STATISTICS IN TRANSITION new series and *SURVEY METHODOLOGY*
Joint Issue: Small Area Estimation 2014
Vol. 16, No. 4, pp. 585–602

SMALL AREA ESTIMATES OF THE POPULATION DISTRIBUTION BY ETHNIC GROUP IN ENGLAND: A PROPOSAL USING STRUCTURE PRESERVING ESTIMATORS

Angela Luna, Li-Chun Zhang[1], Alison Whitworth, Kirsten Piller[2]

ABSTRACT

This paper addresses the problem of producing small area estimates of Ethnicity by Local Authority in England. A Structure Preserving approach is proposed, making use of the Generalized Structure Preserving Estimator. In order to identify the best way to use the available aggregate information, three fixed effects models with increasing levels of complexity were tested. Finite Population Mean Square Errors were estimated using a bootstrap approach. However, more complex models did not perform substantially better than simpler ones. A mixed-effects approach does not seem suitable for this particular application because of the very small sample sizes observed in many areas. Further research on a more flexible fixed-effects estimator is proposed.

1. Introduction

Estimates of demographic characteristics are among the main outputs of National Statistical Institutes (NSIs). In addition to national and regional estimates, for topics such as Labour Force, Household composition or Ethnicity, periodic estimates at lower levels of geographic aggregation are in high demand both for public policy and research purposes.

In census years, given the availability of data for almost all individuals in the population, it is straightforward to produce reliable estimates for small geographic domains. In contrast, during the inter-censal period, updated socio-demographic data can only be obtained via sample surveys or administrative systems. It is generally difficult to obtain reliable direct estimates for small geographic domains from sample surveys due to the small sample sizes. Data from administrative systems do not have this problem but, in contrast, may not cover the topics of interest. Moreover, definitions of the variables and domains in administrative

[1] University of Southampton.
[2] Office for National Statistics ONS-UK.

sources reflect the requirements of the administrative systems, which may be different from those for statistical purposes. This can result in comparability issues with figures obtained from population censuses or household surveys.

From the statistical perspective, the estimation for small domains in the presence of limited, or even null domain-specific sampling data can be framed within the field of Small Area Estimation (SAE). Research in SAE has gained relevance in the last decades due to an increasing demand for small area outputs in the Official Statistics sector, as well as in many others. Readers interested in SAE can find a comprehensive account of methods in Rao (2003). For a review of the most important developments of the last decade, see Pfeffermann (2013).

Implementation of SAE methods in the field of Official Statistics faces specific challenges and specialized research has been encouraged in the European context. Projects such as EURAREA (Eurostat, 2001-2004) and EU-SAE (ESSnet, 2009-2012) provide comprehensive reviews of the available SAE methods with potential applications in a broad set of topics covered by Official Statistics, taking into consideration the specific requirements and characteristics of European statistical systems. Special attention has been given to the use of SAE methods for the measurement of poverty, via collaborative projects such as SAMPLE (Small Area Methods for Poverty and Living conditions Estimates, European Commission, 2008-2011) or AMELI (Advanced Methodology for Laeken Indicators, 2008-2011). The deliverables of all the above mentioned projects are available online.

At present, comparatively few official figures in the region are being produced using SAE methods. In the UK case, the Office for National Statistics (ONS) periodically disseminates small area estimates regarding three main topics: population estimates by age and sex using the Census and its coverage survey; average household income and households in poverty using the Family Resources Survey and administrative data maintained by the Department for Work and Pensions; and unemployment, making use of the Annual Population Survey and the administrative register for Jobseeker's Allowance.

Nonetheless, the interest in understanding the potential gains that can be obtained from a more extensive use of SAE methods in the context of official statistics remains. The ONS established the Census Transformation Programme in January 2015 to take forward the National Statistician's recommendation to make the best use of all available data in the production of population statistics. This involves research into the potential use of administrative data as well as surveys to produce population, household and characteristic information currently provided in the Census. SAE methods provide a framework for intergrating sources. In this paper we investigate the problem of how to obtain estimates of the distribution of the population by ethnic group, in each Local Authority (LA) of England, using proxy and survey data. Such estimates are required by local and central government for planning services and formulating policy. More generally researchers, local authorities, health authorities and other public and private sector organisations could use them to gain an up-to-date picture of the ethnic

composition of local populations and to monitor diversity and anti-discrimination programmes.

Ethnicity is a variable for which the use of Structure Preserving Estimators (SPREE) (Purcell and Kish, 1980) seems natural. Most SAE methods combine existing survey data for the variable of interest with relevant covariate information obtained from censuses or administrative sources, to obtain better estimates than those from the survey alone. For Labour Force status for instance, covariates such as sex, age or level of education can provide some explanatory power, see Molina et al. (2007) and Scealy (2010). In the case of ethnicity, on the other hand, it is difficult to identify such a set of covariates. Instead, for post-censal updates of the LA by ethnicity distribution, the corresponding aggregated census table can always be treated as a proxy for the table of interest.

