Office for National Statistics

GSS Methodology Series No 42

Assessing the Generalised Structure Preserving Estimator (GSPREE) for Local Authority Population Estimates by Ethnic Group in England

Solange Correa-Onel, Alison Whitworth and Kirsten Piller

November 2016

Abstract

The two way table defined by ethnic group and local authority is typical of many census outputs. The methods used in this paper focus on combining census, administrative and survey data and are of wider relevance in developing a statistical framework for estimating population attributes for specific geographies in non-census years. The methods are flexible in that additional sources can be introduced as available and existing sources excluded if they are no longer relevant.

Generalised Structure Preserving Estimation (GSPREE) has been used to produce population estimates for a categorical population characteristic; broad ethnic group (White, Mixed, Asian, Black, Chinese and Other) by local authority in England. The GSPREE approach combines recent survey estimates with more detailed, but outdated, census distributions and also recent information available in administrative sources, for a subsection (but not all) the population.

This paper has three main objectives: i) to demonstrate the use of structural preserving estimation for estimating population by local authority and broad ethnic group for June 2014; ii) to assess the performance of the GSPREE estimators in a validation scenario where the true population distribution is known (i.e. March 2011 Census); iii) to outline further work that may enhance the benefits of incorporating additional administrative data in the estimation process. Therefore, for this paper, GSPREE was applied to produce estimates for both June 2014 and March 2011 under three alternative modelling strategies. These strategies had increasing levels of complexity in terms of estimation approach and use of additional auxiliary sources.

The results for both time periods show that a more complex GSPREE model, accounting for age groups and combining different sources of detailed auxiliary information (census and English School Census), is successful in improving the GSPREE estimates. For the validation scenario, the results show that GSPREE is fairly successful in capturing the population distribution by ethnic group and LA in the 2001-2011 period. Initial assessment of the potential error in the GSPREE estimates (obtained via bootstrap) suggest negligible bias and square root MSE in most areas under the modelling strategies, so further research is required to investigate those showing larger error and to improve measures of accuracy.

Table of Contents

1. Introduction

1.1. Motivation

In March 2014, the National Statistician made a recommendation that the census in 2021 should be predominantly online, making increased use of administrative data and surveys to enhance the statistics from the 2021 Census. This recommendation was endorsed by the Government's formal response, which also set out its ambition that "censuses after 2021 be conducted using other sources of data… sufficiently validating the perceived feasibility of that approach".

It is ONS's ambition to produce the type of information that is collected by a ten-yearly census from an Administrative Data Census. This will require combining administrative and survey data to produce information on population and household characteristics that are currently provided in the census. The goal is to be able to compare outputs based on administrative data and targeted surveys against the 2021 Census to demonstrate to government and other users that the alternative can produce high quality information at a lower cost, and can do so more regularly.

Small area estimation (SAE) provides a tested and transparent mechanism for integrating sources, and thus has potential for expanding population statistics or estimates contributing to an Administrative Data Census. Most social surveys are designed to provide reliable direct estimates at national or regional levels but it is not usually economically viable to obtain sample counts that are large enough to provide robust direct estimates for small population domains. Small area estimation methods work by drawing strength from information across different data sources (including administrative data) and across similar population subgroups in order to obtain robust model-based survey estimates where sample counts are too small for direct estimates. They provide a powerful mechanism for bringing information together across sources and estimating from integrated data.

The census produces three key types of information (the size of the population, households and families and population and housing characteristics). This project applies a structural small area estimation approach for producing outputs on population characteristics. GSPREE has been used to combine different sources for population estimates by broad ethnic group and local authority between census

years. Population estimates by detailed ethnic group and local authority were produced (and published) annually from 2006 until 2009 using the cohort component method (Office for National Statistics, 2005), which updates the census estimates accounting for components of population change (births, deaths and migration). However, the annual estimates were withdrawn in the latter part of the last decade as it was thought that the methods employed had not captured changes in the distributions of population by ethnic group (particularly in some London Boroughs) which had resulted from recent migration patterns.

In this approach survey and administrative data are used to update the census table in a systematic way, drawing upon their particular strengths. The two-way table of ethnic groups by local authorities (LAs) is typical of many census outputs and so the methods studied here are of wider relevance in estimating population attributes from combined sources. These typical two-way tables can be reliably estimated in census years, but currently not many are produced between years due to difficulties in capturing change in population characteristics such as ethnicity.

The methods are flexible in that additional sources can be incorporated as available and existing sources excluded if they become less useful. The models provide parameters which demonstrate the importance of the different sources. As well as deriving estimates between census years (i.e. updating census estimates) the structural approach used, has potential for deriving estimates of population characteristics from survey and administrative sources only (i.e. for the Administrative Data Census). The main purpose of this application is to provide a worked example of these types of methods and to demonstrate the strengths and weakness when applied for a specific population output. A wider framework for estimating population characteristics within the context of the Administrative Data Census will be published by ONS in 2017. This will describe how different data sources can be brought together and used to produce statistics on population characteristics in a systematic way (i.e. taking into account the differences in the nature of the characteristics being estimated and in the information that is available to capture them).

The methods under investigation in this report are the Structure Preserving Estimation (SPREE) method (Purcell and Kish, 1980) and, its extension, the Generalized SPREE (GSPREE; Zhang and Chambers, 2004). These procedures combine auxiliary information (e.g. data from previous population census and/or administrative data) with current survey data to improve the quality of estimates for cells in a contingency table. Measures of accuracy are obtained via resampling methods (e.g. bootstrap), which involves resampling a large number of times from an artificial population.

The rationale of structure preserving estimators is that detailed relationships between cross-classified variables available at a previous time point (usually in a census) provide a good structure for estimation at the current time period, while estimates from a current survey provide reliable and up-to-date total estimates at the aggregate level (or margins). The idea is, therefore, to adjust the cross-classified cells of the contingency table at the current time period preserving the detailed structure from the auxiliary source (census) and the marginal totals from the current survey (Purcell and Kish, 1980).

An important difference between the simple SPREE and the GSPREE is that the former assumes that the distribution of cross-classified variables in a previous time point (e.g. census) is the same as at the current time period (e.g. survey), whereas the latter allows a more flexible structure for this relationship by assuming that the distributions at the two time points are proportional. When the proportionality constant is equal to one, the GSPREE is equal to the SPREE (i.e. the SPREE is a particular case of the GSPREE). In addition, the GSPREE allows for more than one auxiliary source (e.g. census and School Census) to be incorporated in the estimation process. Direct estimates of the cell totals (obtained from surveys) can also be incorporated in the estimation in order to improve the quality of the estimates.

1.2. Background

In this paper a practical application of Generalised Structure Preserving Estimation (GSPREE) method is provided for the categorical population characteristic: ethnic group classified in categories of White, Mixed, Asian, Black, Chinese and Other.

The choice of the broad ethnic groups is mainly driven by the categories used in the census and are harmonised to ensure consistency with those used across surveys and other administrative sources. Broad age groups are used for this initial investigation as estimates at more aggregate levels are generally more robust, and use of the methods for more detailed tables will be considered at a later stage following proof of concept. Chinese is included as a separate category in this study to ensure that methods are assessed for characteristics that are rare in the population.

This application follows an earlier collaborative initiative between the ONS and the University of Southampton to investigate the potential use of small area estimation methods to produce local authority population estimates by ethnic group, making best possible use of *aggregate* information from census, survey and administrative data (Luna-Hernandez, Zhang, Whitworth and Piller, 2015).

In that application, the authors used aggregated data based on the 2011 Census, the 2013 English School Census (ESC), the 2012 and 2013 official mid-year population estimates (MYEs) and the 2012-2013 Annual Population Survey (APS). The reference date for the preliminary estimates produced in that paper was 31st December 2012. Results showed that, overall, the structure preserving estimators were efficient in producing reliable estimates for most LAs. Implementation of the method separately by age group and accounting for the School Census information improved the estimates.

This work addresses some topics raised as potential future work in the previous application. Firstly, the aggregated 2011 Census data are combined with more recent aggregated auxiliary data sources: the 2014 English School Census with January 2014 as reference period. The margins totals are also updated in this application, using the 2014 APS and 2014 MYEs with June 2014 as reference period. This increases the time period between 2011 Census and the margins totals reference dates as an attempt to better capture the dynamic of population changes in terms of ethnicity. In addition, the reference date for the GSPREE estimates is June 2014, which coincides with the MYEs reference date and makes the results more easily comparable. Figure 1 provides a visual representation of the data sources and helps to demonstrate their contribution to the estimates.

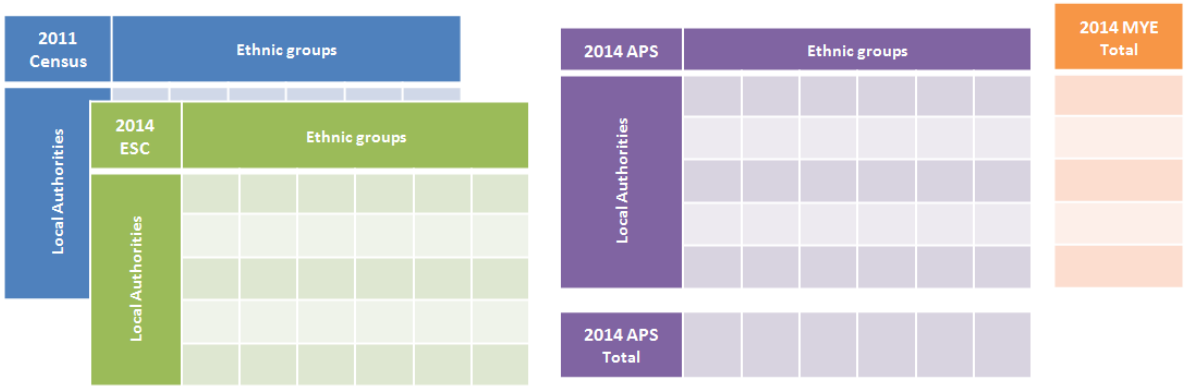


Figure 1. Visual representation of the data sources

This paper also includes a validation study comparing the distribution of the population by LAs and ethnic groups in the 2011 Census to the 2011 GSPREE estimates. The latter are obtained using the 2001 Census as proxy table and 2011 APS, 2013 School Census and 2011 Census as auxiliary information.

This paper, therefore, has three main objectives: i) to demonstrate the use of structural preserving estimation for estimating population by local authority and broad ethnic group for June 2014; ii) to assess the performance of the GSPREE estimators in a validation scenario where the true population distribution is known; iii) to outline further work that may enhance the benefits of incorporating additional administrative data in the estimation process. It is structured as follows: Section 2 presents the GSPREE methodology and Section 3 addresses the application for June 2014, describing the data sources considered, modelling strategies and discussion of results. Section 4 describes the validation study comparing 2011 GSPREE estimates and 2011 Census, with some promising preliminary results. Section 5 presents conclusions and future work.

2. Methods

2.1. Overview

Research in SAE has gained relevance in the recent decades due to an increasing demand for outputs at small area level or for detailed population subgroups (i.e. small domains). A comprehensive account of SAE methods can be found in Rao and Molina (2015). A review of the most important developments of the last decade is presented in Pfeffermann (2013).

