Geospatial machine learning for intercensal small-area population estimates

Evaluation and comparison with traditional methods

Key Messages of Paper

Purpose

- Assess geospatial machine learning models to produce intercensal small area population estimates (SAPE) within Lower Layer Super Output Areas (LSOA), and compare them with ratio change (the current method for official estimates) and benchmarking (an alternative method).
- Compare the performance of models trained at the LSOA level with models trained at a 100m x 100m grid.
- Recommend approaches for SAPE methods that might address issues in the official estimates related to accumulating error across the intercensal period, and extreme error in a small number of problematic areas.
- Seek assurance on our assessment and recommendations from MARP.

Recommendation

- Do not put geospatial machine learning models for LSOA SAPE into production at this time.
- Pursue user engagement to develop:
    - Prioritised SAPE use cases, and user quality requirements for SAPE with which to more robustly define and prioritise the problems with current estimates, and establish quality targets/success criteria for experimental estimates.
- Commission further research to explore (in priority order, and each contingent on the previous point having a successful outcome):
    - Whether ratio change and LSOA geospatial models can be combined to improve upon official intercensal estimates for Census geographies.
    - Whether the LSOA models can be improved further.
    - Whether underestimation issues with the grid models can be solved, and whether their intercensal drift can be quantified.
    - Methods for quantifying uncertainty in SAPE derived from geospatial models, and producing age by sex breakdowns.
    - Whether grid-based geospatial models could also be a general-purpose approach for producing estimates for non-Census geographies e.g. National Parks.

Key Asks of MARP

- Does the panel endorse our recommendation?
- Are there any additional types of evidence the panel would like to see in future work when comparing the quality of experimental intercensal SAPE with official estimates?
- We are particularly interested in feedback on our approach of standardising comparisons between methods by constraining competing small area estimates to trusted LA estimates, and in doing so also omitting the measurement of bias introduced by the LA estimation method.

Executive Summary

The Office for National Statistics (ONS) is exploring possible methodological improvements for intercensal small area population estimates (SAPE) as part of a wider quality improvement plan. We have assessed supervised machine learning (ML) models using geospatially referenced covariate data to produce SAPE for Lower Layer Super Output Areas (LSOAs), and compared them with SAPE produced by the official method and a potential alternative method.

Official LSOA estimates are currently produced using ratio change. They are characterised by high accuracy at the start of the intercensal decade with declining accuracy across the decade, where a small but increasing number of estimates for problematic LSOAs become unusable over time.

We have developed two types of geospatial ML models – LSOA models that predict directly at the LSOA level, and grid models that predict into a 100m x 100m grid (which is then aggregated to the LSOA level).

ML models outperform ratio change 10 years after a Census, but ratio change achieves the most accurate SAPE for most of the intercensal period. Grid models appear to outperform LSOA models, particularly in terms of the number and degree of problematic LSOAs with extremely high error, although they systematically underestimate more than overestimate (particularly for more populous LSOAs).

All methods have some similarity in their problematic areas – particularly LSOAs in London. However, ratio change tends to underperform in urban areas, and geospatial models tend to underperform in smaller rural areas (particularly the LSOA models).

More generally, the strengths and weaknesses of ratio change and geospatial models are in some ways complementary e.g. where the performance of ratio change declines substantially across the intercensal decade, geospatial models are relatively stable; where ratio change has more extreme errors in a small number of LSOAs, geospatial models achieve more consistent performance across a wider range of LSOAs with fewer extreme errors. If an approach can be developed to leverage the complementary strengths of both approaches it may be possible to produce SAPE for Census statistical geographies that have higher accuracy across the intercensal decade, with a more gradual decline in quality over the decade, and fewer and less extreme errors in problematic LSOAs.

Additionally, grid-based geospatial models may also be a good candidate for producing SAPE for non-Census geographies (e.g. National Parks) due to the small and uniform size of grid cells allowing for more accurate aggregation to bespoke geographies than the current approach of aggregating Output Areas, and so potentially more accurate estimates.

Future work would be better supported by more robustly defined problem statements and success criteria for SAPE, informed by user engagement.

Geospatial machine learning for intercensal small-area population estimates

1. Introduction

The Office for National Statistics (ONS) is implementing a plan to improve the quality of population, economic, and social statistics (1). As part of this plan, the Methodology & Quality Directorate (MQD) has been tasked with investigating potential methodological improvements for intercensal small-area population estimates (SAPE) to support conclusions on a future approach for England & Wales. We have been asked to initially focus on estimating the total intercensal population within Lower Layer Super Output Areas (LSOAs) at mid-year per internal prioritisation.

ONS currently publishes population estimates for statistical geographies from the decennial census, and annual intercensal mid-year estimates (MYE) using a cohort component method for Local Authorities (LA), ratio change for Super Output Areas (SOAs), and apportionment for Output Areas (OAs) (2; 3). OA estimates are also aggregated to produce estimates in development for some non-Census geographies e.g. National Parks (4). LA and SOA MYE are badged as National Statistics but OA MYE are only provided as supporting information. Ratio change produces LSOA SAPE by rolling forward Census estimates using annual population change scaling factors calculated from a single administrative source of population stocks. They are characterised by very low error at the start of the intercensal period, but accumulate compounding error across the decade until they can be rebased after the subsequent Census. Ratio change SAPE remain usable for most LSOAs across the intercensal period despite this compounding error, but estimates for a minority of LSOAs can become unusable towards the end of the decade (4).

MQD has previously investigated benchmarking as a potential alternative to ratio change, as well as potential improvements to ratio change itself. Benchmarking is an approach where the small-area population distribution from a single administrative source is used to directly disaggregate known estimates from a higher geographical level. Rather than introducing accumulating error at the LSOA level each year like ratio change, the error profile of benchmarking is invariant as a function of the time elapsed since the last census and instead only covaries over time with the quality of the admin data sourced used as an input. Although, note that all SAPE constrained to LA MYE also incorporate error introduced at the LA level by the cohort component method (regardless of small area method).