When a proxy is available, the SPREE approach allows for an intuitive modelling of the relationship between the so-called association structure, or simply the structure of both the proxy table and the table of interest. The SPREE approach is particularly compelling in the case where the margins of the table of interest are known in advance or can be accurately estimated because, given the margins, the structure is the only unknown component to be estimated. This will be explained in more detail in Section 2.

This application addresses the particular problem of obtaining updated census tables of LA by ethnicity during the inter-censal period. However, it is important to notice that population censuses in general are going through a process of redesign in many European countries. More emphasis is being given to alternative operations based on demographic systems that use information from administrative sources alone or in combination with survey data. In such a context, the potential impact of SAE methods, including the SPREE approach and its extensions, is expected to increase considerably in the future.

The rest of the paper is organised as follows. In the next section, the underpinning idea behind the SPREE approach and the GSPREE extension (Zhang and Chambers, 2004) is discussed in more detail. Section 3 describes the characteristics of the empirical exercise performed to obtain estimates of the distribution by ethnic group and LA in England. Section 4 presents the results of our analysis. Finally, Section 5 discusses the main results and points out some topics for future work.

2. SPREE approach

2.1. Structure Preserving Estimator (SPREE)

Denote by Y the population table of interest, with cells Y_{aj}, where $a = 1,\ldots,A$ indexes the set of areas and $j = 1,\ldots,J$ indexes the categories of the

variable. Define $\zeta_{aj}^{Y} = \log Y_{aj}$. Y can be represented in the form of a saturated log-linear model as:

$$\zeta_{aj}^{Y} = \alpha_0^{Y} + \alpha_a^{Y} + \alpha_j^{Y} + \alpha_{aj}^{Y}, \qquad (1)$$

where $\alpha_0^{Y} = \overline{\zeta^{Y}}_{..}$ (the dot indicating summing over the respective subscript), $\alpha_a^{Y} = \overline{\zeta^{Y}}_{a.} - \alpha_0^{Y}$, $\alpha_j^{Y} = \overline{\zeta^{Y}}_{.j} - \alpha_0^{Y}$ and $\alpha_{aj}^{Y} = \zeta_{aj}^{Y} - \alpha_0^{Y} - \alpha_a^{Y} - \alpha_j^{Y}$, for $a = 1,\ldots,A$, $j = 1,\ldots,J$. Following Purcell and Kish (1980), equation (1) can be used to decompose Y into 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 latter, 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, it can usually be accurately estimated using survey data, if not available from other sources.

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. The term proxy is used here in the customary sense of *proxy variable* as defined in Upton and Cook (2008): "A measured variable that is used in the place of a variable that cannot be measured". 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 dimension $A \times J$. Notice that for demographic characteristics during inter-censal periods, the corresponding tables from the census year are obvious proxies. More generally, proxies can be derived not only from censuses but also from administrative sources.

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

The procedure proposed by Purcell and Kish (1980) to obtain the SPREE of Y is straightforward. The known margins of Y are imposed on X using a multiplicative raking procedure such as the Iterative Proportional Fitting (IPF)

algorithm (see for instance Agresti, 2013, p. 365-366). This ensures that the association structure of the estimated and proxy tables are the same. Fitting a saturated log-linear model with an offset term given by the interactions α_{aj}^{X} is an alternative way to obtain the same estimate (Noble et al., 2002).

However, assuming that the proxy and the table of interest 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 we have called the SPREE approach. The modifications to the initial 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. 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 estimator proposed in Berg and Fuller (2014). Notice that in all the above mentioned cases the allocation structure is imposed by benchmarking the estimates to a set of known margins. The benchmarking has the additional advantage of providing some degree of protection against misspecification of the assumed model (Pfeffermann, 2013).

2.2. Generalized Structure Preserving Estimator (GSPREE)

In some cases, it is possible to have access to a survey estimate of Y. Notice that the small area problem persists because the direct estimates of the cell totals are usually too unstable to be useful, due to small sample sizes. The GSPREE (Zhang and Chambers, 2004) proposes to use such information to *update* the association structure of the proxy table, aiming to reduce the bias of the SPREE. The GSPREE is characterised by the structural equation $\alpha_{aj}^{Y} = \beta\alpha_{aj}^{X}$ for $a = 1,\ldots,A,\ j = 1,\ldots,J.$ Clearly, the SPREE corresponds to the particular case $\beta = 1.$

An estimation procedure for β built directly from the structural equation involves several problems. Small sample sizes can lead to zero survey estimates 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 (GLSM), a model relating the within-area proportions of the proxy table and the table of interest, on the log scale centred around the average of the area. The equation that defines the GLSM is:

$$\eta_{aj}^{Y} = \lambda_j + \beta \eta_{aj}^{X} \tag{2}$$

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

The GLSM 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. Once the estimate $\hat{\beta}$ has been obtained, the GSPREE of Y is calculated by imposing the known row and column margins on the table of exponentiated estimated interactions $\tilde{Y}_{aj} = e^{\hat{\beta} \alpha_{aj}^{X}}$, using IPF.