The idea behind the use of SPREE for this application is that census data provide detailed distribution of the population by LAs and ethnic group (i.e. the cross-classification structure or otherwise known as the association structure), whilst the more recent APS and the MYEs provide the updated column and row totals, respectively, for population by ethnic group at country level (i.e. benchmark totals, otherwise described as the allocation structure). The APS provides also some recent information about the cross-classification structure but this may not be reliable due to small sample counts at this level of disaggregation. The ESC is a second source of auxiliary information explored, providing updated and detailed information of the population by LA and ethnic group for a subset of the population: those between the ages of 5 to 15 years.

The census data are considered a *proxy* for (rather than a correlate of) the population quantities of interest and are used to obtain updated estimates. *Proxy* data are usually distributions of the variable of interest obtained from census or administrative sources for a given set of areas in a different period of time or under non-equivalent definitions (Green et al., 1998). The structure of the cross-classification is incorporated in the modelling process via parameterisation using the logarithm of the auxiliary tabulation (census data) and an interaction term between the columns and the rows of the two-way table. The interaction term is then updated to the available column and row margins (also referred to as benchmark totals) of the target table using iterative proportional fitting (Agresti, 2013, p. 365-366). The SPREE (Purcell and Kish, 1980) and GSPREE (Zhang and Chambers, 2004) methods are formalised in the next Section.

2.2. The Structure Preserving Estimation (SPREE) Method and Its Extensions

Denote by Y the population table of interest with cells Y_{aj}, where $a = 1, \dots, A$ indexes the set of areas and $j = 1, \dots, J$ indexes the categories of the variable. Define $\varsigma_{aj}^{Y} = log\, Y_{aj}$. Thus, Y can be represented in the form of a saturated log-linear model as:

$$\varsigma_{aj}^{Y} = \alpha_0^Y + \alpha_a^Y + \alpha_j^Y + \alpha_{aj}^Y \tag{1}$$

where $\alpha_0^Y = \overline{\varsigma_{..}^{Y}}$, $\alpha_a^Y = \overline{\varsigma_{a.}^{Y}} - \alpha_0^Y$, $\alpha_j^Y = \overline{\varsigma_{.J}^{Y}} - \alpha_0^Y$ and $\alpha_{aj}^Y = \varsigma_{aj}^Y - \alpha_0^Y - \alpha_a^Y - \alpha_j^Y$, for $a = 1, \dots, A$ and $j = 1, \dots, J$. The dot indicates sum over the corresponding subscript.

Following Purcell and Kish (1980), equation (1) can be used to decompose Y in two parts: the *association* structure and the *allocation* structure. The former, corresponds to the terms $\{\alpha_{aj}^{Y}\}$, also called *interactions*, and determines the relationship between rows and columns in the table. In the theoretical case where rows and columns are independent, all the interaction terms are zero. The *allocation* structure, given by the terms α_0^Y, $\{\alpha_a^Y\}$ and $\{\alpha_j^Y\}$, carries information about the scale of the table and the disparities within the sets of rows and columns and is implicitly determined by the row and column margins of the table.

Notice that in the SAE setting, it is easier to obtain information related to the *allocation* structure than to the *association* structure. Even if Y remains unknown, accurate estimates of the row marginal, i.e. the area sizes, can be obtained either from administrative sources or from population estimates. Similarly, given that the column marginal corresponds to the aggregation over the entire set of areas, if not available from other sources, it can usually be estimated using survey data.

Given the margins of Y, (i.e. its allocation structure), a *proxy* of the table of interest, denoted by X, can be used to estimate the association structure of Y. A *proxy* table is, therefore, supposed to contain information for the same set of areas and regarding a similar characteristic as the table of interest. In particular, it is assumed to have the same $A \times J$ dimension. Notice that for demographic characteristics during intercensal periods the corresponding tables from the census year are

obvious available proxies. More generally, proxies are usually derived not only from censuses but also from administrative sources.

For a two-way table, the SPREE of Purcell and Kish (1980) simply uses the observed association structure in the *proxy* table as an estimate for that target table. In other words, denoting by $\{\alpha_{aj}^{X}\}$ the interaction terms for the *proxy* table X that can be defined in analogous way to equation (1), the SPREE is characterised by the *structural equation*: $\alpha_{aj}^{Y} = \alpha_{aj}^{X}$, for $a = 1, \dots, A$ and $j = 1, \dots, J$.

The procedure proposed by Purcell and Kish (1980) to obtain the SPREE of Y is straightforward. The known margins of Y can be imposed on X using a multiplicative raking procedure such as the Iterative Proportional Fitting (IPF) algorithm (see for instance Agresti, 2013, p. 365-366), ensuring that the association structure of X remains unaltered. Equivalently, the known margins can be imposed by fitting a saturated log-linear model with an offset term given by the terms $\{\alpha_{aj}^{X}\}$ (Noble et al., 2002).

However, assuming that the *proxy* and the target tables share exactly the same association structure is clearly restrictive in practice. Other estimators have been proposed to *preserve* in a more flexible way the association structure, leading to what is called the SPREE approach. The modifications to the original SPREE of Purcell and Kish (1980) go in two main directions: i) by relaxing the structural equation of SPREE to consider other types of relationship between the two association structures and ii) by including cell-specific random effects to allow for extra heterogeneity unexplained by the structural equation, which contributes to reduce the potential bias of the synthetic estimator in the presence of departures from the structural equation.

Besides the SPREE, the following estimators can be framed within this approach: the Generalized Structure Preserving Estimator (GSPREE) (Zhang and Chambers, 2004), the Extended Structure Preserving Estimator (ESPREE) (Cinco, 2010) and the nonlinear estimator proposed by Berg and Fuller (2014).

In all estimators mentioned above, the allocation structure is imposed on the final estimates via benchmark to a set of known margins, providing additional protection against misspecification of the assumed models (Pfeffermann, 2013).

The small area problem may still persist even when a survey estimate of the target table Y is available, as the direct estimates of the cell totals are usually too unstable due to the small sample sizes. Such information can be used, however, to *update* the association structure of the *proxy* table and, hence, to reduce the bias of the SPREE. That is the underpinning idea behind the GSPREE proposed by Zhang and Chambers (2004).

The GSPREE is characterised by the structural equation $\alpha_{aj}^{Y} = \beta\alpha_{aj}^{X}$ for $a = 1, \dots, A$ and $j = 1, \dots, J$. Clearly, when $\beta = 1$ the GSPREE corresponds to the SPREE. The authors propose two ways of estimating β: a model-assisted approach based on a Generalised Linear Structural Model (GLSM), and a fully model-based approach assuming Multinomial or Poisson distribution for the sample cell counts in each area. The latter is described here and used in the application presented in Section 3 (as in the first interim report) and maximum likelihood estimators (MLE) of β are then obtained. Once $\hat{\beta}$ is obtained, the GSPREE of the target table Y is calculated by imposing the known row and column margins to the table of estimated exponentiated interactions with cells $\tilde{Y}_{aj} = e^{\hat{\beta}\alpha_{aj}^{X}}$, via iterative proportional fitting.

An estimation procedure for β built directly from the structural equation $\alpha_{aj}^{Y} = \beta\alpha_{aj}^{X}$ involves several problems. Small sample sizes can lead to sample counts of zero for some of the cells, in which case the interaction terms for the survey estimate of Y are not defined. Moreover, even if all cells have a positive estimate, there is not a *natural* distribution that can be assumed for the interactions – as there is for the proportions or the counts – making it difficult to justify a standard approach such as Maximum Likelihood, for instance. Therefore, instead of formulating a model in the interaction scale, Zhang and Chambers (2004) propose to estimate β using the Generalized Linear Structural Model, which relates the within-area proportions of the proxy table and the table of interest on the log scale centred on the average of the

area (see Appendix A for further description of the Generalised Linear Structural Model).

The GLSM is fitted via Iteratively Weighted Least Squares (IWLS) using direct estimates of the within-area proportions and estimates of their variances. In the absence of estimates of the variance of the direct estimators, it is possible to obtain fully model-based estimates of β. One option, suggested in Zhang and Chambers (2004) is to assume a multinomial distribution for the sampling cell counts in each area, and obtain an estimator of β using Maximum Likelihood (ML). Notice that this approach implicitly assumes that the sampling design of the survey is ignorable for Y. Otherwise, direct estimates of the proportions can be used instead of the observed proportions.

Fully model-based estimates of β under the GSPREE structural assumption can also be obtained assuming a Poisson distribution for the sampling counts y_{aj}. The equation:

$$\log Y_{aj} = \gamma_a + \lambda_j + \beta\alpha_{aj}^X \tag{2}$$

with $\sum_j \lambda_j = 0$ is equivalent to the structural equation of the GSPREE. Both the γ_a and the λ_j terms for $a = 1, \ldots, A$ and $j = 1, \ldots, J$, are nuisance parameters. It is possible to fit (2) in a standard software using log-linear models and obtain the corresponding ML estimator of β. By doing so, it is implicitly assumed that the structural equation of the GSPREE holds for the table of direct estimates as well, or at least, that the value of β that better relates the table of interest and the proxy table does not change when the former is substituted by its direct estimate.

The value of $\hat{\beta}$ is then used to obtain the GSPREE of the target table Y by calculating the table of estimated exponentiated interactions $\tilde{Y}_{aj} = e^{\hat{\beta}\alpha_{aj}^X}$. The known row and column margins (or benchmark totals) are then imposed on this table using IPF.

In the application presented in Section 3 we follow the fully model-based approach using Equation (2) and a Poisson distribution for the cell counts in order to simplify

the fitting process. By doing so, the estimation process can be subject to misspecification of the variance structure of the sampling errors. Nevertheless, using an argument similar to that for the generalised estimating equation approach in Liang and Zeger (1986), it is possible to show that in such a case the estimator of β, although not fully efficient, would remain unbiased.

For more information on methods and applications, refer to Zhang and Chambers (2004), Luna-Hernandez (2014) and Luna-Hernandez, Zhang, Whitworth and Piller (2015).

3. Application: 2014 Population Estimates by Local Authority and Ethnic Group in England

In this Section, the fully model-based GSPREE approach is applied to different scenarios in order to assess the best way to make use of aggregate information available from several data sources.

As mentioned in previous sections, the GSPREE aims at allowing more flexibility by relaxing the assumption concerning the longevity of the census distribution and by allowing new information from administrative sources to be incorporated in the estimation process as it becomes available. The GSPREE estimates are ultimately benchmarked to the LA mid-year population estimates and to the ethnic group population estimates at the country (England) level.

In this application, aggregate data from the 2011 Census and 2014 English School Census[1] are used as *proxy* tables. Survey data for the target table Y to be used in the GSPREE approach are obtained from the 2014 APS (January to December). Population estimates by LA (row margin) are obtained by the 2014 official MYEs with reference period of June 2014. The population estimates by ethnic group (column margin) are obtained from the 2014 APS with reference date equal to the period mid-point, June 2014. Thus, the GSPREE estimates have June 2014 as reference period.

