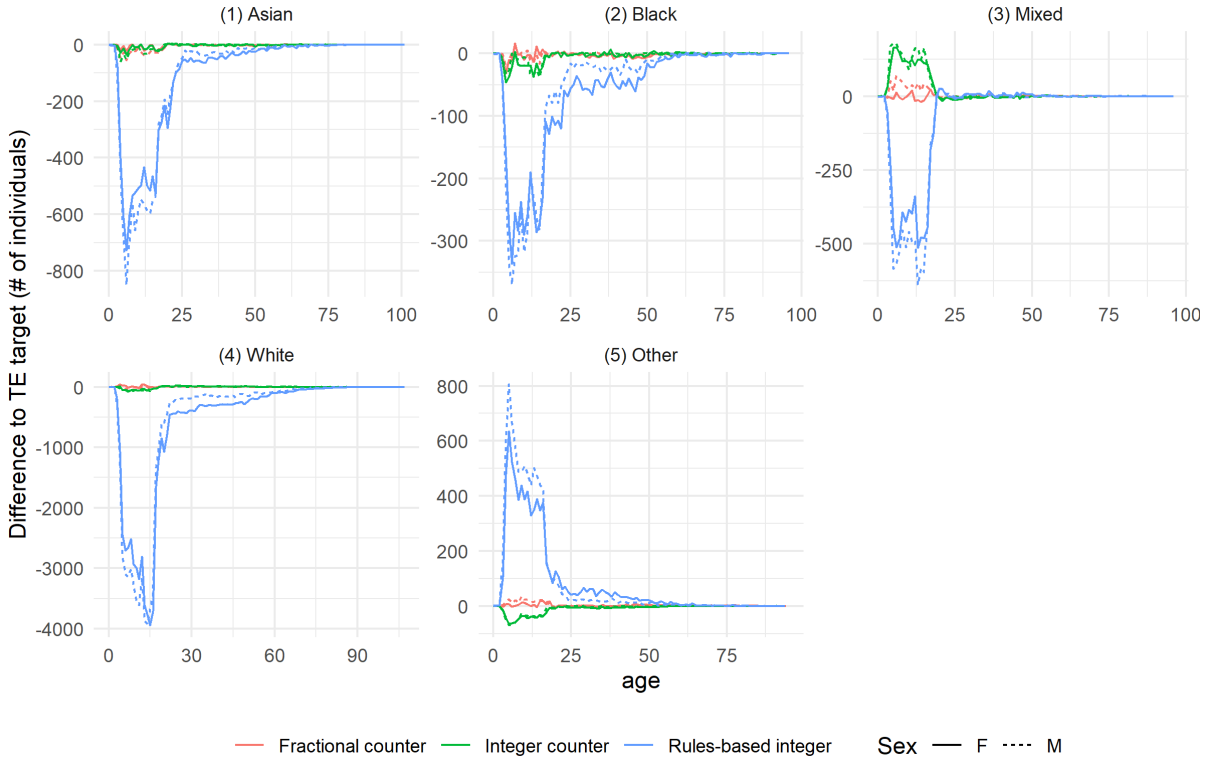

Figure 24 Line plot showing the absolute differences between the true ethnicity (TE) subset and the fractional counter (red-line), integer counter (green-line) and rules-based integer (blue-line) approaches for both male (dashed-line) and female (solid-line) for the 5 categories of ethnicity.