MQD is now developing geospatial machine learning (ML) models in collaboration with Data Science & Flexible Analysis (formerly the Data Science Campus) as a potential alternative to ratio change at the LSOA level. Geospatial ML models employ a supervised learning approach where observed population estimates are used to train models to predict unobserved estimates using a range of spatially referenced covariate data. This includes environmental remote sensing data like night-time light intensity and air quality, high resolution land use and building characteristics, and spatially referenced administrative and survey data where relevant e.g. admin-based population counts for statistical geographies.

Supervised ML models with geospatially referenced data are increasingly being used to produce SAPE in place of more traditionally-used linear mixed-effects models in areas where population data collection is spatially incomplete or where population estimates aren't available at small-area resolution (5). In our use case the observed LSOA estimates come from the Census, and the "unobserved" LSOA estimates are those at mid-year i.e. we implement our models as if SAPE are available once per decade but are otherwise temporally incomplete. Geospatial ML models have the potential to address some of the shortcomings of ratio change by using a range of high-resolution covariate data that may more closely and robustly covary with small area population size than the single administrative source used by ratio change.

In this paper we present progress on the development of several geospatial ML models for producing intercensal SAPE, compare them with previous work that explored benchmarking and ratio change, and make recommendations for future work.

2. Background

Colleagues in MQD have investigated benchmarking and potential improvements to ratio change in parallel to our geospatial work (recently reviewed by the ONS Methods and Research Assurance Group but not published externally). This work concluded that (i) the current implementation of ratio change used to produce intercensal SAPE is preferable to several alternative implementations using different data inputs, and (ii) benchmarking is slightly better than ratio change towards the end of the intercensal period due to how ratio change accumulates error over time. Subsequent work (partially presented in this paper) has shown that the performance of benchmarking is relatively stable over the intercensal period compared to ratio change but varies with input data quality.

In previous work we reported on the development of proof-of-concept geospatial models for predicting SAPE for the Census household population only i.e. excluding residents of communal establishments (6; 7). Overall, our previous results showed potential in both LSOA and grid-based models but highlighted challenges including the robustness of model performance metrics calculated with a single train and test data split, how models handled spatial variation in household occupancy rates (e.g. number of people per dwelling), high values for the maximum bias achieved, and understanding the temporal stability of the models.

3. Data

Table 1 *Data used to train and optimise geospatial models, and to assess all methods.*

ID	Source	Supplier	Spatial resolution	Reference date/period	Release frequency
1	**Visible Infrared Imaging Radiometer Suite (VIIRS) night-time light intensity**	Earth Observation Group	300m x 500m grid (approximately)	2011-2021	Annual
2	**UK air pollution**	Department for Environment, Food, and Rural Affairs	1x1 km	2011 & 2021	Annual
3	**Domestic gas and electricity consumption**	Department for Energy Security & Net Zero	LSOA	2011-2021	Annual
4	**Rural-urban area classification**	Office for National Statistics	OA/LSOA	Census day 2011; Census day 2021	Decadal
5	**Land cover maps**	UK Centre for Ecology & Hydrology	10m resolution raster	2021	Annual
6	**AddressBase land, property, and street data**	GeoPlace	Individual properties	2011-2021	Approx. every 6 weeks
7	**Ordnance Survey buildings**	Ordnance Survey	Individual properties	2011 & 2021	Daily
8	**Ordnance Survey highways**	Ordnance Survey	Individual road	2011 & 2021	Monthly
9	**Ordnance Survey Land Features**	Ordnance Survey	Individual land parcel	2021	Daily
10	**Statistical population dataset v5.1 (SPD)**	Office for National Statistics	LSOA & by individual properties	2021	Annual
11	**Admin-Based Living Arrangements Dataset v5.1 (ABLAD)**	Office for National Statistics	Individual properties	2021	Annual
12	**Patient Register (PR)**	Department for Health & Social Care	LSOA	2011	Discontinued
13	**Personal Demographic Service (PDS)**	Department for Health & Social Care	LSOA & by individual properties	2021	Annual
14	**Census**	Office for National Statistics	LSOA (2011, 2021) and resident microdata (2021)	2011 & 2021	Decadal
15	**Census-based mid-year estimates 2021**	Office for National Statistics	LSOA and LA	2021	Annual

4. Methods

4.1 Method assessment

We assess the quality of SAPE using absolute relative bias (ARB), calculated per LSOA as the absolute value of:

$$100 * \left(\frac{(LSOA\ estimate - LSOA\ true\ value)}{LSOA\ true\ value}\right)$$

Where in our case the "true values" are estimates taken from Census-based LSOA MYE 2021 for our main 2021 results, or Census 2011 and 2021 LSOA estimates when we calculate temporal scaling factors for intercensal model drift. However, note that in some analyses we present the signed value of relative bias to assess over- and under-estimation.

Correlations between SAPE are calculated using Spearman's rank correlation as the estimates are not normally distributed.

To standardise comparisons across methods we constrain all LSOA estimates to sum to the LA estimates in the assessment data for the relevant reference period. This means the ARB measures we report exclude any error introduced by the choice of LA constraint in the assessment year, and thus describe how well each method reproduces the true population distribution at the LSOA level i.e. they are not representative of overall bias in published estimates.

We also classify ranges of ARB into categories of increasingly large bias to cluster LSOAs into groups for investigation: (i) <10, (ii) 10 to <25, (iii) 25 to <50, (iv) 50 to <100, and (v) >100.

We have defined all categories except the first (<10) as problem areas i.e. areas where the ARB is too high. However, note that this is a preliminary working definition that needs to be tested with user input.