The ethnic groups considered in this exercise are: White, Mixed/Multiple Ethnic Groups, Asian/Asian British, Black/African/Caribbean/Black British, Chinese and Other. These categories are fully harmonised with the census, APS and ESC data sources. As mentioned in Section 1.2, ethnic groups considered here are mainly the broad classification used in the census, with a separate category for Chinese to allow the assessment of methods for sparse categories. Also, focusing the analysis on six key broad ethnic groups rather than a more detailed classification, allows the methods to be tested for proof of concept in this preliminary assessment of the GSPREE. Application of the methods for more detailed categories will be undertaken

[1] Access to and use of information from the English School Census is authorised by data sharing regulations i.e. Statistics and Registration Service Act 2007 (Disclosure of Pupil Information) (England) Regulations 2009.

at a later stage of the project. The analysis is conducted for 324 LAs in England, as City of London and Isles of Scilly local authorities are excluded.

More information on the data sources used in this application is presented in Section 3.1, along with some initial exploratory analysis. Section 3.2 presents a brief description of the modelling approach adopted when applying the GSPREE method. A comparison of GSPREE point estimates based on different scenarios is presented in Section 3.3. Finally, the GSPREE estimators are assessed in terms of bias and mean square errors (MSE) via bootstrap in Section 3.4.

3.1. Data Sources

In order to produce population estimates for the target table (LA by ethnic group) using GSPREE, *proxy* and survey estimates for the cross-classified table of interest and estimates of the corresponding row and column margins (i.e. benchmark totals or allocation structure) are required. Further details of each of those components are presented below.

Proxy Information

Proxy information for the distribution of Ethnicity at the LA level in England is obtained from the 2011 Census and the 2014 ESC for the population attending school.

The 2011 Census provides estimates of the counts of people and households who are defined as usual residents of England and Wales on the 27th March. The estimated coverage rate for people in the 2011 Census was 93%. The observed counts were adjusted by overcount and undercount, taking into account the characteristics of individuals and households who were missed from the census enumeration process.

The 2014 ESC has almost full coverage of children between the compulsory school ages of 5 and 15 in State-maintained schools and non-maintained special schools. Independent schools and home educated children are not covered though and this can result in some differences between the population estimates for children in compulsory school age obtained from the School Census and other sources. In 2014, 7% of pupils (all ages) in England went to an independent school. The ESC

has collection periods in January (when information on pupil's ethnicity is obtained), in May and in October of a specific year.

As the ESC only provides good coverage for children between 5 and 15 years old, it can be said that, for this empirical exercise, there is one source of *proxy* information for individuals in the age groups 0-4 and 16 and above (the 2011 Census); and two sources for those between 5 and 15 years (the 2011 Census and the 2014 School Census). In order to allow different modelling approaches for each age group, age-group specific *proxy* tables of LA by ethnic group are also produced. The English School Census is only considered in the model in combination with the age-group specific census table.

Survey Estimates

The APS is a household survey that is designed to provide information at a local level on many demographic and socioeconomic topics. It produces updated population estimates by ethnic group, as it contains detailed information on ethnicity and it includes information for all LAs in England (except for Isle of Scilly). It has also the largest sample size among the periodic surveys available. However, the APS sample counts are too small (or null) to provide reliable estimates for all ethnic groups at LA level.

The APS data are released quarterly (January to December; April to March; July to June; and October to September) and contain approximately 250,000 individuals. It contains the Labour Force Survey (LFS) data and the boost samples to the LFS. The boost sample for England is called the English Local LFS (ELLFS) and has been designed to give a minimum sample size of economically active individuals for each local education authority. The APS data for England therefore consist of four successive quarters from the LFS, plus the ELLFS boost.

Both the LFS and the ELLFS use a rotational sampling design involving waves. For the LFS, a sample of households is interviewed quarterly for five waves, inducing an 80% of overlap between samples of consecutive quarters. For the ELLFS a sample is interviewed once a year for four waves.

The households are included in the APS only the first or fifth time that they are interviewed, so that each respondent appears in the dataset only once. Non-private

households (some communal establishments, armed forces accommodation, etc.) are excluded from the sampling frame. For England, the households are sampled through the Royal Mail Postcode Address File (PAF) and the National Health Service (NHS) communal accommodation list.

The APS data corresponding to January – December 2014 are used in this study. The reference point is taken as the period mid-point, so approximately 30th June 2014. In analogy with the procedure applied to the 2011 Census, age-group specific survey tables of LA by ethnic group were produced for each of the following age groups: 0-4, 5-15 and 16 and over.

Benchmark Totals (or Allocation Structure)

Updated totals for the population by LA, and by ethnic group at country level (i.e. the row and column totals) are used within the GSPREE method to update the cross classified cells of the contingency table (see Section 2). Estimates of the LA populations are obtained from the official MYEs. These estimates are produced using the cohort component method, which uses information on components of population change to update the most recent census population. In this method, the previous year's population estimates by sex, age and local authority of usual residence is aged on by one year. Births within the 12 months to the reference date are added to the population and deaths are removed. The net flows of migration are accounted for, including internal (cross-border and between LAs) and international flows. There are also adjustments for special populations (armed forces and prisoners) who are not represented in the data sources used for the components of population change.

The 2014 MYEs at LA level are used to calculate the row margin. The reference date of such estimates is 30th June of the corresponding year, which is consistent with the reference period of the other sources involved in this analysis.

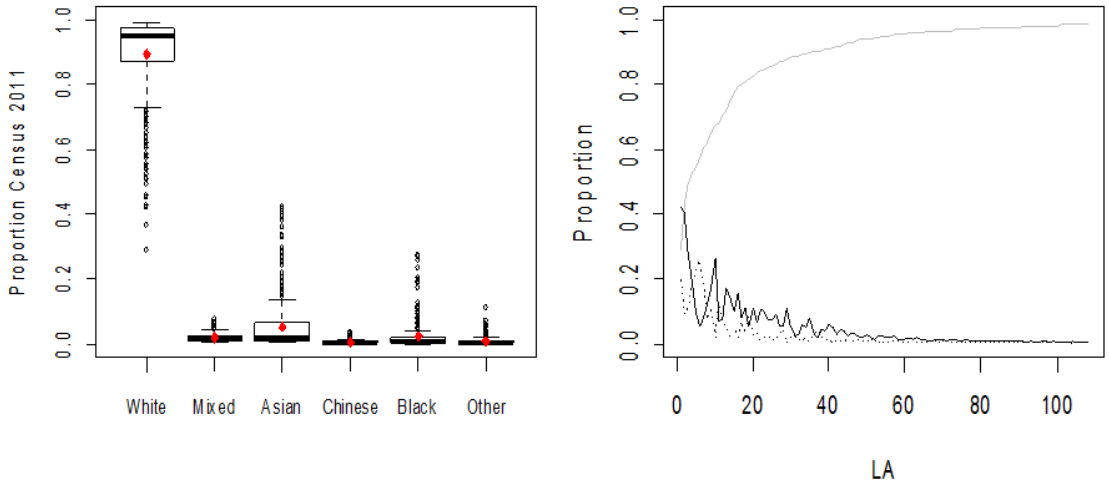
Direct estimates of the total population by ethnic group obtained from the APS at the country level (England) are used as the column benchmark totals in this study. These estimates differ slightly from the MYE as the survey weights are based on an APS-defined population, which does not cover non-private households.

Descriptive Analysis

This analysis demonstrates how the distribution of the population by ethnic groups and LA compares across the data sources described above. It shows the uneven distribution of the population across ethnic groups and the difference between the cross classified two way table (i.e. ethnicity by LA) in the two proxy sources (i.e. 2011 Census and 2014 School Census).

According to the 2011 Census, the distribution of the population by broad ethnic groups in England (excluding City of London and Isles of Scilly local authorities) shows that the category White makes up 85.42% of the population, followed by Asian with 7.10%, Black with 3.48%, Mixed with 2.25%, Other with 1.03% and, finally, Chinese with 0.72%. The total population is 53,002,878.

Figure 2 shows the distribution of LAs population by ethnic groups in England. It can be seen that LAs differ considerably in terms of proportions of White, Asian and Black, with several LAs showing proportions of Asian and Black much higher than the country proportion.



(a) (b)

Figure 2. Distribution of local authority population by ethnic group in the 2011 Census.
(a) The box plot shows the distribution of local authority population by ethnic group. The diamond is the national proportion.
(b) Shows the proportions of White (solid grey line), Asian (dotted black line) and Black (solid black line) ethnic groups across local authority. One in each three LAs are shown, in ascending order of proportion of White.

Figure 3 shows the distribution of within LA population proportions in the 2011 Census and in the 2014 ESC for the 5-15 age group. It can be seen that, on average, the proportions of Black and Other in the 2014 ESC tend to be higher when compared to 2011 Census, whereas the proportion of Chinese tend to be lower. It is noteworthy that the corresponding plot for 2011 Census versus 2011 ESC (not shown) indicates almost identical patterns to those observed in Figure 3 in all ethnic groups. The differences shown therefore are most likely due to differences in the coverage of the data sources rather than a change in the distribution of population by ethnic group between 2011 and 2014.

Figure 4 compares the association structure $\{\alpha_{aj}^{X}\}$ obtained from both tables (2011 Census and 2014 ESC) for each ethnic group. The two association structures are not identical (particularly for Mixed, Chinese and Other) and combining the two proxy tables can potentially improve the GSPREE estimates in terms of precision. This is considered in Section 3.2 as an alternative modelling strategy (see Model 3).

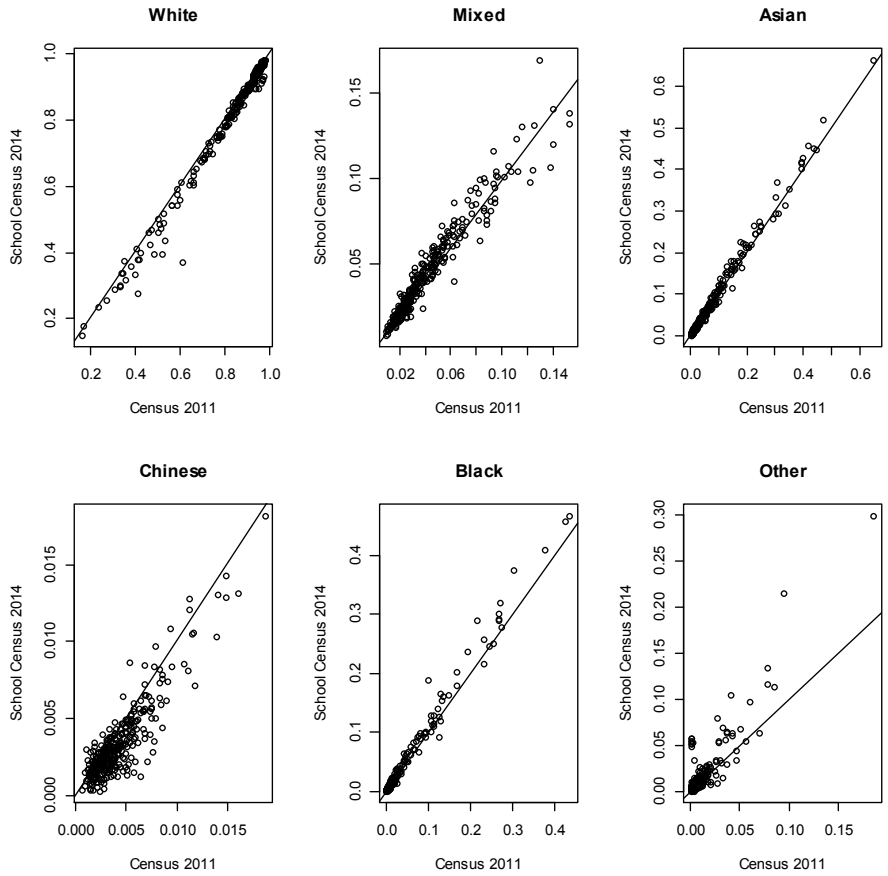


Figure 3. Comparison between 2011 Census (5-15 age group) and 2014 English School Census in terms of population proportions by local authority and ethnic group. Line: Y=X.