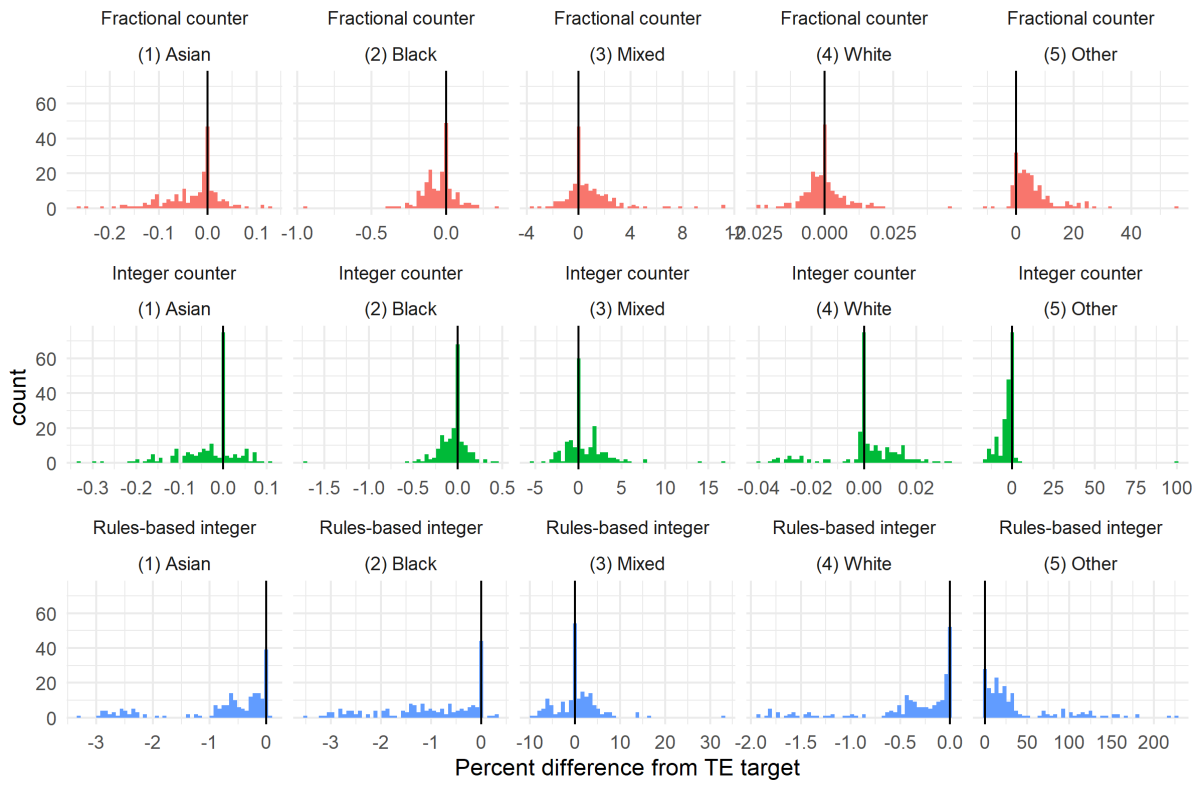

*Figure 25 Frequency histogram of the percentage difference between admin-based population counts by ethnicity*age*sex and the true ethnicity (TE) benchmark for the integer counter (green), fractional counter (red) and rules-based integer (blue).*

True ethnicity with conflicts subset (national)

National ethnicity (TEC)

We further subset the individuals to those that are in the previous TE subgroup but also have conflicting records identified in the admin-sources, to create the true ethnicity with conflicts (TEC) subset. For this subset, the fractional counter performs significantly better than both alternatives, followed by the modelled integer counter, with the rules-based integer approach performing the worst (Table 23).

Table 23. Error metrics for the integer, fractional, and rules-based counters for national population by five-category ethnicity ('true ethnicity with conflicts' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0388	0.0355	508	466
Integer	0.1407	0.1110	1846	1457
Rules-based integer	0.6854	0.5822	8991	7637

Looking at the aggregated 5-category ethnicity distributions for the TEC subset, we see that the fractional counter is closer for 4 out of the 5 categories, with the integer counter closer for a single category and the rules-based integer approach distribution being the furthest from the TEC target (Table 24).

Table 24. Percentage distribution of five-category ethnicity for the integer, fractional, and rules-based counters for national population ('true ethnicity with conflicts' subset)

	Asian (1)	Black (2)	Mixed (3)	White (4)	Other (5)
Target Population	22.81	15.04	29.04	30.52	2.59
Fractional Counter	21.74	14.57	29.85	30.28	3.56
% Difference (fractional)	-1.07	-0.47	0.81	-0.24	0.96
Integer Counter	21.79	14.06	34.59	29.22	0.34
% Difference (integer)	-1.02	-0.99	5.55	-1.29	-2.26
Rules-based integer	11.47	9.69	16.60	34.96	27.27
% Difference (r-b integer)	-11.33	-5.35	-12.44	4.45	24.68

National ethnicity by sex (TEC)

Disaggregating the TEC subset by sex, the fractional counter continues to outperform the integer counter and the rules-based integer approach (Table 25).