In the absence of estimates of the variance of the direct estimators, it is also possible to obtain fully model-based estimates of β. One possibility, mentioned 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, assuming a multinomial distribution for the direct estimates of the cell totals.

Despite not being mentioned in Zhang and Chambers (2004), 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}. It is straightforward to show that the equation:

$$\log Y_{aj} = \gamma_a + \lambda_j + \beta \alpha_{aj}^{X} \tag{3}$$

with $\sum_j \lambda_j = 0$ is also equivalent to the structural equation of the GSPREE. Both the γ_a and the λ_j terms for $a = 1, \ldots, A$, $j = 1, \ldots, J$, are nuisance parameters. It is possible to fit (3) in a standard software using log-linear models and obtain the corresponding ML estimator of β. As with the fitting using the GLSM, this

approach assumes that the structural equation also holds for the table of sample counts.

In the application presented in Section 4 we followed a fully model-based approach in order to simplify the fitting process. By doing so, we can be incurring in a 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.

3. Empirical exercise: distribution of the population by Ethnicity at LA level in England

An empirical exercise was conducted with the aim of producing small area estimates of the distribution of the population by ethnic group for each LA in England. Given that some of the sources of information used in this exercise are subject to disclosure control, it was necessary to perform all the data analysis in a Safe Room of the Virtual Microdata Laboratory (VML) of ONS. Thus, in accordance with ONS standards and the principles set out in the Code of Practice for Official Statistics, full account has been taken of requirements to safeguard confidentiality and uphold relevant data security standards. All the calculations hereby presented are the responsibility of the authors.

This section starts with a description of the data sources used: the proxy table, the table of survey estimates and the benchmark totals for the columns and row margins. A description of the variable of interest and the definition of categories across the different sources is then provided. Finally, the models that were involved in the fitting process are presented.

3.1. Sources of information

Proxy Information

Proxy information for the distribution of Ethnicity at the LA level can be obtained for England from several sources. For this empirical exercise, aggregate data from the 2011 Census and the English School Census[3] were used.

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

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

The English School Census targets the population attending school in England and it is carried out every year. It mostly covers the population between 2 and 19 years old, with almost full coverage of children between the compulsory school ages of 5 and 15. The main school census in January collects information on the pupil's ethnicity, which is not asked about in the two other collection periods in June and August. Whereas state maintained schools and non-maintained special schools are included, independent schools are not covered. This can result in some differences between the population estimates for children in compulsory school age obtained from this and other sources.

As the English School Census only provides a good coverage for children between 5 and 15 years old, it could be said that for the empirical exercise there is one source of proxy information for individuals in the ages *0-4* and *16 or more*, and two sources for those *between 5 and 15 years*. In order to use the appropriate models for each age group, age-group specific Census tables of LA by Ethnic group were produced. Regarding the School Census, the empirical exercise hereby presented used information collected in January 2013.

Survey estimates

Most household surveys carried out by the ONS collect demographic data. For this empirical exercise, the Annual Population Survey (APS) is used for the updated estimates for the population by ethnic group. The APS contains detailed information on ethnicity, has the biggest sample size among the periodic surveys and, except for the Isle of Scilly, it includes information for all Local Authorities in England.

The APS is a household survey that is designed to provide information at a local level, on many demographic and socio-economic topics. The data sets are published quarterly (January to December; April to March; July to June; and October to September) and contain approximately 250,000 individuals. They contain the Labour Force Survey (LFS) data and the boost samples to the LFS. The boost 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 set for England therefore consists 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. Notice that the households are included in the APS only the first time they are interviewed, so that each respondent only appears in the data set 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. This empirical exercise uses the data

corresponding to July 2012 – June 2013**.** The reference point is taken as the midpoint, so approximately the 31st of December 2012, which ties in with the School Census data.

As with the data from 2011 Census, a survey table from the APS survey was produced for each one of the age groups *0-4, 5-15* and *16 or more*.

Benchmark totals

Estimates of the LA population sizes can be obtained from the official mid-year population estimates. These estimates are produced using the cohort component method, which uses information on components of population change to update the most recent census population. The previous year's population estimate by sex, age and LA 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 internal (cross border and between LA) 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 2012 and 2013 mid-year population estimates at LA level were used to calculate the row marginal. As the reference date of such estimates is 30th of June of the corresponding year, an average of the mid-year population estimates for 2012 and 2013 would provide an estimate of the population close to the 31st of December 2012, consistent with the reference period of the other sources involved in this exercise.

The direct estimates of the total population size by ethnic group, obtained from the APS at the national level, are used as the column benchmark totals in this exercise. Neither for the ethnic group nor the LA margins, a disaggregation by age group was considered.