4.2 Machine learning methods

4.2.1 Algorithms

We have trained tree-based regression models using Random Forest (RF) (8) and XGBoost (XGB) (9). Tree-based methods are advantageous for use cases with many predictive features like geospatial SAPE as they can automatically fit interactions between features without manual specification. However, neither algorithm can extrapolate predictions meaningfully beyond observed values in the training data. We model population density rather than population size, changes of which over a decadal timescale are primarily characterised by shifts in distribution rather than range.

While we have implemented geospatial models using both RF and XGB, please note that for brevity we only present geospatial results from XGB in this paper as the results from both are qualitatively similar.. Summary of traditional methods can be found in Annex section A2.

4.2.2 The bottom-up approach

Our geospatial models are based on a hybrid of the WorldPop bottom-up and top-down approaches for population size estimation using tree-based algorithms (10; 11; 12). In short, the bottom-up approach is where models are trained on known estimates for a sample of areas at a lower-level geography to predict new estimates for unsampled lower-level areas, all of which can then be aggregated to produce new estimates at a higher-level geography. In contrast, the top-down approach is where models are trained on known estimates at a higher-level of geography to predict new estimates at a lower-level geography that are constrained to sum to the known estimates at the higher-level. In our approach we train models on known estimates at a lower-level geography at one point in time (Census) to predict new estimates at the same lower-level geography at a later point in time (intercensal mid-year), and constrain them to known estimates at a higher-level geography (LA). We have developed models at two lower-level geographies, (i) LSOA models trained on Census LSOA estimates that directly predict the population density of LSOAs, and (ii) grid models trained on Census microdata that predict population density within a 100m x 100m resolution grid that we then aggregate to the LSOA level. In both cases we use land area calculations to convert model predictions of population density into nominal population values before constraining to known LA estimates.

In both LSOA and grid models we predict the natural log of population density (population per hectare). We predict population density to improve model generalisation by making the target variable more comparable across all areas; LSOA size varies substantially and density is independent of LSOA size. We use density in the grid model primarily for convenience as grid cells have a standardised area of 1 hectare and so population size and density are equivalent (our method does not yet distinguish between habitable and uninhabitable area for the purposes of density calculations). The natural log is used to transform population density from right-skewed to a gaussian distribution (a positive constant of 1 is added to all grid cells to avoid log(0) errors). Model predictions of log population density are then transformed back to their original scale and multiplied by land area to produce nominal LSOA population counts. LSOA population counts are used to create a weighting layer to disaggregate LA MYE into their constituent LSOAs by taking the model-based population count for each LSOA as a proportion of the sum of all model-based population counts for all LSOAs within a LA. This yields model-based LSOA estimates that are constrained to sum to the LA MYE total population. Note that some of our model drift results are referenced to Census day 2011 and 2021 rather than mid-year and so we instead constrain them to the Census LA estimates for the relevant years (see section 4.5).

4.3 Model training and testing framework

4.3.1 Training and testing

We train and test our main models using *k*-fold nested cross-validation that contains an outer loop for training and testing models, and an inner loop for feature selection and hyperparameter optimisation (*k*=5) (13). Using cross-validation gives more robust and unbiased estimates compared to a single train-test split (14). Areas are split into folds using a simple random sample of LSOAs for both our full LSOA and grid models, but split by region for our reduced drift models as some LSOAs were split, merged, or otherwise altered across 2011-2021. Where grid cells are intersected by an LSOA boundary the cells are assigned to the LSOA that makes up the largest proportion of the cell by land area. Note that we first mask cells that do not contain any residential addresses (properties that are or could be lived in) recorded in AddressBase and assign them zero population without modelling; this both makes the grid models more computationally tractable and avoids issues caused by zero-inflated data.

All models are trained to minimise Mean Absolute Error (MAE) in log population density predictions as this most closely resembles the quality metrics used to assess model-based SAPE.

ARB metrics are calculated on a final set of estimates for all LSOAs collated from all held-out validation fold predictions, and so are an average of results across several different models. This means that our geospatial results should be interpreted as an assessment of a model fitting process rather than of any specific model produced from that process.

4.3.2 Feature selection

We implement reverse feature elimination with cross-validation (RFECV). In RFECV we select the optimal number of features using cross-validation by iteratively removing features in reverse order of feature importance (see section 4.4 for importance method) until only one feature is left, and then select the number that yields lowest MAE across all cross-validation folds (note that the number of features is a hyperparameter but we do not tune it like other hyperparameters as per section 4.3.3). The model is then retrained on all folds and the optimal feature set is selected via reverse feature elimination of the least important features until the optimal number of features is reached.

4.3.3 Hyperparameter tuning

We implement model hyperparameter tuning using Bayesian optimisation to search a pre-defined multidimensional hyperparameter space to select the values that minimise MAE in our nested cross-validation framework (15).

4.4 Model explainability

We interpret our models by quantifying the importance of the features used by the models to make predictions. To ease interpretation, we categorise model features into "themes" and sum feature importance scores within themes to produce theme importance scores. The theme categories are: (i) Administrative population, (ii) Building/Property, (iii) Energy, (iv) Environmental, and (v) Infrastructure and area classification. Please see Annex Table A1 for more specific feature information).

We measure feature importance in our XGB models using Gain, which measures the average reduction in error across all model splits that use the feature in question. Features with a higher Gain score affect model accuracy more than features with a lower score. Reported Gain scores are averages taken from the *k*-fold cross-validation models.

4.5 Temporal stability (2011-2021)

We have investigated the performance of geospatial models in 2011 and 2021 to enable more robust comparison with ratio change in 2021 by using geospatial models with the same Census lag as ratio change.