Table 25. Error metrics for the integer, fractional, and rules-based counters for national population by five-category ethnicity and sex ('true ethnicity with conflicts' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0485	0.0425	318	279
Integer	0.1415	0.1110	928	728
Rules-based integer	0.6868	0.5822	4505	3819

National ethnicity by sex and age (TEC)

Further breaking down the predicted weights by age, the fractional counter performs the best, with the rules-based integer approach performing the worst (Table 26). This performance deficit by the rules-based integer approach is clear from Figures 26 and 27, with a broad underestimation across the age range for categories 1-3 (Asian, Black, and Mixed) by the rules-based approach combined with an overestimation for categories 4-5 (White and Other).

Table 26. Error metrics for the integer, fractional, and rules-based counters for national population by five-category ethnicity, age, and sex ('true ethnicity with conflicts' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.1253	0.0602	10	5
Integer	0.3655	0.1390	28	11
Rules-based integer	1.5042	0.6243	117	49

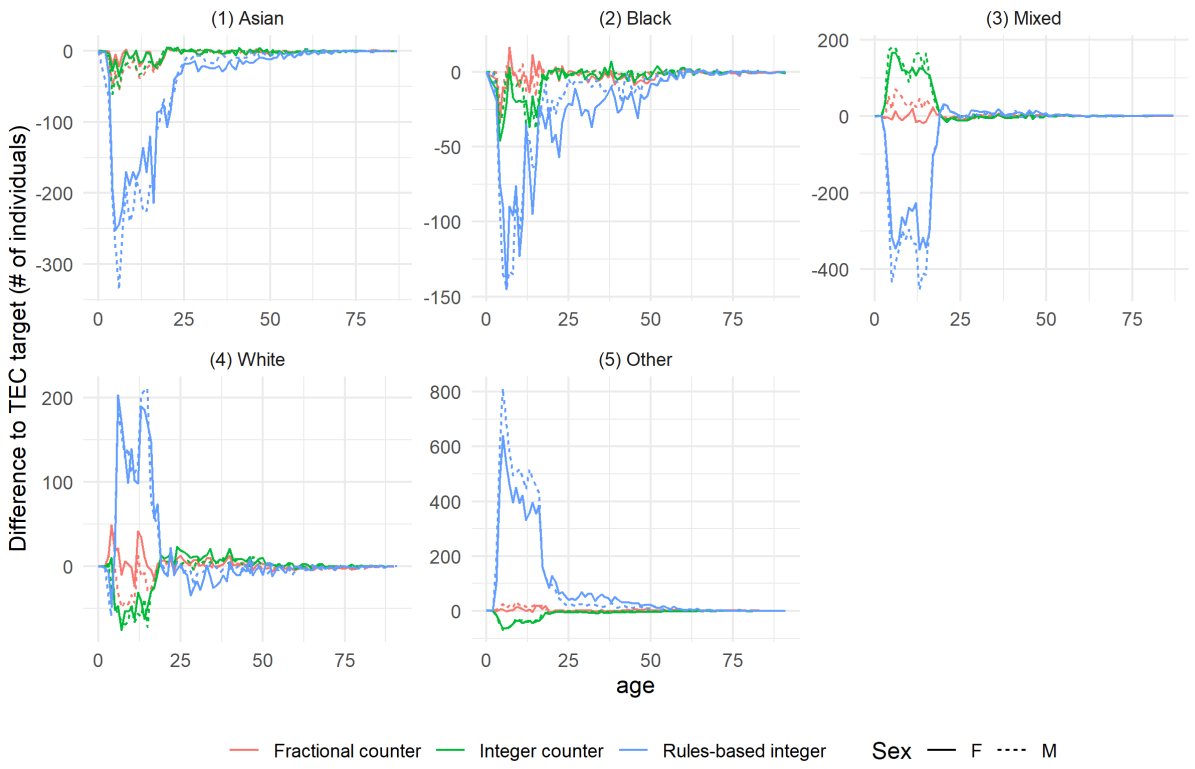

Figure 26 Line plot showing the absolute differences between the true ethnicity and conflicts (TEC) subset and the fractional counter (red-line), integer counter (green-line) and rules-based integer (blue-line) approaches for both male (dashed-line) and female (solid-line) for the 5 categories of ethnicity.