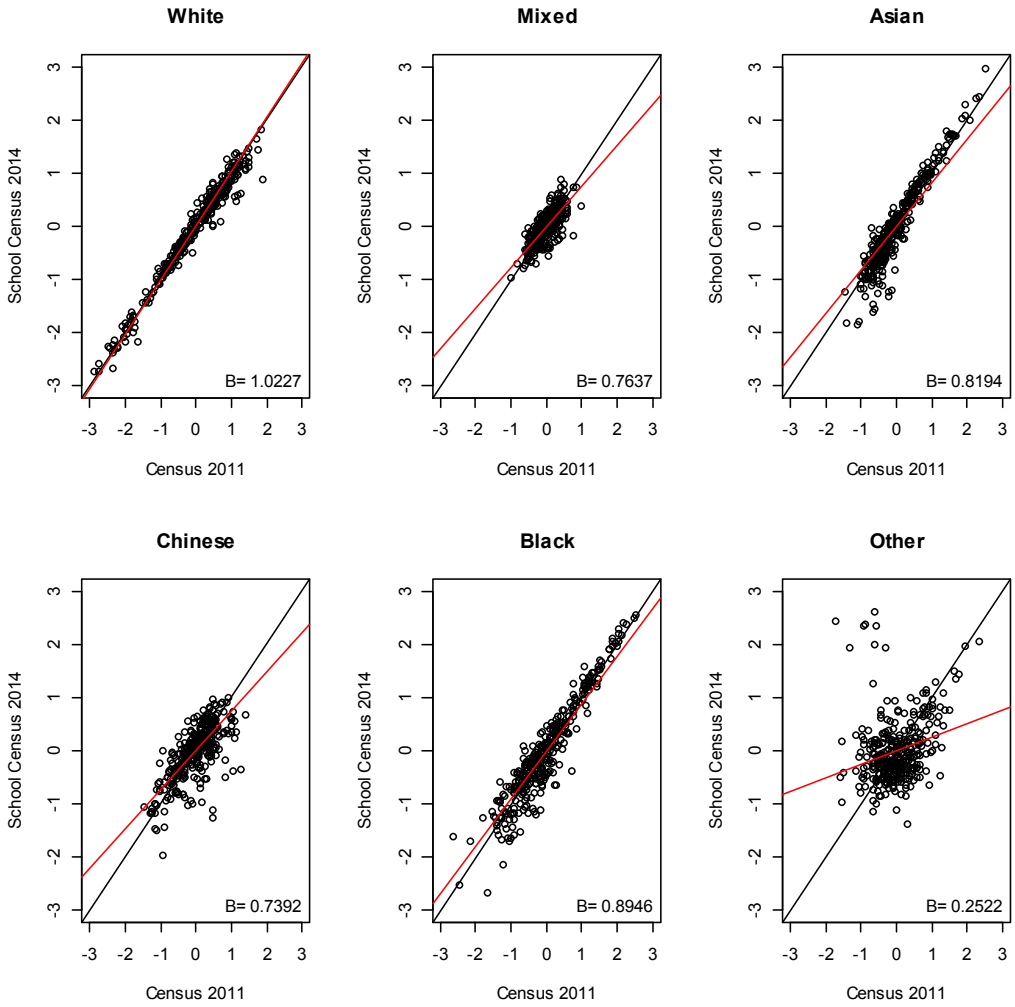


Figure 4. Comparison between 2011 Census (5-15 age group) and 2014 English School Census in terms of association structure by local authority and ethnic group. Line: Y=X.

The actual sampling fractions of the 2014 APS in some LAs can be very small. An implicit sampling fraction was obtained dividing the observed sample size by the corresponding projected population total in each LA. This varies between 0.2% and 2.3%, with an average of 0.4%. In addition, preliminary analyses based on the 2014 APS show that this LA versus ethnic group table shows that 46% of the LAs have at least one cell with zero count. No LA has cells with zero count in the 2011 Census (excluding Isle of Scilly).

The low proportions of individuals in the population belonging to categories as Chinese (0.72%) or Mixed (2.25%) and the relatively small APS sample sizes observed in most LAs contribute to the large number of zero cell counts in the target table.

Table 1 shows the population percentages by ethnic groups in England based on 2011 and 2014 APS (2014 APS is used as the column total in this application) and on the 2011 Census. It shows that the distribution of the population by ethnic groups has been reasonably stable from 2011 to 2014. These figures should be used with caution however as comparisons using the APS estimates should account for the corresponding standard errors.

Table 1. Population percentages by ethnic group according to 2011 Census, 2011 APS and 2014 APS. England.

	Ethnic Group					
	White	**Mixed**	**Asian**	**Chinese**	**Black**	**Other**
2011 Census	85.42	2.25	7.10	0.72	3.48	1.03
2011 APS (Oct-Sep: weighted)	86.44	1.44	6.55	0.50	3.33	1.73
2014 APS (Jan-Dec: weighted)	85.43	1.67	7.17	0.54	3.35	1.85

3.2. Modelling Approaches

As described in the previous section, two different sources of *proxy* information are available for the age group 5-15: 2011 Census and 2014 ESC. To identify the best way to incorporate the aggregate *proxy* information in the structure preserving methodology, three *fixed effects* models with increasing level of complexity are considered. All models are built on the expression defined in equation (2).

The first approach (Model 1) considers the 2011 Census as the only source of *proxy* information and both the census and the APS survey tables are produced for LA versus ethnic group. In the second approach (Model 2), the age group disaggregation (0-4, 5-15 and 16 and above) is accounted for. The *proxy* and the survey tables are produced for LA versus ethnic group with independent fitting in each age group. Model 3 is similar to Model 2, but it combines both the 2011 Census and the 2014 ESC to obtain a third *combined proxy* table. In Model 3, the weight δ given to the census *proxy* table is estimated separately for each age group via numerical optimisation and it ranges from zero to one. Note that the ESC association structure for the 5-15 age group is used for the 0-4 and 16 and above age groups when creating the combined proxy table. The three modelling approaches are summarised in Table 2. More details in Office for National Statistics, Luna-Hernandez and Zhang (2015).

Table 2. Modelling approaches. Fixed effects Models.

Fixed Effects GSPREE Approaches	Proxy Information	Disaggregation - Proxy and Survey Tables	Age Group
Model 1	2011 Census	LA x Ethnicity	None
Model 2	2011 Census	LA x Ethnicity | Age Group	0-4 5-15 16 and above
Model 3	2011 Census and 2014 School Census	LA x Ethnicity | Age Group	0-4 5-15 16 and above

3.3. Results

In this section, the GSPREE point estimates are obtained under the three scenarios shown in Table 2. For all models, estimation of the association structure (β) between *proxy* and survey tables is obtained via Poisson Maximum Likelihood Estimation based on Equation (2), as mentioned in section 2.2. Results are presented in Table 3. For all modelling strategies, the estimated proportionality constant $\hat{\beta}$, is very close to 1. This indicates that, for the two-way table LA versus ethnic group, the association structure based on the proxy table (2011 Census for Models 1 and 2; and combination of 2011 Census and 2014 ESC for Model 3) requires small or no adjustment in order to derive the association structure for the reference time period (June 2014). A potential reason for this is the small time difference between APS and proxy tables, as $\hat{\beta}$ is determined from the relationship between the 2014 APS detailed cross-classification and the proxy tables. Another important point is that the GSPREE captures the change across all ethnic groups simultaneously by estimating only one proportionality constant, $\hat{\beta}$. A more efficient method would provide different $\hat{\beta}$ values for different ethnic groups. This is discussed further in section 5 as future work.

In Model 3, the estimated weights given to the 2011 Census proxy table, $\hat{\delta}$, are 0.74, 0.69 and 0.85 for age groups 0-4, 5-15 and 16 and above, respectively. It can be seen that, for all age groups, the contribution of the census table in the combined *proxy* table is larger than the contribution of the ESC table, potentially due to the small time difference between APS and proxy tables.

The last three columns in Table 3 contain the information to perform a formal statistical test (Likelihood Ratio Test - LRT) comparing the three models in increasing order of complexity, as explained in Section 3.2. The differences in the deviances between the models were larger than the critical values, indicating that a more complex model leads to a slightly better fit. In this application, Model 3 would be the recommended option, although the estimated within-area proportions are similar under all three models. For illustration, the within-area GSPREE proportions estimated under Model 2 and Model 3 are compared in Figure 5. Measures of variability for the GSPREE estimates are obtained via resampling process (bootstrap) and are discussed in the next Section.

Table 3. Model Fitting Results. GSPREE Fixed Effects Models.

Model	Age group	Estimated Coefficients	Deviance	Difference Deviance	Critical Value (5% Sig.)
1) LA x Ethnicity Census 2011	-	β =0.99	4840.67 (7517.30*)	-	
2) LA x Ethnicity | Age Census 2011	0-4	β =0.93	1581.10	2) vs 1): 35.77	5.991
	5-15	β =0.94	2468.27		
	16 or above	β =1.00	3432.15		
3) LA x Ethnicity | Age Census 2011 & School Census	0-4	β =0.91; δ =0.74	1568.42	3) vs 2): 95.30	7.815
	5-15	β =0.94; δ =0.69	2443.75		
	16 or above	β =0.98; δ =0.85	3374.05		

* Deviance of Model 2 with β =0.99 in each age group.