4.5.1 Geospatial drift models

Not all the features we use in our full 2021 models are available in 2011, and so we train "reduced" LSOA models in 2011 on a subset of features that are available in both years, and a small number of building attribute features available after 2011 where these are available at a later date and are unlikely to change over short timescales e.g. building height. While not all the features are available in both years, we believe enough of the most important features are available to make the results broadly generalisable to our full LSOA models (Annex Table A1). However, we have not been able to train grid-based drift models as too many features are unavailable at the grid level for 2011. We have been able to produce results for both XGB and RF reduced LSOA models covering:

- "0-year lag" models trained and tested on 2011 data (Census day)
- "10-year lag" models trained on 2011 data (Census day) and tested on 2021 data (Census day when investigating performance across 2011-2021; mid-year when producing 2021 mid-year estimates for main comparison with full models and traditional methods)

4.5.2 Quantifying model drift

We quantify model drift as a scaling factor of the median ARB between two reference dates calculated as the median ARB value of the later model divided by the ARB value of the earlier model. This measure incorporates changes in covariate data quality between two timepoints, changes in the distributions within the covariate and target data, and temporal change in the fitness of the modelled function describing the relationship between the features and the target.

4.5.3 Sensitivity of drift to features

To investigate the sensitivity of drift factors to feature type, we have trained a series of alternative drift models, each excluding a different theme of features as defined in section 4.4 (without performing automatic feature selection as in our main models). We compare the drift factors for each alternative feature set to assess whether the exclusion of a particular category of features makes models drift more or less.

5. Results

We calculate ARB at the LSOA level for all methods and present these figures both overall and broken down in various ways to investigate patterns of bias by method.

It is important to consider the Census lag when comparing between methods. Ratio change and the geospatial drift model are assessed 10 years after they were initiated/trained using Census 2011 data (10-year lag). In contrast, the other geospatial models were trained using data from the year in which they are assessed, and benchmarking only uses data from the assessment year (0-year lag). This means that ratio change and the drift model incorporate error due to the time elapsed between 2011 and 2021 that benchmarking and the other geospatial models do not, and so comparisons between methods with different Census lags should be made with caution.

5.1 Headline results

5.1.1 Bias at mid-year 2021

Geospatial ML models (both LSOA and grid) have lower ARB across the board in 2021 than traditional methods when comparing within 0-year lag methods (XGB LSOA and XGB grid vs. benchmarking) and within 10-year lag methods (XGB drift vs. ratio change) (Table 2, Figure 1).

The geospatial ML models have substantially lower maximum ARB values than the traditional methods, particularly the grid model which is the only method to achieve a sub-100 maximum ARB. While the grid model has a lower maximum ARB compared to the LSOA model, the two methods otherwise have similar ARB quartiles. The drift model has higher ARB than the full LSOA model as the drift model has been allowed to drift across the intercensal period and uses a reduced feature set. The relatively lower quality of covariate data in 2011 likely also contributes to a less performant model.

The overwhelming majority of LSOAs are in the <10 ARB category for all methods (Table 3). There are fewer LSOAs in each increasingly higher ARB group for all methods, and the geospatial methods have relatively fewer LSOAS in the highest groups than the traditional methods – the grid model has no LSOAs in the highest ARB group at all.

Table 2 *Absolute relative bias of the LSOA estimates (quartiles and max) for all methods as measured at mid-year 2021 against Census-based mid-year estimates. Lowest values for each column in bold.*

Method	Version	Census lag (years)	P_{25} ARB	P_{50} ARB	P_{75} ARB	Max ARB
Ratio change	PR 2011-20, PDS 2021	10	1.53	3.30	6.04	258.52
Benchmarking	PDS ‘21	0	1.46	3.19	5.87	252.28
Geospatial ML	$LSOA_{XGB}$	0	1.22	2.57	**4.64**	143.1
	$Grid_{XGB}$	0	**1.17**	**2.52**	4.67	**85.52**
	$Drift_{XGB}$	10	1.40	3.04	5.60	182.47

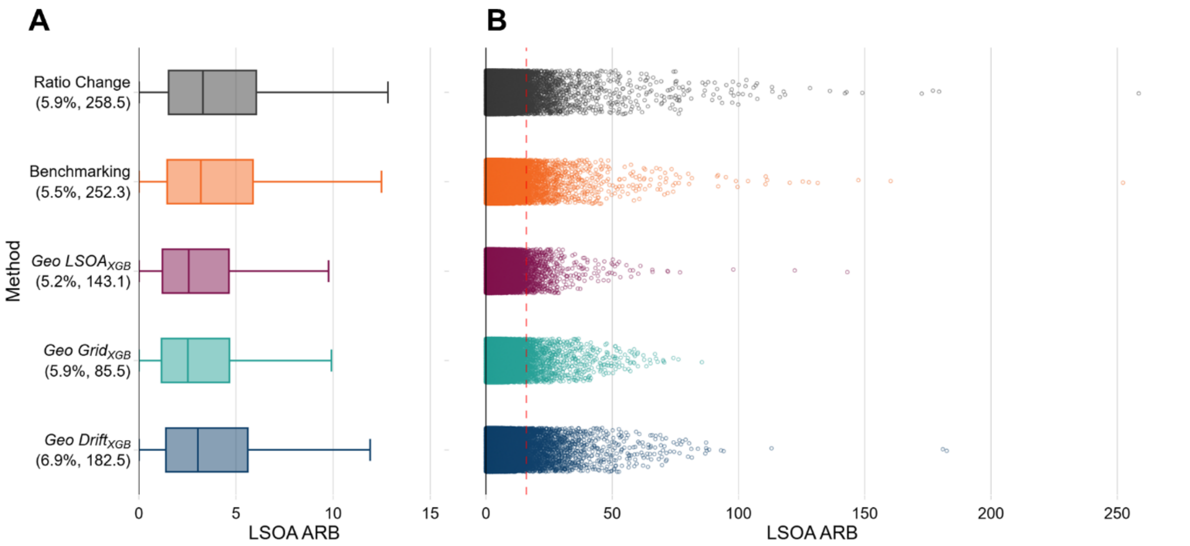

Figure 1 *ARB distributions of LSOA estimates as measured at mid-year 2021 against Census-based mid-year estimates. (A) Min, quartiles, and max ARB excluding LSOAs more than 1.5*IQR above the third quartile (parenthesis on the y-axis denote the percentage of LSOAs excluded, and the maximum ARB). (B) ARB for all LSOAs (vertical jittering for visibility; dashed red line at ARB=15 shows where the left panel ends).*