3.2. Definition of the categories of the variable

The variable Ethnic group is collected in England in a very detailed way. The APS collects information regarding 18 subcategories of Ethnicity, grouped in 7 main categories: White, Mixed/multiple ethnic groups, Asian/Asian British, Black/African/Caribbean/Black British, Chinese, Arab and Other ethnic group. The 2011 Census uses 18 subcategories grouped in 5 main categories, with Chinese included within Asian and Arab within Other. Finally, the English School Census considers a classification similar to the one of the Census, except there is not a specific subcategory for Arab and Chinese is included as a subcategory within Other instead of within Asian.

To use a classification that is fully compatible with the three aforementioned sources, this empirical exercise uses the classification: White, Mixed/multiple ethnic groups, Asian/Asian British, Black/African/Caribbean/Black British, Chinese and Other.

3.3. Models

In order to produce an estimate for the table of interest using the GSPREE, only a proxy and a survey estimate of the table of interest, and the corresponding set of row and column margins are required. However, as it was described in the previous section, for the age group *5-15* two different sources of proxy information are available in this case. To study how to better use these sources, the following three models, with increasing level of complexity, were considered:

- **Model 1:** uses the 2011 Census as the only source of proxy information. Both the proxy and the survey tables are aggregated at the LA versus Ethnicity level, without considering the age group.
- **Model 2:** uses the 2011 Census as the only source of proxy information. The proxy and the survey tables are split by age group and an independent fitting is performed for each one of the three age groups mentioned above. The three estimates of the population counts are summed up to produce one estimate of the target table. The table of estimates after aggregating by age group is then benchmarked to the column and row margins.
- **Model 3:** uses both 2011 Census and the English School Census as sources of proxy information. In analogy to Model 2, an independent fitting is performed for each age group. Each one of those fittings goes through two steps:
    - **Step 1:** construction of an auxiliary structure that is a convex linear combination of the two available structures. The coefficient of the 2011 Census structure in the convex combination, denoted by δ, is found via numerical optimisation, as the value that minimizes the deviance of the fitting of the model defined by equation (3) for that particular age group.
    - **Step 2**: Estimation of the table of interest for that age group, using the survey data, the auxiliary structure built in step 1 and the GSPREE.

    As for Model 2, the three estimates of the population counts are summed up to produce one estimate of the population table. The benchmark of rows and columns is only applied over this last table estimate.

Notice that Model 2 is a particular case of Model 3 where $\delta = 1$. Therefore, a Likelihood Ratio Test (LRT) can be used as a diagnostic tool to compare their fitting. Given that Model 1 does not fit the three age groups independently, it is not possible to consider it nested in either of the other two models. However, an approximate Likelihood Ratio Test (LRT) between Model 1 and Model 2 is performed by approximating the former as a particular case of Model 2 with the same β in all age groups.

4. Results

Possibly due to the sampling design of the APS, no appreciable differences were observed in the within-LA distributions of ethnicity calculated from the sampling counts, or from direct estimates of the population counts. Therefore, sampling counts were used as input for the models. Poisson and Multinomial Likelihoods were used for the estimation of β, the latter being closer to a Simple Random Sampling design stratified by LA. The estimates of β obtained under the two distributions differ only at the third decimal point. Here we present only the results for the Poisson MLE.

The ethnicity variable has a very unequal distribution in the population. Aggregating the data of the 2011 Census for the areas under consideration, the category *White* is dominant with 85.42% of individuals, followed by *Asian* (7.10%), *Black* (3.48%), *Mixed* (2.25%), *Other* (1.03%) and finally *Chinese* (0.72%). How different LAs deviate from that global distribution can be observed in Figure 1. Notice that for categories *Asian* and *Black* it is possible to find some areas with proportions considerably higher than the global proportion. Moreover, notice that in such areas, non-white individuals are predominantly from one of the two above mentioned categories instead of evenly distributed. Meanwhile, for the categories *Mixed*, *Chinese* and *Other,* the proportions are uniformly low in all local authorities.

The actual sampling fractions of the APS in some LAs can be quite small. An implicit sampling fraction was calculated by dividing the observed sample size by the corresponding projected population total in each LA. This varies between 0.05% and 2.5%, with an average of 0.8%.

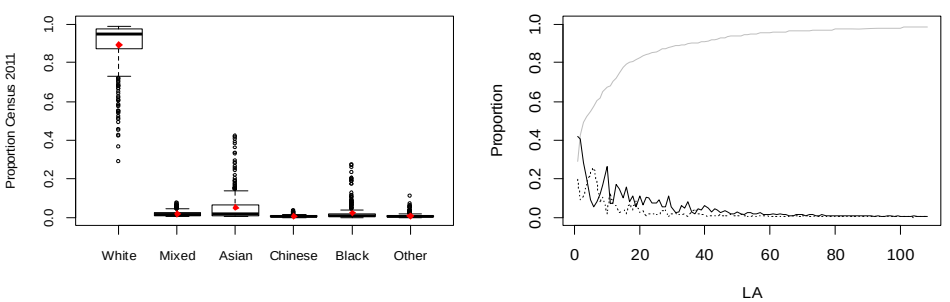

Figure 1. Distribution of Ethnicity by LA in the 2011 Census. (a) Boxplot proportions in each category by LA. Red diamond: mean. (b) Detail of the largest categories. Lines: *White*: continuous grey. *Asian*: dotted black. *Black*: continuous black. After sorting the LAs according to the proportion of *White,* one of each three LA was included in the plot