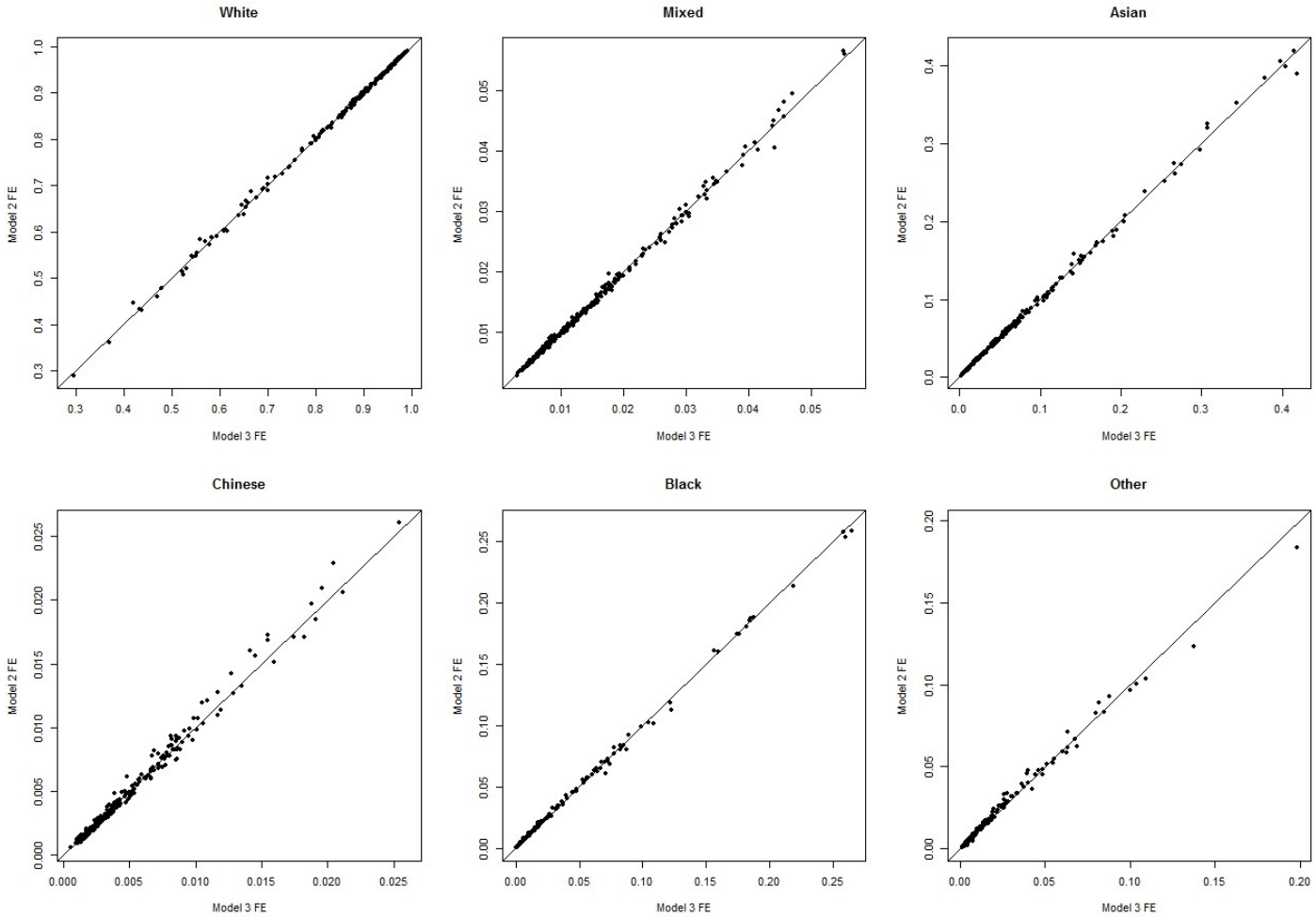


Figure 5. Local authority population proportions by ethnic group according to Model 2 and Model 3. Fixed effects GSPREE. Line: Y=X.

3.4. Bias and Mean Square Error Evaluation for the GSPREE Estimators

Estimation of accuracy (bias and mean square error) under the GSPREE approach can be obtained via resampling methods (e.g. bootstrap). In this Section, the GSPREE estimators are assessed in terms of bias and mean square error via a bootstrap. This requires deriving a synthetic population from which a large number B of bootstrap samples is withdrawn. For each sample, the GSPREE estimated proportions are obtained and compared to the "true" proportions from the synthetic population across all bootstrap samples. This type of assessment gives a general appreciation of the level of variability in the GSPREE estimators and it is closely related to the characteristics of the particular synthetic population generated.

The GSPREE estimators obtained under the three models described in Section 3.2 are assessed in terms of bias and mean square error based on a specific synthetic population using bootstrap. These are referred to as finite population bias (FP-Bias) and finite population mean square error (FP-MSE), where the synthetic population is kept fixed and only the sample changes at each bootstrap iteration.

It has been decided to generate the synthetic population under a mixed effects GSPREE approach to allow for extra variability across area and ethnic groups. This model is more complex and involves adding explicitly in equation 2 area-ethnicity specific random effects with group-specific variances. This means that, to generate the synthetic population, the model explicitly accounts for the effect of each area-ethnic group on the estimated proportions. However, when GSPREE is applied to each of the bootstrap samples, the fixed effects model in Equation 3 is fitted under the three modelling strategies showed in Table 2.

To obtain a synthetic population under the mixed effects GSPREE, variance components for each ethnic group need to be estimated. This is obtained from the two *proxy* tables 2014 ESC and 2011 Census. The ESC is considered as a large sample from the true population for the age group 5-15 and the mixed effects GSPREE is applied. In this application, the estimated variance components for White and Black are negative and they are replaced with the minimum value estimated for the other four groups (0.0196). The estimates used are: 0.0196 for White, 0.0196 for Mixed, 0.0531 for Asian, 0.1515 for Chinese, 0.0196 for Black and 0.7593 for Other. These variance estimates are used in all three age groups (as there is no ESC information available for 0-4 and 16 and above) to generate the synthetic population compositions (or proportions) from which 5,000 bootstrap samples are selected.

Given that the implied sampling fractions of the APS are negligible, bootstrap samples were randomly generated from a plausible population composition, instead of randomly selecting them from the fixed synthetic population, making the process quicker. For this, Multinomial and Poisson sampling are applied, but results are shown only for the Multinomial case (which assumes the observed sample size in each area as fixed), as the findings are similar.

Table 4 shows average FP-Bias and average square root FP-MSE by ethnic group for Models 1, 2 and 3 (fixed effects GSPREE) using Multinomial sampling. Even

though the average FP-Bias is close to zero for all ethnic groups, non-negligible FP-bias and square root FP-MSE are observed in some areas for all fixed effects estimators.

The way the synthetic population is generated for the bootstrap procedure is crucial for the MSE estimates, further research is required to ensure the best method is used to generate this. A definitive methodology for MSE estimation could be implemented from a more realistic synthetic population, being set up using census information, for example.

Table 4. Average FP-Bias and square root FP-MSE by ethnic group for fixed effects GSPREE estimators (Models 1, 2 and 3). Multinomial sampling.

		Ethnicity					
Measure	**Model**	**White**	**Mixed**	**Asian**	**Chinese**	**Black**	**Other**
Average FP-Bias	Model 1 FE	-0.00087	0.00001	0.00035	0.00042	0.00049	-0.00040
	Model 2 FE	-0.00094	0.00005	0.00038	0.00044	0.00052	-0.00044
	Model 3 FE	-0.00091	0.00014	0.00029	0.00047	0.00044	-0.00043
Average Square Root FP-MSE	Model 1 FE	0.01076	0.00194	0.00776	0.00185	0.00334	0.00804
	Model 2 FE	0.01077	0.00190	0.00784	0.00187	0.00336	0.00806
	Model 3 FE	0.01075	0.00187	0.00767	0.00191	0.00325	0.00813

4. Validation Study: Comparing 2011 Census to 2011 GSPREE Population Estimates by Local Authority and Ethnic Group in England

A validation study is conducted whereby GSPREE estimates are derived for March 2011 using 2001 Census data and auxiliary data near to the reference data in order to update the cross tabulation structure and margins. A comparison can then be made between the GSPREE estimates for 2011 and the census estimates. Also the longer time period between the previous census and the reference date of the GSPREE estimates allows a clearer assessment of the performance of the GSPREE method and contribution of the auxiliary (proxy) data sources on the estimates.

The idea of the validation study is to produce GSPREE estimates for a time period in which a reliable estimate of the target table is available, so that it can be considered as the "true" within-area ethnicity distribution. In this way, it is possible to further investigate the error underlying the GSPREE estimates and better define modelling strategies to meet business needs.

The scenarios considered in this validation study are similar to those used to obtain the 2012 (see Section 1) and 2014 (see Section 3) GSPREE estimates in terms of the format of the target table, modelling strategies and integration of data sources. Details of the validation study are presented in the following sub-sections.

4.1. Data Sources

The aim of this validation study is to produce 2011 GSPREE estimates for the within-area ethnicity distribution based on 2001, 2011 and 2013 aggregate auxiliary data. The target table is the cross-classification; LA by broad ethnic group (White, Mixed, Asian, Chinese, Black and Other). The 2001 Census and 2013 ESC are used as *proxy* tables and the 2011 APS (from October 2010 to September 2011) and 2011 Census provide the column and row totals, respectively, to apply GSPREE. The within-area ethnicity distribution based on 2011 Census is a reliable estimate of the target table, and is then assumed as the true within-area ethnicity distribution for 2011 for error calculation purposes. The data sources are combined such that March 2011 is the reference period, which is in line with the census reference period.

It is worth noting that, in real GSPREE applications, the MYEs would be used as row totals because the target table based on the census data is unknown for the time

period of interest. In this validation, the 2011 Census row totals are used to exclude uncertainty caused by the process of updating the census estimates to the mid-year reference date. Their use allows a more direct assessment of how GSPREE estimates capture changes in the within-area ethnicity distribution over the 2001-2011 period. In addition, the 2013 ESC is used as *proxy* table instead of the 2011 ESC (which is closer to the required reference date of the estimates), for practical reasons. This should not have a significant impact on the results due to the small time difference (and, consequently, small changes on the association structures) between the two School Censuses.

4.2. Modelling Strategies

Similarly to the models described in Section 3, three models are assessed in the validation study in decreasing order of complexity: Model 1, which uses only the 2001 Census as *proxy* table; Model 2, which is similar to Model 1 but with independent fittings in each age group; and Model 3, which is similar to Model 2 but considers both the 2001 Census and the 2013 English School Census as *proxy* tables. Table 5 summarises the three modelling strategies.

Table 5. 2011 GSPREE Estimates. Modelling Approaches. Fixed Effects Models.

Fixed Effects GSPREE Approaches	**Proxy Information**	**Disaggregation - Proxy and Survey Tables**	**Age Group**
Model 1	2001 Census	LA x Ethnicity	None
Model 2	2001 Census	LA x Ethnicity | Age Group	0-4 5-15 16 and above
Model 3	2001 Census and 2013 School Census	LA x Ethnicity | Age Group	0-4 5-15 16 and above

4.3. Results

Model Interpretation

Estimation of the association structure (β) between *proxy* and survey tables in all modelling strategies is obtained via Poisson Maximum Likelihood Estimation based on Equation (2) (see Section 2.2). Results are presented in Table 6.

In terms of the estimated proportionality constant $\hat{\beta}$, it can be seen that there is some departure from 1 for all modelling strategies. This indicates that, for the two-way LA by ethnic group table, the survey information is successful in updating the association structure in the proxy table (2001 Census).

The contribution of the recent auxiliary information (2013 ESC) is evident in Model 3, where the estimated weights given to the 2001 Census proxy table, $\hat{\delta}$, are 0.22, 0.27 and 0.66 for age groups 0-4, 5-15 and 16 and above, respectively. This shows that the contribution of the 2001 Census in the combined *proxy table* is small, reducing reliance in the census in favour of more up-to-date information from the ESC. This is true for all age groups. Note that these weights are considerably smaller than those found in previous GSPREE applications.

With the inclusion of the more recent ESC information the $\hat{\beta}$ values for the 0-4 and 5-15 age groups have been estimated closer to 1 compared to the $\hat{\beta}$ values for Model 2. This indicates that the School Census is bringing the association structure of the auxiliary information closer to that of the survey. The large time difference between recent auxiliary information and proxy table has allowed GSPREE to capture changes in the population distribution across ethnic groups.

Table 6. 2011 GSPREE Estimates. Model Fitting Results. GSPREE Fixed Effects Models.