Table 3 *Number of LSOAs within each ARB group as measured at mid-year 2021 against Census-based mid-year estimates (percentages in parentheses).*

Method	**<10 ARB**	**10 to <25 ARB**	**25 to <50 ARB**	**50 to <100 ARB**	**≥100 ARB**
Ratio change	32,086 (*89.95*)	3023 (*8.47*)	417 (*1.17*)	123 (*0.34*)	23 (*0.06*)
Benchmarking	32,444 (*90.95*)	2807 (*7.87*)	346 (*0.97*)	64 (*0.18*)	11 (*0.03*)
$LSOA_{XGB}$	33,909 (*95.06*)	1602 (*4.49*)	140 (*0.39*)	19 (*0.05*)	2 (*0.01*)
$Grid_{XGB}$	33,599 (*94.19*)	1721 (*4.82*)	302 (*0.85*)	50 (*0.14*)	0 (*0*)
$Drift_{XGB}$	32,352 (*90.69*)	2656 (*7.45*)	552 (*1.55*)	109 (*0.31*)	3 (*0.01*)

5.1.2 Estimates

Experimental LSOA SAPE plotted against true LSOA values (Census-based MYE 2021) are presented in Figure 2. Although estimates from all methods are highly correlated (Table 4; Figure 4), there are several structural differences in the patterns of bias across LSOA population size between methods. Geospatial LSOA models (including drift; Figure 2C, 2E) and benchmarking (Figure 2B) have relatively symmetrical patterns of over- and under-estimation. Ratio change (Figure 2A) has more over-estimates than under-estimates (particularly for the most extreme over-estimates). Conversely, the grid model (Figure 2D) has more under-estimates than over-estimates (particularly for the most extreme under-estimates which occur in more populous LSOAs). Notably, it has no over-estimates in the two worst ARB categories.