Given the low proportions of individuals belonging to categories such as *Chinese* or *Mixed*, as well as the small sample sizes observed in most LAs, some of the cells of the observed survey composition have zero observations, making it impossible to calculate their interaction terms directly. In principle, the presence of some sample zero cells does not necessarily cause a problem in terms of the estimation of the parameter of the GSPREE, when a fully model-based approach such as the one described in Section 2.2 is employed for this task. However, this means that the plausibility of the structural equation for this particular variable cannot be empirically checked using a scatterplot between interactions of the survey and proxy compositions. For illustration purposes, the pairs of interactions at the LA level for the 2001 Census and the 2011 Census in England are shown in Figure 2. Notice how, except for the category *Other*, interaction terms from the same composition 10 years before can still work fairly well as linear predictors. Unless period 2011-2013 behaves in a substantially different way than 2001-2011, it could be expected for the structural equation to hold at least approximately.

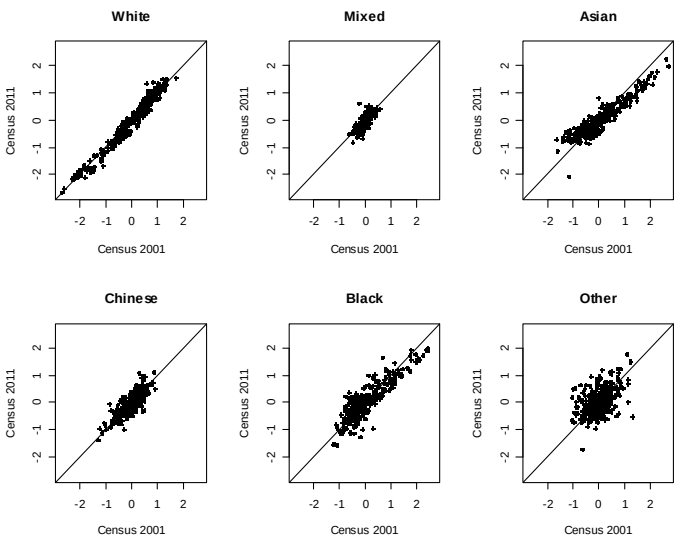

Figure 2. Interaction terms of the composition LA by Ethnicity , Census 2001 and 2011. Line: Y=X

The three fixed effects models stated in Section 3.4 were fitted to the data, using both the SPREE and GSPREE. In each case the estimated coefficient of the GSPREE estimator, β, is very close to 1, i.e. this estimator and the SPREE almost coincide. We therefore omit the results for the latter. The main results for the GSPREE are presented in Table 1. The last three columns contain the information to perform a LRT comparing the models in increasing order of complexity, as explained in section 3.4. In all cases there is evidence indicating that the more complex model leads to a slightly better fit. However, the estimates of the within-area distribution obtained using the three models are very close. For illustration, scatterplots between those obtained with Models 1 and 3 are presented in Figure 3.

Table 1. Fitting results. Fixed effects models

Model	Age group	Estimated Coefficients	Deviance	Difference Deviance	Crit. value 5% Sig.
1) LA x Ethnicity Census 2011	-	b=1.007	4639.41 7358.58*	-	
2) LA x Ethnicity|Age Census 2011	0-4	b=0.990	1714.81	2) vs 1) 15.06	5.991
	5-15	b=0.963	2441.20		
	16 or more	b=1.010	3187.51		
3) LA x Ethnicity|Age Census 2011 & School Census	0-4	b=0.974; d=0.780	1703.21	3) vs 2) 56.51	7.815
	5-15	b=0.958; d=0.677	2414.87		
	16 or more	b=0.998; d=0.913	3168.93		

* Deviance of a Model 2 with b=1.007 in each age group.