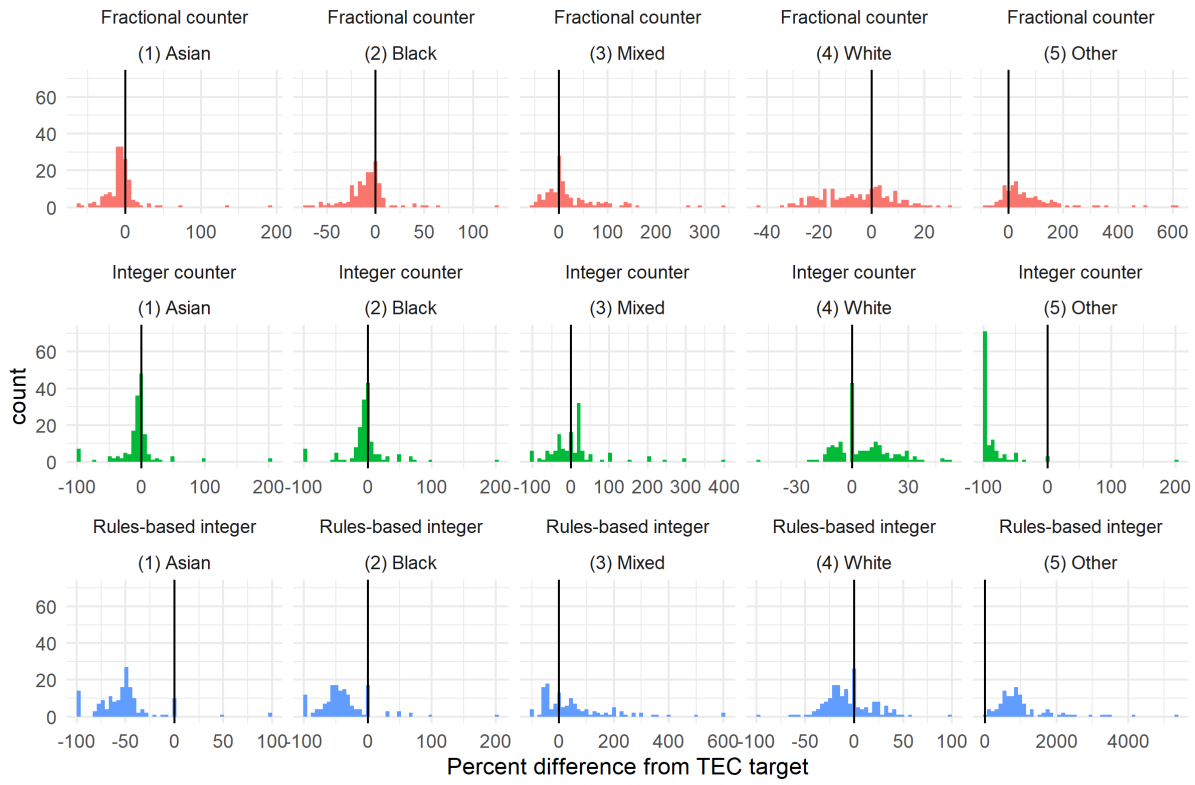

*Figure 27 Frequency histogram of the percentage difference between admin-based population counts by ethnicity*age*sex and the true ethnicity and conflicts (TEC) benchmark for the integer counter (green), fractional counter (red) and rules-based integer (blue).*

LA results

Target population subset (LA)

LA ethnicity (TP)

Sub-setting the TP ethnicity weights by LA, we find that the fractional and integer models perform the same, with the rules-based integer approach performing slightly better however this difference in performance compared to the modelled approaches is significantly smaller than that seen in the previous disaggregated metrics (Table 27).

Table 27. Error metrics for the integer, fractional, and rules-based counters for LA population by five-category ethnicity ('target population' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.0621	0.0336	2718	1470
Integer	0.0621	0.0336	2718	1470
Rules-based integer	0.0533	0.0260	2332	1139

LA ethnicity by age and sex (TP)

Disaggregating the ethnicity totals for the TP by LA, and sex we see that the fractional approach performs the best, closely followed by the integer counter, with the rules-based integer approach performing the worst (Table 28; Fig 28).

Table 28. Error metrics for the integer, fractional, and rules-based counters for LA population by five-category ethnicity, age, and sex ('target population' subset).