Model	Age group	Estimated Coefficients	Deviance	Difference Deviance	Critical Value (5% Sig.)
1) LA x Ethnicity Census 2001	-	$\hat{\beta}$=0.92	8782.09*	-	
2) LA x Ethnicity | Age Census 2001	0-4	$\hat{\beta}$=0.79	1888.91	2) vs 1) 109.79	5.991
	5-15	$\hat{\beta}$=0.86	2696.11		
	16 and above	$\hat{\beta}$=0.94	4087.28		
3) LA x Ethnicity | Age Census 2001 and 2013 School Census	0-4	$\hat{\beta}$=0.85; $\hat{\delta}$=0.22	1717.60	3) vs 2) 797.08	7.815
	5-15	$\hat{\beta}$=0.93; $\hat{\delta}$=0.27	2386.99		
	16 and above	$\hat{\beta}$=0.92; $\hat{\delta}$=0.66	3770.62		

* Deviance of Model 2 with $\hat{\beta}$=0.92 in each age group.

When performing the Likelihood Ratio Test based on information from the last three columns in Table 3 (see Section 3.2), it can be said that there is strong evidence that a more complex model leads to a better fit. In this validation, Model 3 would be the recommended option, although the estimated within-area proportions are similar under Model 2 and Model 3 (See Figure 6).

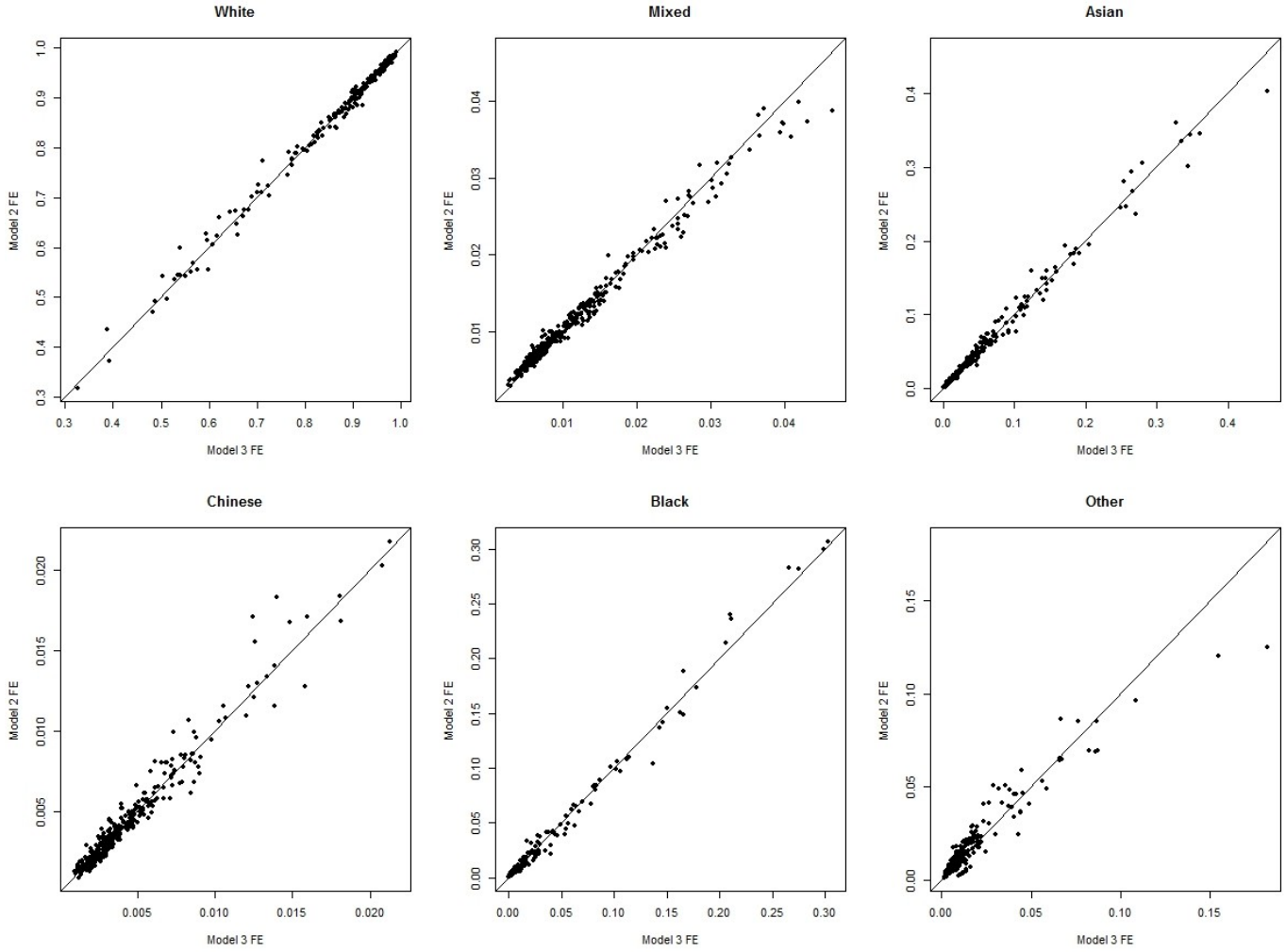


Figure 6. 2011 GSPREE Estimates. Local authority population proportions by ethnic group according to Model 2 and Model 3. Fixed effects GSPREE. Line: Y=X.

Bias, MSE and Coefficient of Variation

Measures of variability for the GSPREE estimates are obtained via resampling process (bootstrap). As mentioned in Section 3.4, the performance of the GSPREE estimates are closely related to the synthetic population from which the bootstrap samples are selected. Table 7 shows the average bias and square root MSE by ethnic group (for the 2011 GSPREE estimates) for Models 1, 2 and 3 under Multinomial sampling. Overall, all three models show very similar performance, with small bias and square root MSE in most LAs, but some non-negligible estimates are observed in some LAs. Figure 7 illustrates the accuracy of the estimates in terms of percent coefficient of variation – CV (square root MSE divided by the point estimate).

On average, the CVs are reasonable for White, Mixed and Asian ethnic groups. For Chinese, Black and Other, the CVs are extremely high for at least half of the LAs.

Table 7. 2011 GSPREE Estimates. Average FP-Bias and square root FP-MSE by ethnic group for fixed effects GSPREE estimators (Models 1, 2 and 3). Multinomial sampling.

		Ethnic Group					
Measure	**Model**	**White**	**Mixed**	**Asian**	**Chinese**	**Black**	**Other**
Average FP-Bias	Model 1 FE	-0.00355	0.00035	0.00089	0.00044	0.00012	0.00174
	Model 2 FE	-0.00350	0.00041	0.00098	0.00042	0.00005	0.00163
	Model 3 FE	-0.00231	0.00037	0.00096	0.00040	-0.00019	0.00078
Average Square Root FP-MSE	Model 1 FE	0.01943	0.00242	0.01043	0.00161	0.01237	0.01138
	Model 2 FE	0.01940	0.00237	0.01050	0.00156	0.01229	0.01130
	Model 3 FE	0.01766	0.00223	0.00986	0.00160	0.01114	0.01040

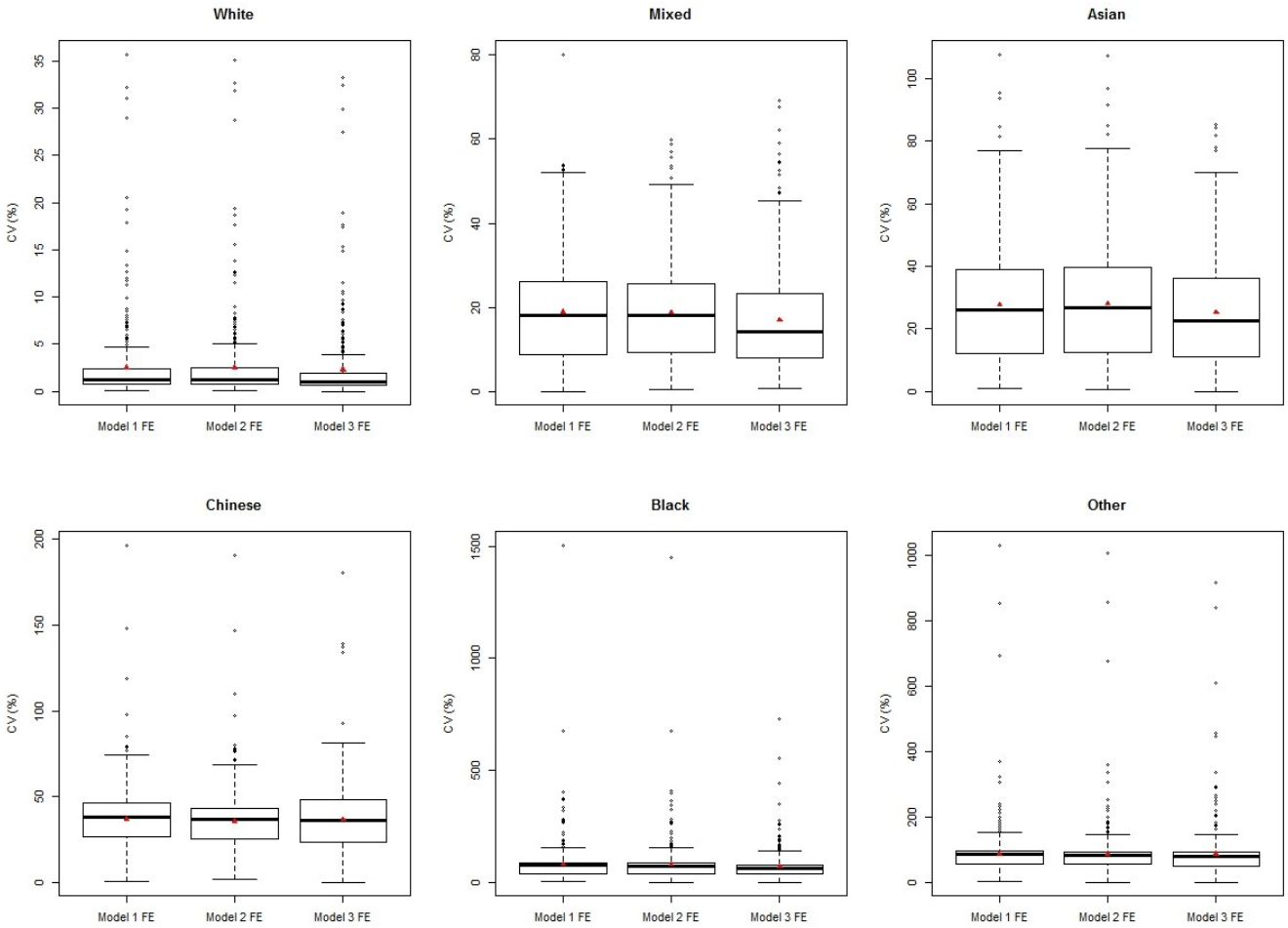


Figure 7. 2011 GSPREE Estimates. Coefficient of variation by ethnic group for fixed effects GSPREE estimators (Models 1, 2 and 3).