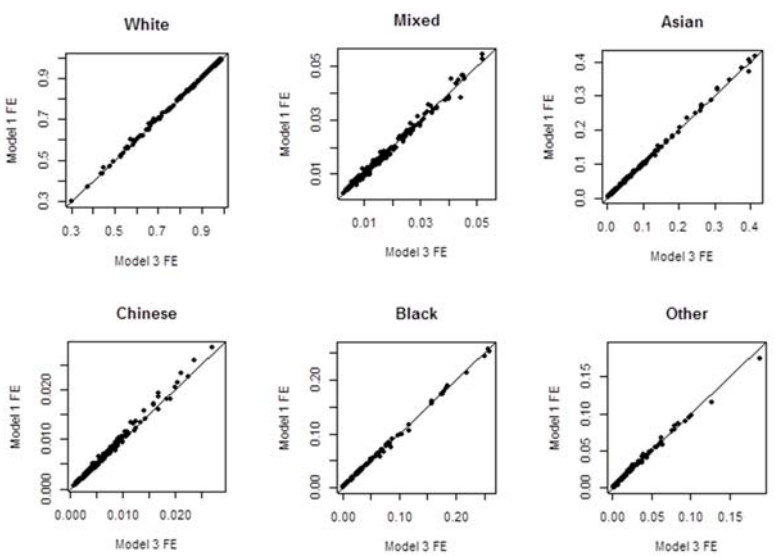

Figure 3. Estimates of the within-area distribution. Fixed effects GSPREE. Model 1 and Model 3. Line: Y=X

On the other hand, it is expected that if the sample sizes are big enough, an estimator based on a mixed effects model would be less biased than its fixed effects counterpart. For each of the models, we attempted to calculate the mixed effects version of the GSPREE proposed in Zhang and Chambers (2004) but it was impossible to achieve convergence in the estimation of the variance-covariance matrix of the random effects, possibly due to the generally low sampling fractions. As an alternative, we fitted a fully parameterised mixed effect GSPREE, with a log-link and a Poisson sampling distribution, similar to the one described by equation (3) but including cell-level independent random effects with category-specific variances, as an extension of Model 1. Only for three of the six

categories positive estimates of the variance components were found. The estimates are 0.029, 0.014 and 0.101, for *White*, *Mixed* and *Asian* respectively. For the other categories, the corresponding variance component estimates were set to zero. The issue of negative variance component estimates for some but not all the categories will be discussed further in Section 5.

A set of scatterplots comparing the estimates obtained under the mixed effects version of the GSPREE estimator for Model 1 and the fixed effects version for Model 3 are presented in Figure 4. Differences in the estimated proportions are observed, especially for the categories Mixed, Asian and Chinese. Notice that even though for the last three categories the variance component estimate was zero, the two estimators do not coincide due to the IPF. Figure 4 does not suggest a bad performance of Model 3 in terms of bias, when compared to the mixed effect estimator.

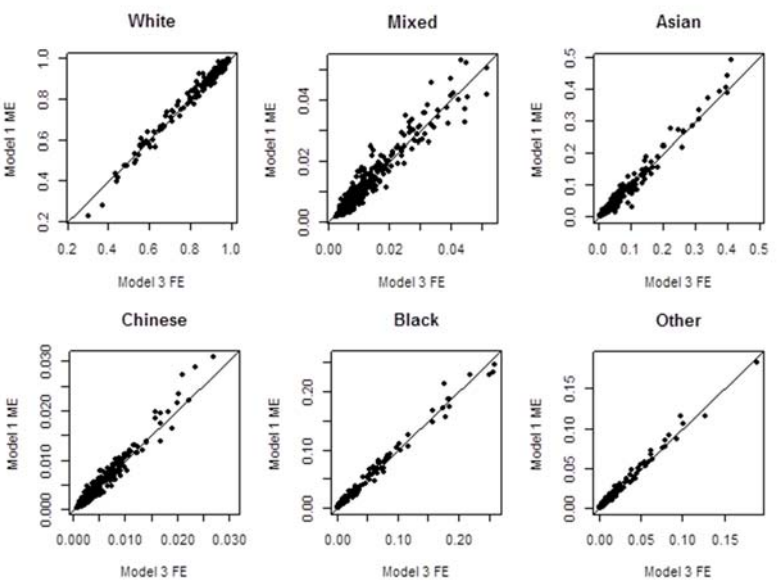

Figure 4. Estimates of the within-area distribution. Mixed effects GSPREE for Model 1, Fixed effects GSPREE for Model 3. Line: Y=X

4.1. Mean Square Error (MSE) evaluation

To assess the performance of the different estimators in terms of their Finite Population Mean Square Error (FP-MSE), a semi-parametric bootstrap approach was applied. The bootstrap samples were randomly generated from a plausible population composition, instead of randomly selected from a fixed synthetic population. Both approaches should perform similarly given that the implied sampling fractions of the APS are negligible but the former is considerably quicker. Two sampling designs were used: Multinomial, assuming the same observed sample size in each area as fixed, and Poisson sampling with random sample size. As counts by age are required to fit Models 2 and 3, independent

samples were generated for each age group, and the aggregate of the three samples was used to fit the Fixed and Mixed effects estimators under Model 1.

The initial idea was to generate the population composition under a mixed effects model split by age. However, as mentioned before, it was impossible to obtain positive variance estimates for those models, and even in the case of Model 1, only three of the six categories have a positive variance component estimate. Using such variance components estimates could lead to an overly optimistic scenario for the GSPREE because of a lack of heterogeneity.