Counter	RMSE	MAE
Fractional	18	5
Integer	18	5
Rules-based integer	21	6

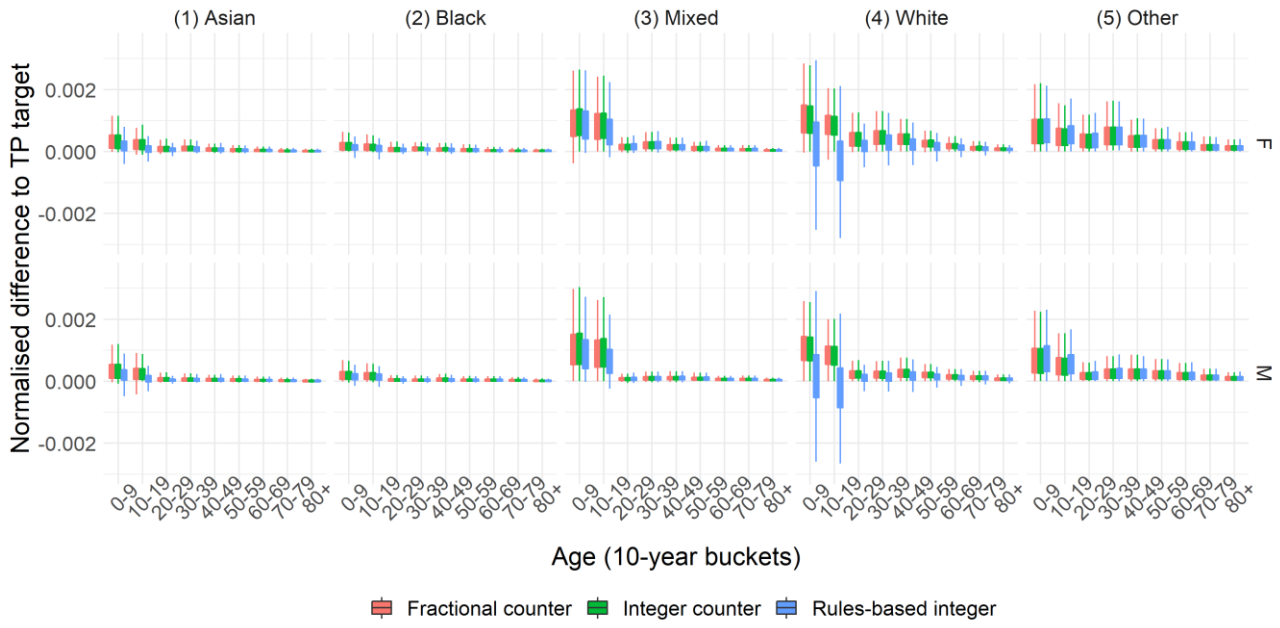

Figure 28 Box and whisker plot showing the distribution of normalised differences to the target population (TP) benchmark from the fractional (red) and integer (green) counters and the rules-based integer (blue) approach across 10-year age buckets for males (top panels) and females (bottom panels) and ethnicity. Outliers have been removed for clarity. Similar numbers of outliers found in each category, with the outliers for (4) White M/F being significantly larger than the other categories.

True ethnicity subset (LA)

LA ethnicity (TE)

Sub-setting the TE ethnicity weights by LA, we find that the fractional and integer models perform far better than the rules-based integer approach, with the fractional counter performing the best (Table 29).

Table 29. Error metrics for the integer, fractional, and rules-based counters for LA population by five-category ethnicity ('true ethnicity' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	1.04E-06	1.16E-07	0.0455	0.00508
Integer	1.23E-06	6.57E-08	0.0536	0.002874
Rules-based integer	0.0137	0.00872	599	382

LA ethnicity by age and sex (TE)

Further sub-setting the TE subpopulation by age and sex, we find that the fractional counter continues to perform the best, followed by the integer counter and with the rules-based approach performing the worst (Table 30; Fig 29).

Table 30. Error metrics for the integer, fractional, and rules-based counters for LA population by five-category ethnicity, age, and sex ('true ethnicity' subset).