2011 GSPREE Estimates versus 2011 Census

For validation purposes, the GSPREE estimates based on Model 3 will be used. Before assessing the error in the 2011 GSPREE estimates when compared to the 2011 Census estimates, it is important to verify how the within-area ethnicity distributions changed from 2001 to 2011 based on censuses data and assess if the 2011 GSPREE estimates are in line with the findings based on the 2011 Census. Figure 8 compares the 2001 and 2011 Censuses proportions by LA and ethnic group and, on average, the population proportions by LA increased from 2001 to 2011 for all ethnic groups, except for White. Figure 9 shows 2001 Census and 2011 GSPREE proportions by LA and ethnic group. Overall, the 2011 GSPREE estimates are successful in capturing changes for the largest ethnic groups (White, Asian and

Black) in the 2001-2011 period, as Figure 9 shows similar patterns to those in Figure 8. However, for the Other, Mixed and Chinese ethnic groups the GSPREE estimates do not fully reflect the patterns observed in the 2011 Census.

Table 8 shows summary of the distribution of the difference between the 2011 GSPREE (Model 3) estimates and the 2011 Census across LAs for each ethnic group. On average, the differences are reasonable, although considerable differences can be observed for some LAs and ethnic groups (e.g. Mixed and Other). To complement the information in Table 8, the summary distribution of the relative differences is shown in Table 9, and using box-plot format in Figure 10.

The findings from this validation study are promising; over this longer time period the GSPREE method has captured some of the changes in the distribution of ethnicity by LAs. It has also highlighted issues for further research, for example to improve the LAs showing large differences in the ethnic group distributions when compared to the 2011 Census.

As a further study, it would be appropriate to investigate how the GSPREE distribution of ethnicity compares to the census distribution when analysing London and non-London groups separately, as changes in ethnic distributions could be different in London due to differences in migration patterns. The large CVs may be partly due to incorrect specification of the synthetic population used so to estimate the variance so again this could be a focus for further research.

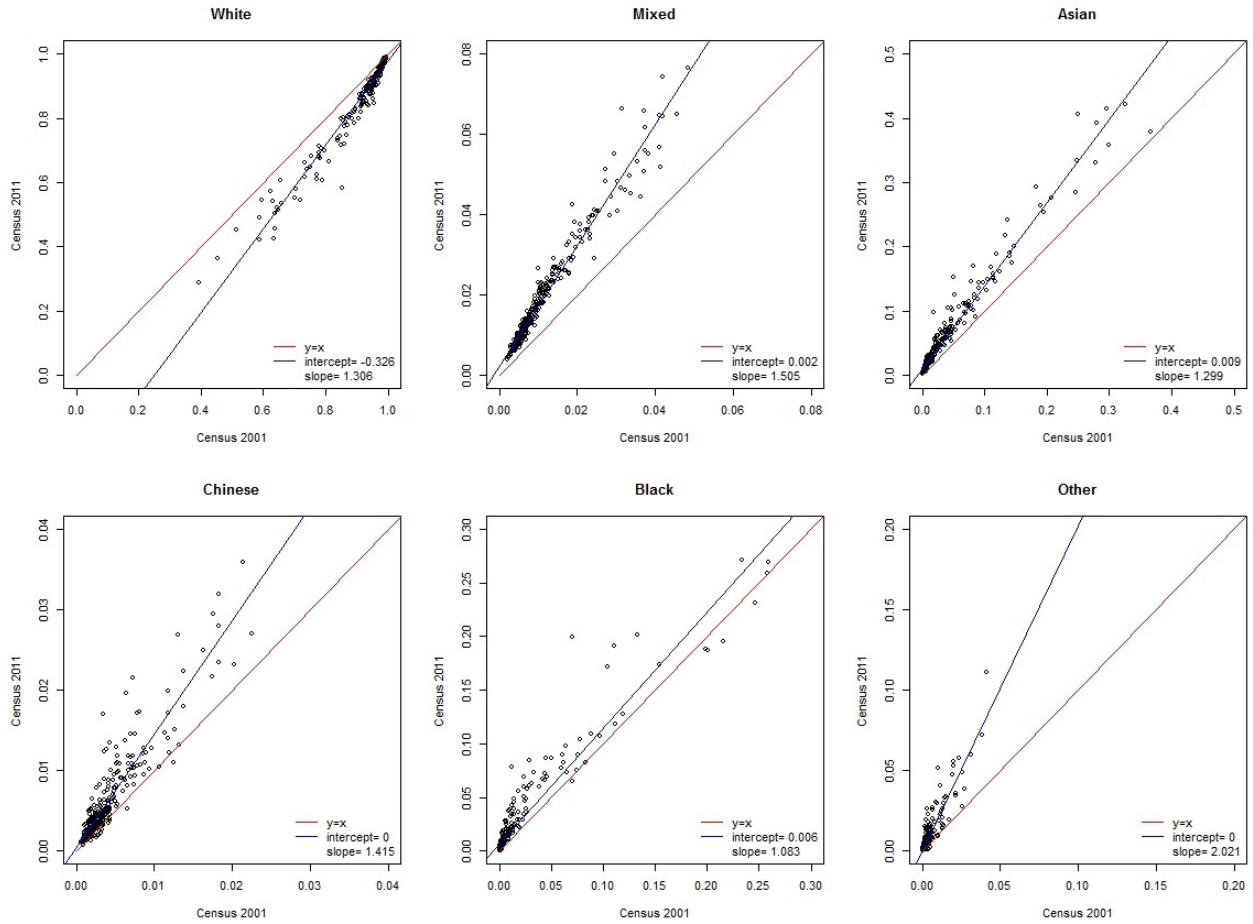


Figure 8. 2001 and 2011 Censuses proportions by ethnic group and LAs.

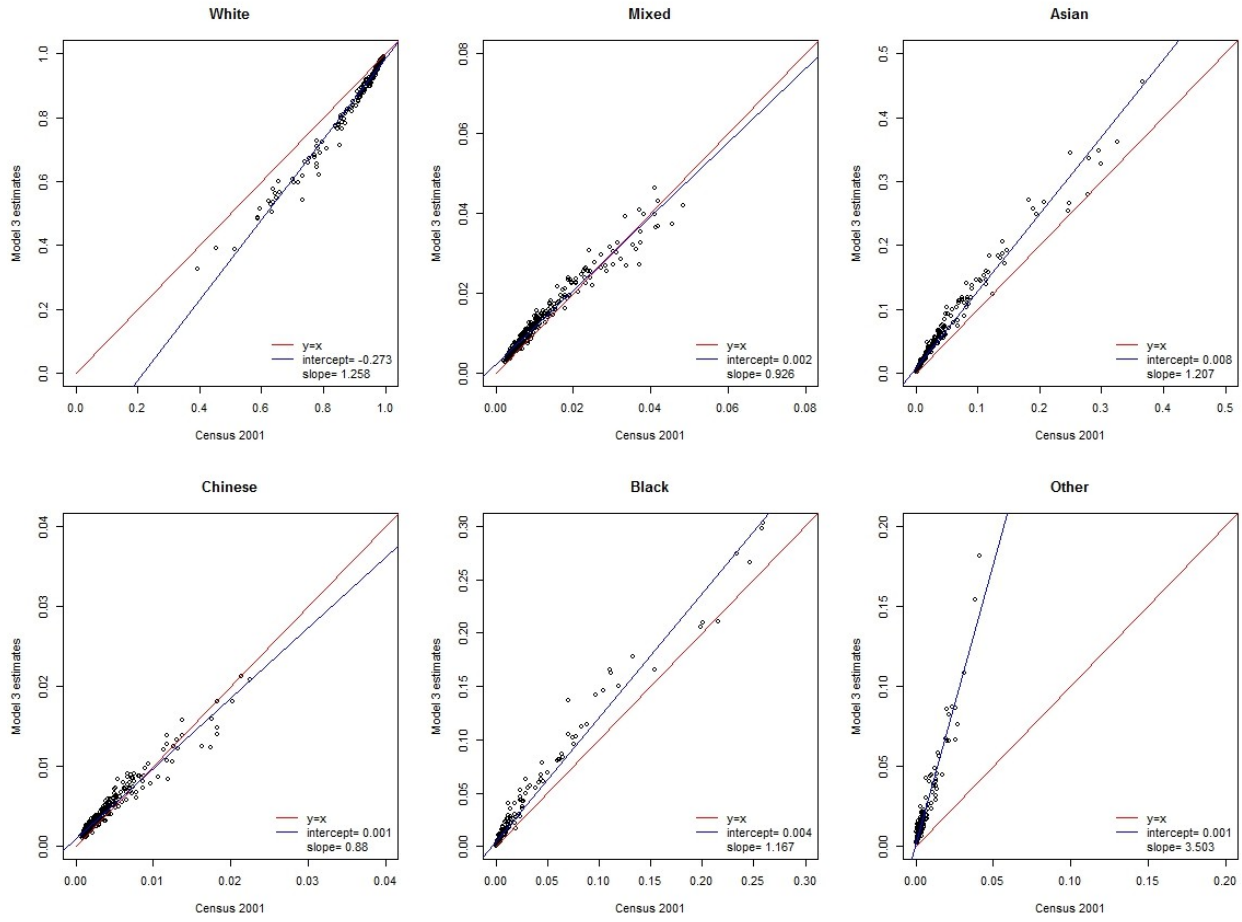


Figure 9. 2001 Census and 2011 GSPREE (Model 3) proportions by ethnic group and LAs.

Table 8. Distribution of the differences[1] across LAs by ethnic group.
2011 GSPREE (Model 3) versus 2011 Census. England.

	Ethnic Group					
	White	**Mixed**	**Asian**	**Chinese**	**Black**	**Other**
Minimum	-18019	-20183	-18781	-6812	-11645	-3410
1st Quartile	17	-1248	-1004	-313	-370	273
Median	472	-631	-195	-60	-93	513
Mean	1674	-1329	-891	-344	-254	1143
3rd Quartile	1946	-319	-22	7	0	900
Maximum	23935	-53	19389	265	11820	24042

[1] 2011 GSPREE - 2011 Census.
Note: Number of local authorities in the analysis is 324.

Table 9. Distribution of the relative differences[1] (%) across LAs by ethnic group.
2011 GSPREE (Model 3) versus 2011 Census. England.

	Ethnic Group					
	White	**Mixed**	**Asian**	**Chinese**	**Black**	**Other**
Minimum	-14.41	-57.51	-52.36	-72.14	-85.54	-48.94
1st Quartile	0.02	-37.83	-21.67	-29.65	-23.60	67.60
Median	0.38	-33.46	-12.23	-14.49	-11.96	132.12
Mean	1.11	-32.98	-11.36	-12.96	-10.38	184.10
3rd Quartile	1.41	-28.17	-1.48	2.23	-0.14	222.73
Maximum	22.08	-10.96	52.10	66.75	116.52	4031.01

[1] 100 x (2011 GSPREE - 2011 Census)/ 2011 Census.
Note: Number of local authorities in the analysis is 324.