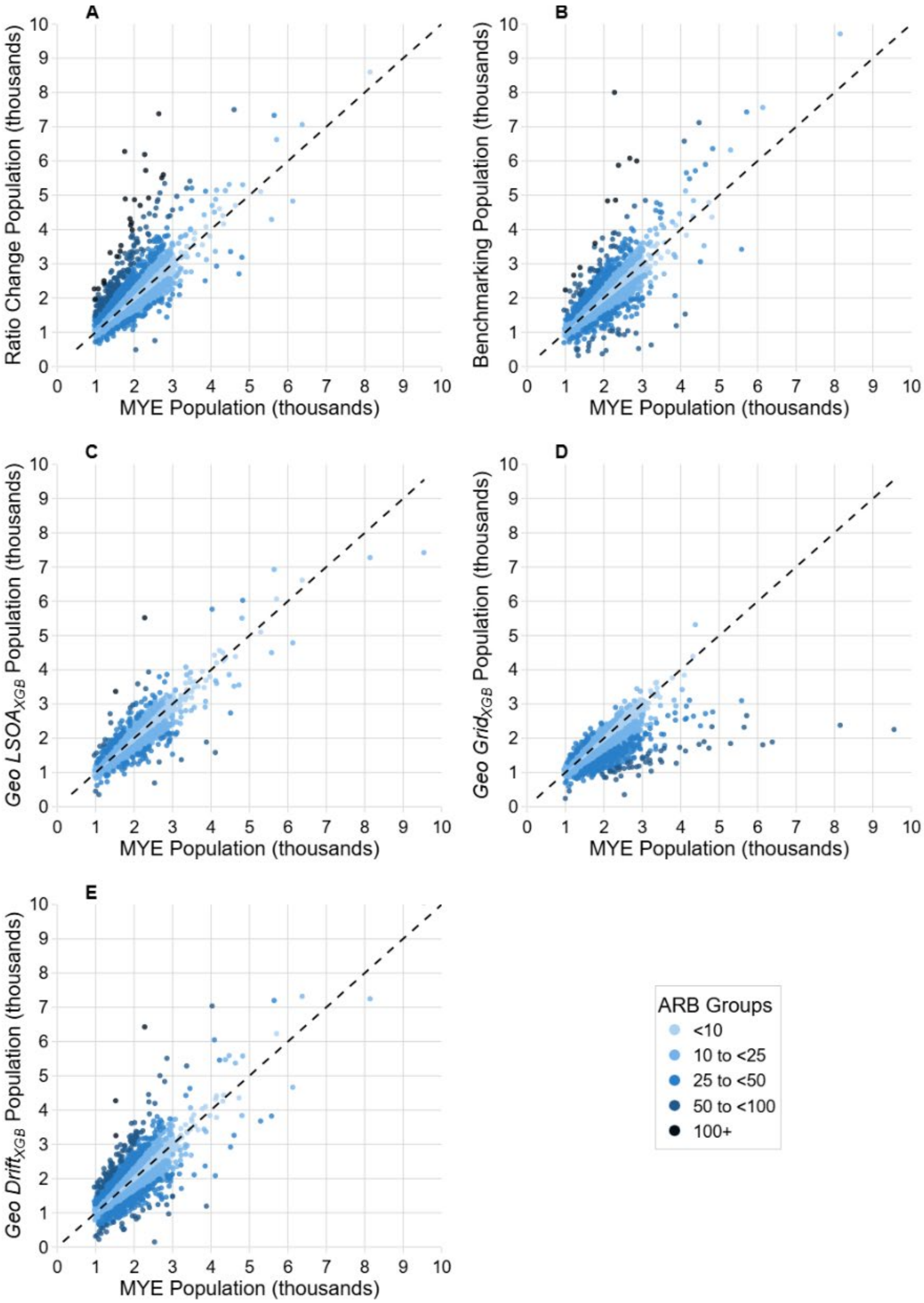

***Figure 2** Experimental LSOA SAPE plotted against Census-based mid-year estimates in 202; (A) ratio change, (B) benchmarking, (C) geospatial LSOA$_{XGB}$, (D) geospatial Grid$_{XGB}$, (E) geospatial Drift$_{XGB}$. Each point represents one LSOA and is colour coded according to its ARB group. Dashed diagonal lines show x=y whereby points above are over-estimates and below are under-estimates.*

Table 4 *Spearman's rank correlation coefficients for pairwise comparisons of LSOA SAPE for all methods at mid-year 2021. All coefficients are significant at p<0.001.*

	Ratio change	Benchmarking	*LSOAXGB*	*GridXGB*
Ratio change				
Benchmarking	0.944			
LSOAXGB	0.942	0.968		
GridXGB	0.908	0.928	0.950	
DriftXGB	0.935	0.957	0.961	0.929

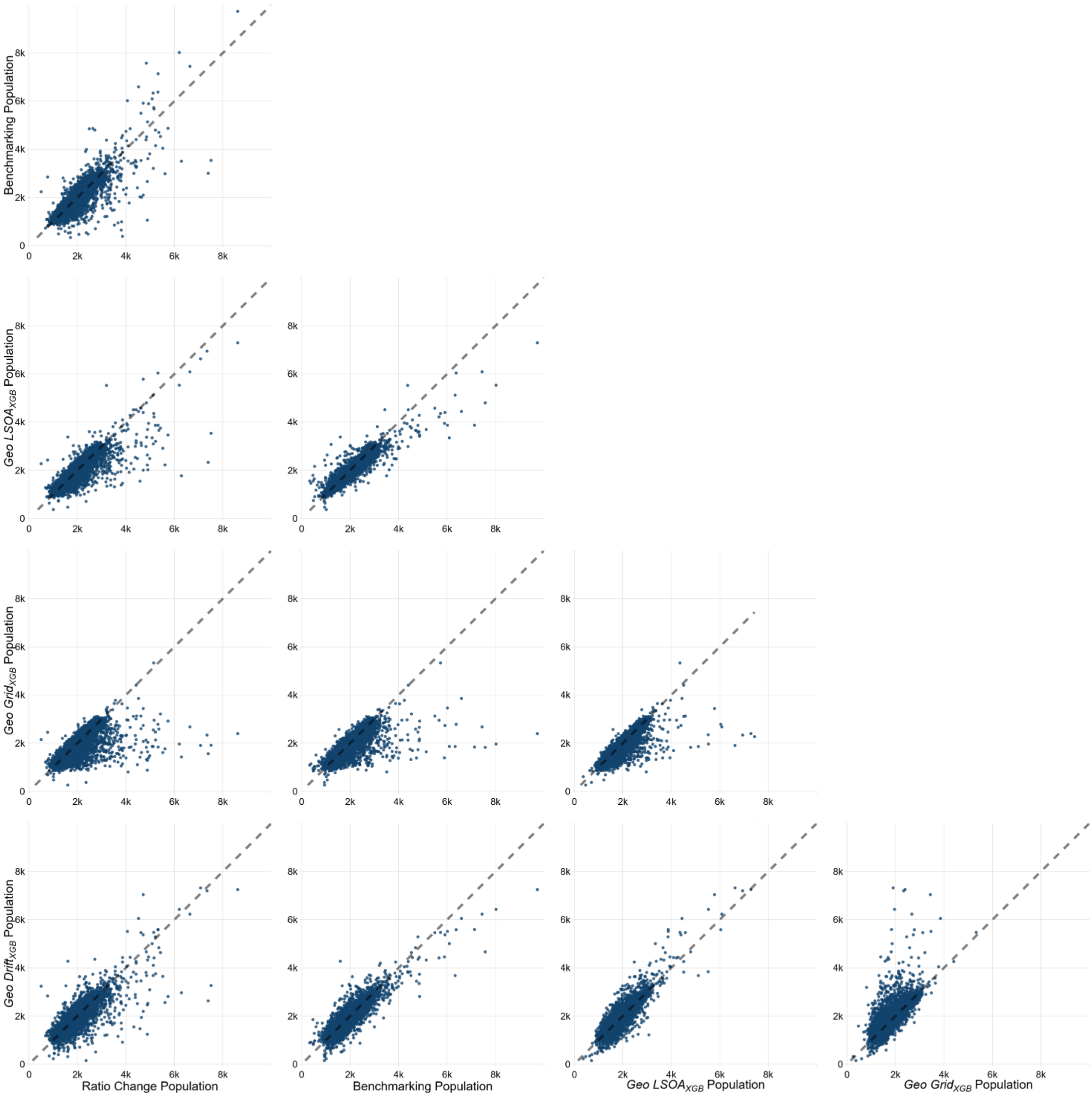

Figure 3 *Pairwise comparisons of LSOA SAPE between all methods. Dashed diagonal lines are lines of equality (x=y). Points are individual LSOAs.*

5.2 Feature importance

The Administrative Population and Building/Property features are the most important feature themes used by the full 0-year lag models (Figure 4). However, while the two themes are similarly scored for the LSOA model, the Building/Property theme is much more important for the grid model. Concomitantly, the importance scores for the Environmental and Infrastructure themes for the grid model are higher than the LSOA and drift models. The drift model also differs from the full 0-year lag LSOA model in that the Building/Property theme is much lower than Administrative Population and Energy is much higher in importance.