An alternative set of variance components was obtained from the two proxy tables, School Census 2012-2013 and Census 2011, by considering the School Census as a big sample from the true population in the age group 5-15 and using the methodology of the mixed effects GSPREE estimator. The estimated variance components are 0 for *White*, 0.02 for *Mixed*, 0.05 for *Asian*, 0.12 for *Chinese*, 0 for *Black* and 0.79 for *Other*. To allow for extra heterogeneity in all the categories, the two zero estimates were replaced by the minimum positive estimated value, 0.02. These estimates were used in all age groups, to generate the population composition from which the bootstrap samples are generated.

Despite the two zero estimates, we have found the set of variance components estimated using the two auxiliary sources more plausible than the one obtained from the sampling data under Model 1 in section 4.1, when taking into consideration the category specific heterogeneity observed in Figure 2. This could be seen as evidence against the performance of the mixed effects estimator presented in the previous section. It is possible that, even under Model 1, a synthetic estimator needs to be used given the small sample sizes in the cells of the survey composition.

The results in terms of FP-Bias and FP-MSE obtained under Poisson or Multinomial sampling were very similar, possibly due to the impact of the benchmarking on reducing the variability associated to the random area sample size in the case of the Poisson sampling. We will therefore omit one set of the results. The results for the Multinomial sampling are presented in Table 2 and Figures 5, 6 and 7.

Table 2. Average FP-Bias and Square root FP-MSE

		Ethnicity					
Measure	Model	White	Mixed	Asian	Chinese	Black	Other
Average FP-Bias	Model 1 FE	-0.00157	0.00013	0.00162	0.00028	-0.0001	-0.00036
	Model 2 FE	-0.0008	-0.00001	0.00119	0.00022	-0.0002	-0.0004
	Model 3 FE	-0.00156	0.00011	0.00164	0.00029	-0.00009	-0.0004
	Model 1 ME	-0.00158	0.00018	0.0016	0.00031	-0.00013	-0.00039
Average Square Root FP-MSE	Model 1 FE	0.00948	0.00195	0.00691	0.00158	0.00354	0.00717
	Model 2 FE	0.01177	0.00251	0.00895	0.00161	0.00418	0.00651
	Model 3 FE	0.00951	0.00189	0.007	0.00162	0.00357	0.00717
	Model 1 ME	0.00974	0.00187	0.00709	0.00167	0.00355	0.00716

Overall, there is no estimator that performs substantially better than the others, either in terms of FP-Bias or FP-MSE. Even though the average bias for each category is close to zero, according to Table 2, for specific areas there is bias in the estimation of the within-area distribution in all the fixed effects estimators, as it can be seen from Figure 5. The mixed effects estimator under Model 1 seems unable to correct this bias, given that the estimates with bigger biases are those for LAs with small sampling fractions. See Figure 6.

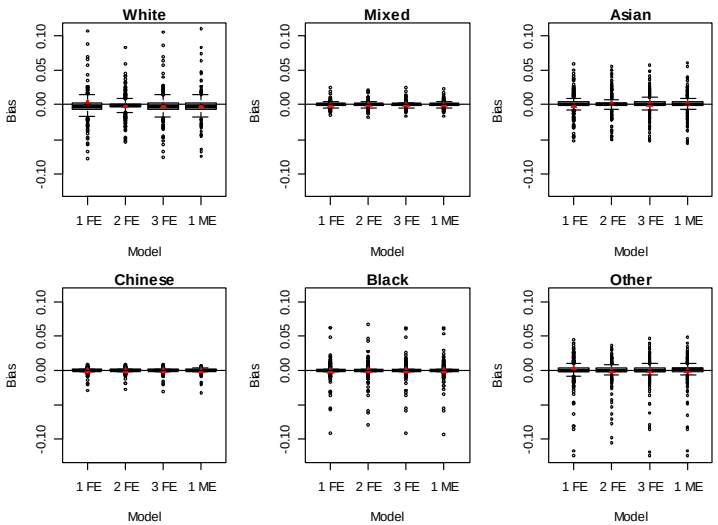

Figure 5. FP-Bias with respect to the simulated population composition. Red triangle: Mean

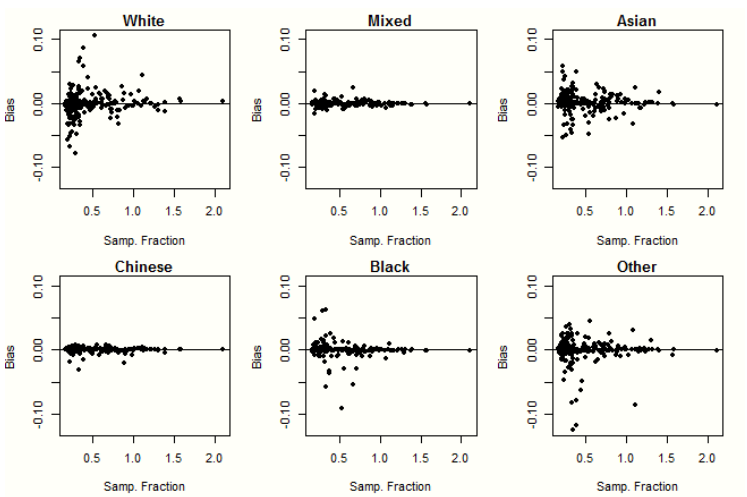

Figure 6. Implicit sampling fraction Vs. FP-Bias of the mixed effects estimator (Model 1)