Counter	RMSE	MAE
Fractional	0.828	0.211
Integer	1.102	0.237
Rules-based integer	14.22	3.096

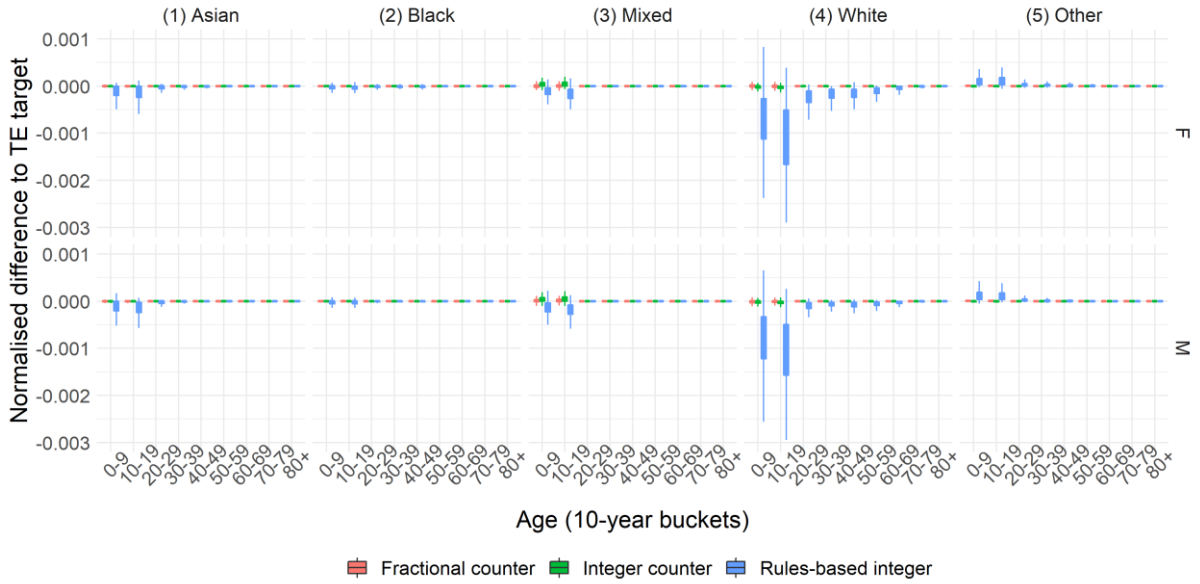

Figure 29 Box and whisker plot showing the distribution of normalised differences to the true ethnicity (TE) benchmark from the fractional (red) and integer (green) counters and the rules-based integer (blue) approach across 10-year age buckets for males (top panels) and females (bottom panels) and ethnicity. Outliers have been removed for clarity. Similar numbers of outliers found in each category, with the outliers for (4) White M/F being significantly larger than the other categories.

True ethnicity with conflicts subset (LA)

LA ethnicity (TEC)

Sub-setting the TEC ethnicity weights by LA, we find that the fractional and integer models perform far better than the rules-based integer approach, with the fractional counter performing the best according to the RMSE/NRMSE and the integer counter performing best when comparing the MAE/NMAE (Table 31).

Table 31. Error metrics for the integer, fractional, and rules-based counters for LA population by five-category ethnicity ('true ethnicity with conflicts' subset).

Counter	NRMSE	NMAE	RMSE	MAE
Fractional	0.000243	2.78E-05	0.0467	0.00534
Integer	0.000286	1.57E-05	0.055	0.00302
Rules-based integer	0.00683	0.00123	1.31	0.236

LA ethnicity by age and sex (TEC)

Further sub-setting the TEC subpopulation by age and sex, we find that the fractional counter continues to perform the best, followed by the integer counter and with the rules-based approach performing the worst (Table 32).

Table 32. Error metrics for the integer, fractional, and rules-based counters for LA population by five-category ethnicity, age, and sex ('true ethnicity with conflicts' subset).