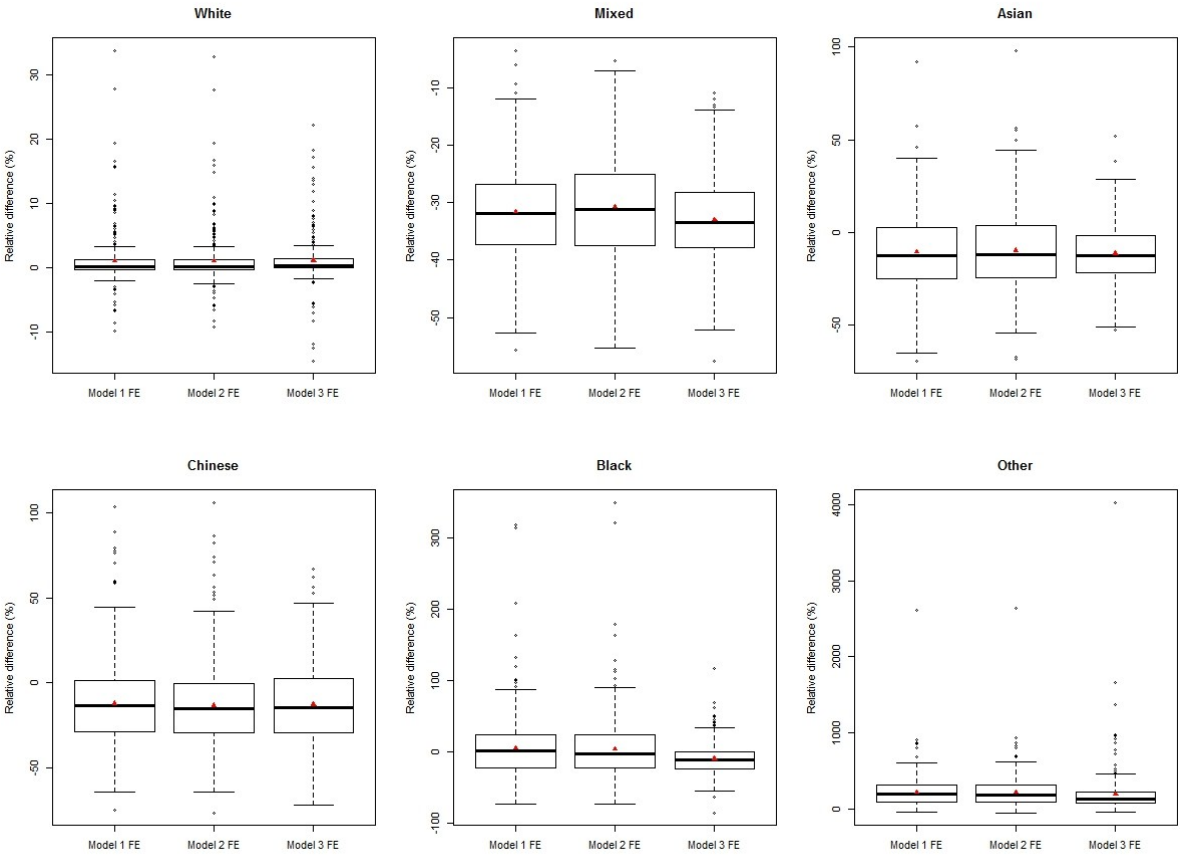


Figure 10. Relative differences across LAs by ethnic group. 2011 GSPREE (Model 3) versus 2011 Census. England.

5. Concluding Remarks and Further Work

In this work, a feasibility study is conducted to produce small area estimates of the within-area distribution of ethnicity by local authority in England using the Generalised Structure Preserving Estimators (GSPREE). This is the first time this approach has been implemented in the UK.

In contrast to other demographic and socioeconomic characteristics, ethnicity is a variable for which there is not a clear set of covariates identified in the literature that could be used as predictors for the typical model-based approaches used in small area estimation. In fact, unless *proxy* information is involved, it seems difficult to expect a good performance of a small area estimator in this context. Structure Preserving Estimators can be applied in this situation, given that *proxy* compositions can be obtained either from the last population census or from other sources, such as the School Census.

The pure SPREE method proposed by Purcell and Kish (1980) assumes that the detailed structure of the target cross-tabulation is the *same* as that shown in the auxiliary tabulation (census data). However, this assumption is difficult to justify in practise, as the structure found in the census data is likely to become out of date and departures from this assumption can result in biased SPREE estimators. The Generalised Structure Preserving Estimation (GSPREE) approach (Zhang and Chambers, 2004) aims at relaxing this assumption. The GSPREE expresses the estimators via log-linear models and is more flexible than the SPREE, allowing for more recent information from other surveys or data sources (if available) to be incorporated in the estimation process.

In this application, three alternative models to produce population estimates by ethnic groups using GSPREE are formulated with increasing levels of complexity in terms of estimation strategy and use of additional auxiliary sources. Overall, application of GSPREE with a combined *proxy* table (based on ESC and on census) and with separate fits by age group shows improvement on the estimators, although further research is needed to potentially enhance the contribution of the auxiliary information.

Similar findings were identified when using 2012-2013 auxiliary data (in the previous application), but the Likelihood Ratio Tests for 2011 and 2014 gave stronger evidence in favour of Model 2 (when compared to Model 1) and Model 3 (when compared to Model 2) where difference in the deviances were larger. The mixed effects GSPREE used to generate the synthetic population also produced negative variance estimates for White and Black ethnic groups in the previous analysis. A validation of the GSPREE method has also been conducted comparing the 2011 GSPREE estimates of the within-area ethnicity distribution (based on 2001, 2011 and 2013 auxiliary information) to the within-area ethnicity distribution based on the 2011 Census. Results show that, overall, the GSPREE method is fairly successful in capturing changes in the within-area ethnicity distribution in the 2001-2011 period, although further research is needed to improve measures of accuracy and the use of the ESC information (or another auxiliary information).

Future Work

The GSPREE methodology is very flexible and different ways of accounting for the auxiliary information can be incorporated in the method. Areas for further research include:

1) Evaluation of the GSPREE estimators in terms of bias and MSE. The performance of the estimators is closely related to the plausibility of the characteristics of the synthetic finite population (or population composition) from which the bootstrap samples are extracted (see Section 3.4). Additional work in this direction is needed, addressing alternative scenarios that would allow a more complete and realistic evaluation, increasing the knowledge on the performance of the proposed estimators.
2) Alternative modelling strategies (given that the GSPREE estimates only use one proportionality constant for all ethnic groups). For example the model could be fitted separately to London and non-London LAs to allow for different patterns of population change due to migration, and its consequent impact on ethnic group distributions. Similarly, a hierarchical approach could be adopted by fitting separate models for large ethnic groups (e.g. White) to avoid the estimated proportionality constant ($\hat{\beta}$) to be dominated by them.

3) Strategies for estimating more detailed population subcategories (e.g. more detailed ethnic groups or a three way table such as the population by LA, ethnic group and age groups), and how the estimates can be used to measure change in the distribution of local authority population by ethnic group over time.
4) Inclusion of additional sources of data as auxiliary sources as they become available, with the aim of reducing reliance on the census distributions.

Note that the issue of not having large enough sample sizes to fit mixed effects models is not a problem of this particular application, but rather a problem that all applications of small area estimation face sooner or later when the aim is to produce estimates at increasingly lower levels of disaggregation. In this sense, it is important to address the problem of how to improve synthetic predictors (such as GSPREE) as a priority.

6. Acknowledgements

The Office for National Statistics is grateful to Angela Luna-Hernandez (University of Southampton) and Professor Li-Chun Zhang (University of Southampton and Statistics Norway) for development and implementation of methods and for continuous advice and support.

7. References

Agresti, A. (2013). Categorical Data Analysis. John Wiley & Sons.

Berg, E. J. and Fuller, W.A. (2014). Small Area Prediction of Proportions with Applications to the Canadian Labour Force Survey. Journal of Survey Statistics and Methodology, **2**, 227–56.

Cinco, M. (2010). Intercensal Updating of Small Area Estimates. Unpublished PhD thesis. Massey University.

Green, A., Haslett, S., and Zingel, C. (1998). Small Area Estimation Given Regular Updates of Census Auxiliary Variables. In Proceedings of the New Techniques and Technologies for Statistics Conference, 206–211.

Luna-Hernandez, A. (2014). On Small Area Estimation for Compositions Using Structure Preserving Models. Unpublished PhD upgrade document, Department of Social Statistics and Demography, University of Southampton.

Luna-Hernandez, A., Zhang, L., Whitworth, A. and Piller, K. (2015) Small Area Estimates of the Population Distribution by Ethnic Group in England: A Proposal Using Structure Preserving Estimators. Statistics in Transition New Series and Survey Methodology Joint Issue: Small Area Estimation 2014. Vol. 14, No.4, pp. 585-602.

Molina, I., Ayoub, S. and Lombardia, M.J. (2007). Small Area Estimates of Labour Force Participation under a Multinomial Logit Mixed Model. Journal of the Royal Statistical Society, A, **170**, 975–1000.

Noble, A., Hasett, S. and Arnold, G. (2002). Small Area Estimation via Generalized Linear Models. Journal of Official Statistics, **18,** 45–68.

Office for National Statistics (2016). Annual assessment of ONS's progress towards an Administrative Data Census Post 2021. Available at

https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusannualassessments

Office for National Statistics (2005). Making a population estimate in England and Wales. Available at http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=575

Office for National Statistics, Luna-Hernandez A. and Zhang, L. (2015) Interim Report on Combining Census and Survey Data to Estimate Local Authority Population by Ethnic Group. ONS Internal Report.

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. Statistical Science, **28**, 40–68.

Purcell, N. J. and Kish, L. (1980). Postcensal Estimates for Local Areas (or Domains). International Statistical Review, **48**, 3-18.

Rao, J.N.K. and Molina, I. (2015). Small Area Estimation. Second Edition. John Wiley & Sons.

J. Scealy (2010). Small Area Estimation Using a Multinomial Logit Mixed Model with Category Specific Random Effects. Research paper. Australian Bureau of Statistics.

Zhang, L.C. and Chambers, R. (2004). Small area estimates for cross-classifications. Journal of the Royal Statistical Society, B, **66**, 479–496.

Appendix A

The Generalized Linear Structural Model for estimating β Zhang and Chambers (2004) relates the within-area proportions of the proxy table and the table of interest on the log scale centred on the average of the area.

The equation that defines the GLSM is:

$$\eta_{aj}^{Y} = \lambda_j + \beta \eta_{aj}^{X} \tag{3}$$

Where $\eta_{aj}^{Z} = \log \theta_{aj}^{Z} - J^{-1} \sum_k \log \theta_{ak}^{Z}$, $\theta_{ak}^{Z} = Z_{ak} \Big/ \sum_l Z_{al}$ for $Z = X, Y$, and $\sum_j \lambda_j = 0$.

The terms in the decomposition given in equation (1) satisfy $\sum_j \alpha_j^{Z} = 0$ and $\sum_j \alpha_{aj}^{Z} = \sum_a \alpha_{aj}^{Z} = 0$ for $Z = X, Y$. Moreover, $\alpha_j^{\theta} = \alpha_j^{Y}$ and $\left\{\alpha_{aj}^{\theta}\right\} = \left\{\alpha_{aj}^{Y}\right\}$. Using these arguments it is straightforward to show that $\eta_{aj}^{Z} = \alpha_j^{Z} + \alpha_{aj}^{Z}$ for $Z = X, Y$, and therefore,

that equation (3) is equivalent to the structural equation of the GSPREE. The λ_j are nuisance parameters with no practical interest.

The GLSM in equation (3) is fitted via Iteratively Weighted Least Squares (IWLS) using direct estimates of the within-area proportions $\hat{\theta}_{aj}^{Y}$ and estimates of their variances. By doing so, it is implicitly assumed that the structural equation of the GSPREE holds for the table of direct estimates as well, or at least, that the value of β that better relates the table of interest and the proxy table does not change when the former is substituted by its direct estimate.