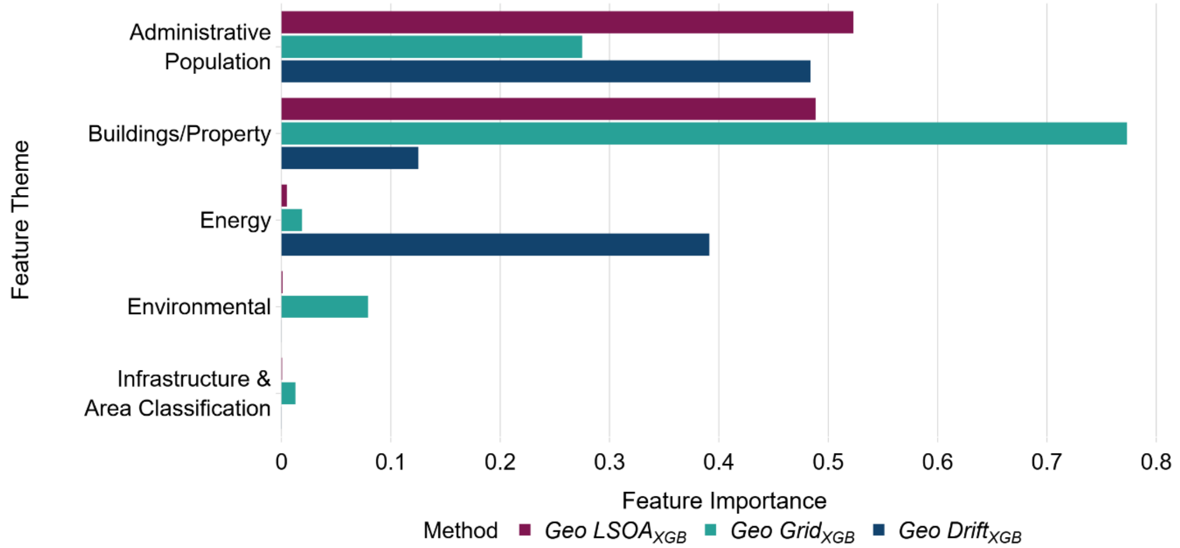

Figure 1 *Feature theme importance scores (sum of the feature gain scores within themes) for geospatial models.*

5.3 Bias breakdowns at mid-year 2021

We have explored patterns in SAPE bias across various area characteristics such as area classifications, spatial distribution, and population type/size as follows.

5.3.1 Rural Urban Classification

The traditional methods both have higher levels of bias for urban areas than rural (Figure 5), and the areas with the highest levels of bias tend to be almost entirely urban (Figure 6). In contrast, the difference between urban and rural areas is less consistent for the geospatial models, and they generally have higher ARB for smaller rural areas than larger rural (Figure 5). Notably, the LSOA models (LSOA & drift) have similar ARB profiles in both smaller rural areas and urban areas, and suffer higher bias in smaller rural areas than the traditional methods (Figure 5). Additionally, the areas with the highest ARB for geospatial LSOA models are more likely to be smaller rural unlike the traditional methods (Figure 6). Overall, the geospatial models achieve lower bias than the traditional methods in urban areas, but their advantage is inconsistent in larger rural areas, and only the grid model outperforms traditional methods in smaller rural areas.