Counter	RMSE	MAE
Fractional	1.549	0.736

Integer	2.061	0.827
Rules-based integer	7.950	3.015

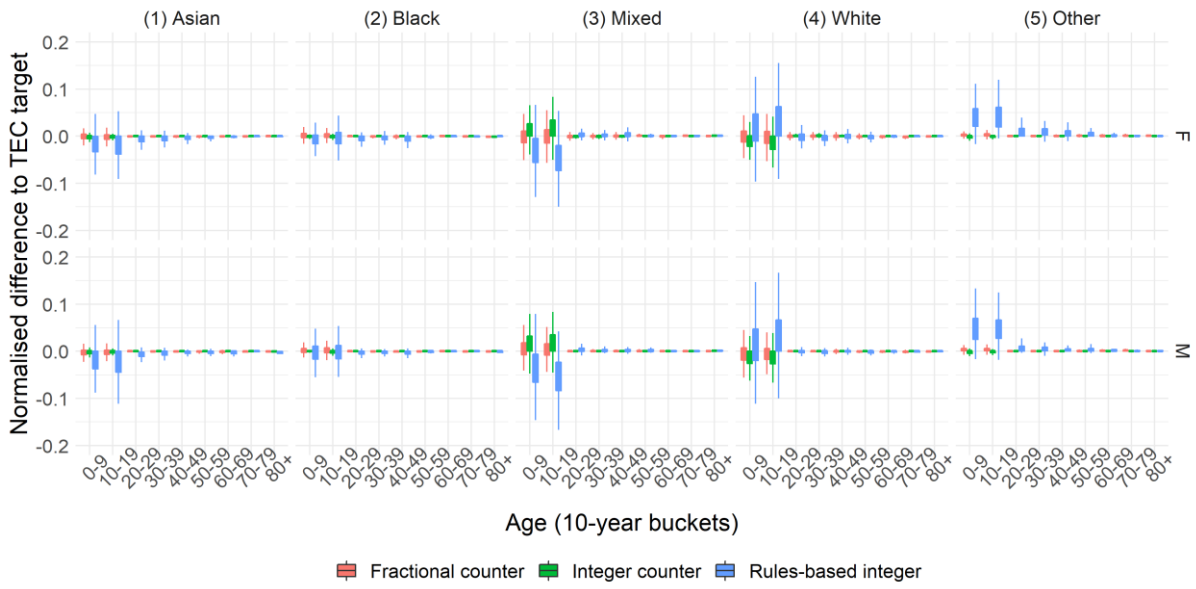

Figure 30 Box and whisker plot showing the distribution of normalised differences to the true ethnicity and conflicts (TEC) benchmark from the fractional (red) and integer (green) counters and the rules-based integer (blue) approach across 10-year age buckets for males (top panels) and females (bottom panels) and ethnicity. Outliers have been removed for clarity. Similar numbers of outliers found in each category, with the outliers for (4) White M/F being significantly larger than the other categories.

Summary of standardised ethnicity analysis

At the highest aggregate level (e.g. 5-category ethnicity) the fractional and integer counters perform similarly on the target population (TP) subset, returning similar metrics and producing similar national ethnicity distributions, with the rules-based integer approach appearing to perform better when considering the metrics and national ethnicity distributions. The rules-based integer approach continues to outperform the fractional and integer counters for the TP subset for ethnicity*sex, ethnicity*sex*age and ethnicity*LA, however, at the LA*ethnicity*age*sex level, the fractional counter performs the best, followed by the integer counter, with the rules-based integer approach the worst. Sub-setting the population to those with a true Census matched ethnicity (TE) and those with conflicts (TEC), the fractional counter performs the best, followed by the integer counter, with the rules-based integer approach performing the worst. This is the case for all aggregations of these subset populations, with the rules-based integer approach significantly worse than both counters. The difference in performance between the modelled approaches (fractional/integer) and the rules-based approach is far greater than the difference between the models for these subset populations, suggesting that either modelled approach would be preferable to the rules-based at these levels.

Holdout performance for placement and ethnicity models

We calculated record-level performance on the holdout set of TAC and TEC subsets described in the methods section to assess whether the model-based methods (fractional and integer) outperform the rules-based methods when the data used to

train the models is excluded. This is common practice when assessing machine learning models, but in our main results we leave the training data in our test data. This was done to to allow comparison with other methods that use the entire dataset, and justified on the basis that the training data is a small fraction of the overall data and its inclusion does not alter our conclusions. The data for the results that exclude the training data are not presented here, but we include these record-level performance metrics from smaller holdout test sets (that contain no training data).

The S2 placement models more comprehensively and accurately classify correct locations when there are conflicting values than the rules-based method (Table 33), as do the S3 ethnicity models (Table 34).

Table 33. Performance metrics on TAC subset of holdout set. Most performance counter in each metric highlighted. Metrics not applicable to specific counters are shaded in black.

Counter	Precision	Recall	Log loss
Fractional			0.61
Integer	0.62	0.71	
Rules	0.45	0.5	17.12

Table 34. Performance metrics on TEC subset of holdout set. Most performance counter in each metric highlighted. Metrics not applicable to specific counters are shaded in black.

Counter	Precision	Recall	Log loss
Fractional			0.49
Integer	0.73	0.81	
Rules	0.28	0.31	23.1