5. Discussion

In this paper, we present a feasibility study to produce Small Area estimates of the within-area distribution of Ethnicity by LA in England, using the GSPREE. It is the first time this approach has been attempted for this type of problem in the UK. Unlike other demographic and socio-economic characteristics, Ethnicity is a variable for which there is no clear set of covariates identified in the literature, which could be used as a predictor. In fact, unless a proxy is involved, it seems difficult to expect good performance of a Small Area Estimator in this context. Structure Preserving Estimators can be used, given that proxy compositions can be obtained either from the last population census or from other sources, such as the School Census.

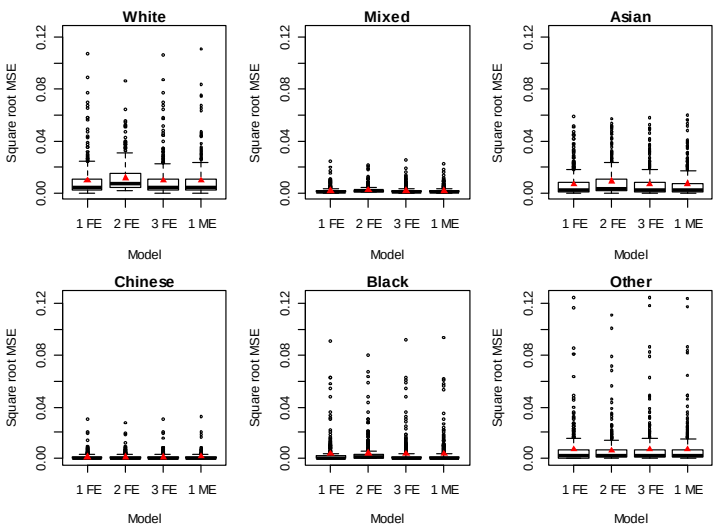

Figure 7. Square root FP-MSE with respect to the simulated population composition. Red triangle: Mean

In this work, we formulated three alternative models to produce the desired estimates with the GSPREE. However, in terms of Bias and FP-MSE, no substantial improvement was obtained by using more complex models or different sources of information. Moreover, given the small sample sizes available from the APS, synthetic estimates seem the only possible alternative in this case.

Notice that the lack of sample size to fit a mixed effects model is not a problem only of this application but rather one which all applications of SAE face sooner or later, if the aim is to produce estimates at increasingly lower levels of aggregation. In this sense, work to improve the synthetic predictor is of highest priority. Currently, we are working on a more flexible version of the fixed effects GSPREE and we expect to be able to evaluate it against the other estimators included in this paper, in the near future.

When it comes to mixed effects modelling, a particular problem we encountered with these data is that the variance component estimate can be

negative for some but not all the categories, when the model allows for category-specific variance components. A possible remedy is to impose a common variance component. However, further study is needed in order to determine whether this or another random effects modelling strategy can be suitable.

Evaluation of the estimators in terms of their Bias and FP-MSE is also a topic for future work. The conclusions and quality of the evaluation is closely related to the plausibility of the characteristics of the artificial finite population, or as in our case, of the artificial population composition, from which the bootstrap samples are extracted. Additional work is still necessary in this area in order to formulate alternative scenarios that can be used to select a model, as well as to increase our knowledge on the performance of the proposed estimators.

REFERENCES

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

BERG, E. J., FULLER,W. A., (2014). Small Area Prediction of Proportions with Applications to the Canadian Labour Force Survey. Journal of Survey Statistics and Methodology, 2 (3), 227–56.

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

LIANG, K-Y., ZEGER, S. L., (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73, 13–22.

MOLINA, I., AYOUB S., LOMBARDIA, M. J., (2007). Small Area Estimates of Labour Force Participation under a Multinomial Logit Mixed Model. Journal of the Royal Statistical Society: Series A (Statistics in Society), 170 (4), 975–1000.

NOBLE, A., HASETT, S., ARNOLD, G., (2002). Small Area Estimation via Generalized Linear Models. Journal of Official Statistics, 18(1):45–68.

PFEFFERMANN, D., (2013). New Important Developments in Small Area Estimation. Statistical Science, 28 (1), 40–68.

PURCELL, N., KISH, L., (1980). Postcensal estimates for local areas (or domains). International Statistical Review, 48(1), 3–18.

RAO, J. N. K., (2003). Small Area Estimation. John Wiley & Sons.

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

UPTON, G., COOK, I., (2008). *A Dictionary of Statistics*. Oxford University Press.

ZHANG, L. C., CHAMBERS, R., (2004). Small area estimates for cross-classifications, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(2), 479–496.