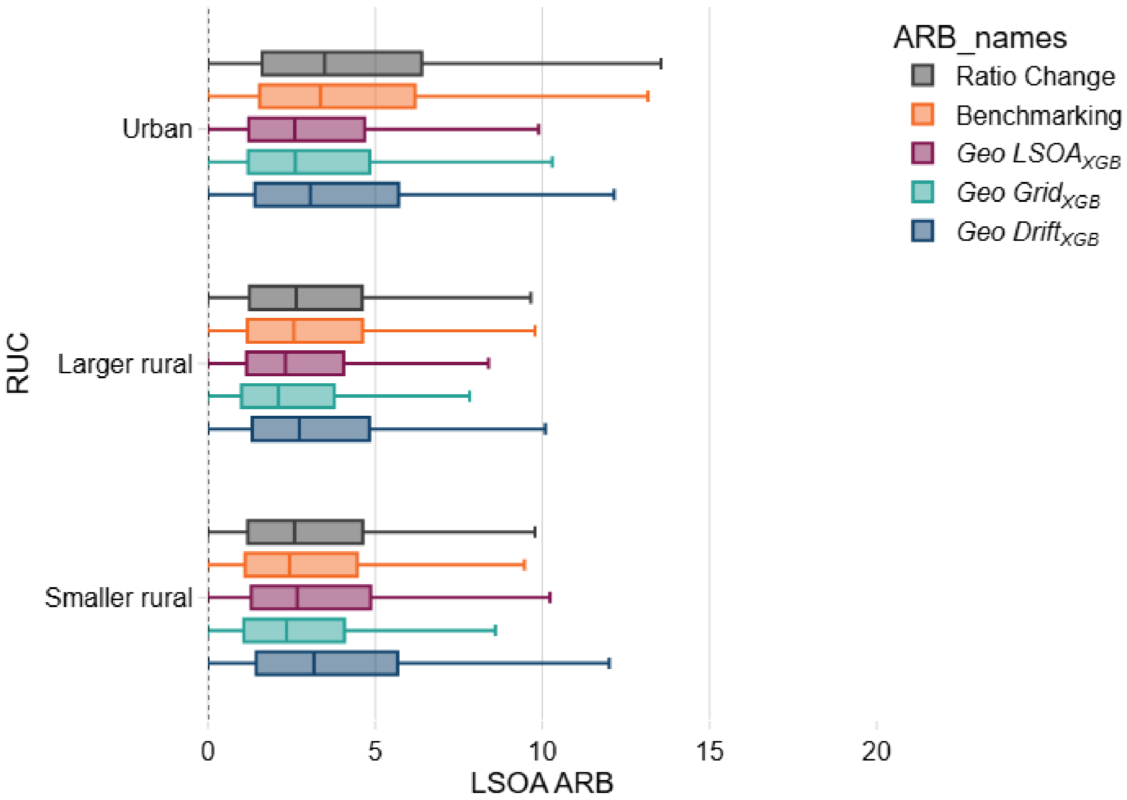

Figure 2 *ARB distributions of LSOA estimates as measured at mid-year 2021 against Census-based mid-year estimates, broken down by Rural Urban area Classification (2021). Box and whiskers exclude LSOAs more than 1.5*IQR above the third quartile.*

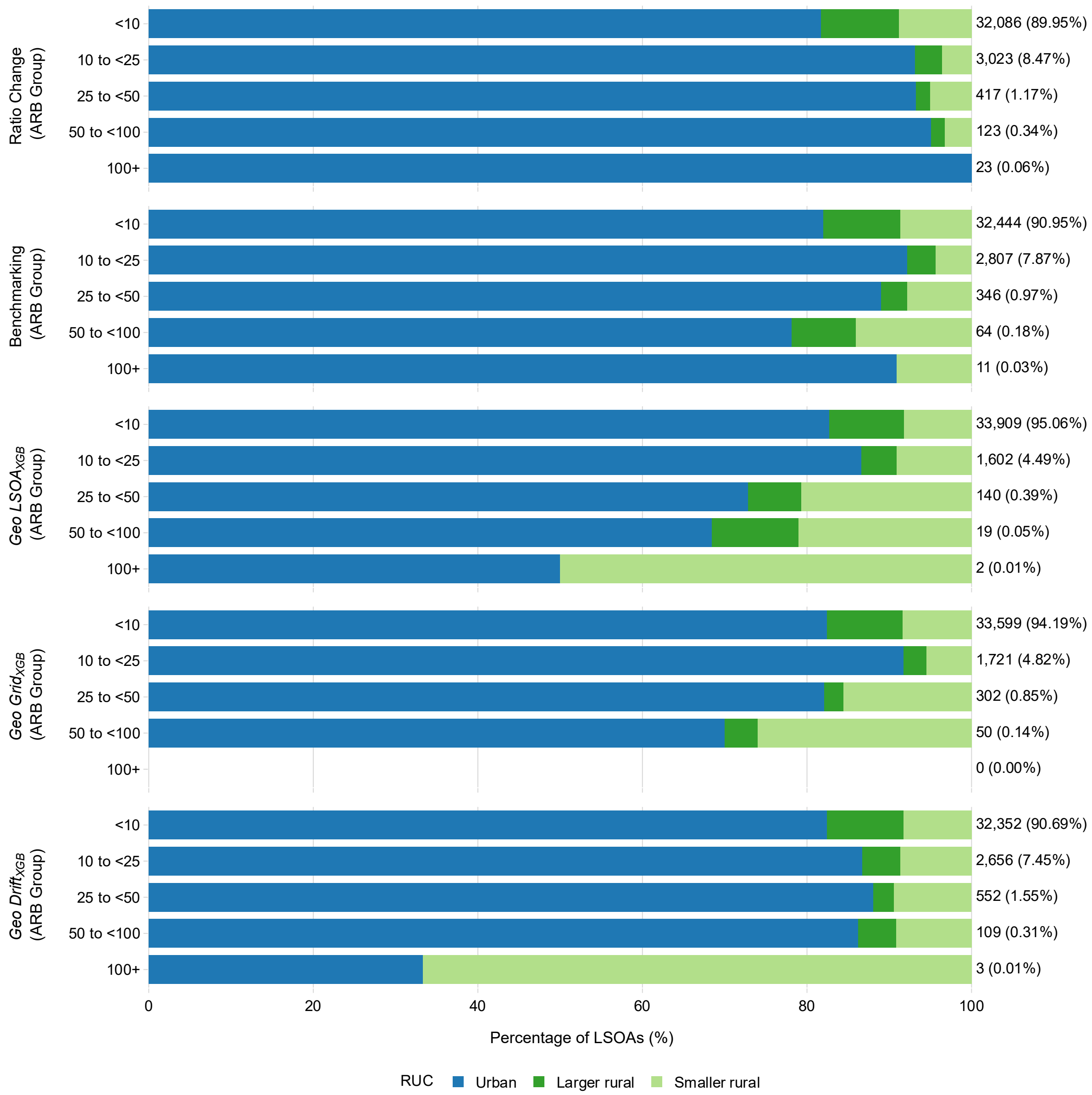

***Figure 3** Percentages of LSOAs in each ARB group that belong to each Rural Urban area Classification category (2021) for each method. Numbers to the right of each bar display the count of LSOAs in each group for each method (percentage in parentheses).*

5.3.2 Geography

The relative performance of geospatial and traditional methods across sub-national regions is reasonably consistent with the national results with some exceptions. While the geospatial models outperform the traditional methods in most regions, only the non-drift models have an advantage in London, and all the methods perform similarly in Wales (Figure 7). Notably, all the methods perform substantially worse in London than they do in all other regions.

Problem areas (LSOAs with ARB≥10) make up a minority of LSOAs in most LAs for all methods, but are a majority of LSOAs for a small number of LAs (predominantly in London) (Figure 8). The geospatial models tend to have fewer LAs with high concentrations of problem areas than the traditional methods (Figure 8). Amongst the geospatial models, the grid model has relatively higher concentrations of problem areas in LAs outside of London compared to the LSOA model (Figure 8).

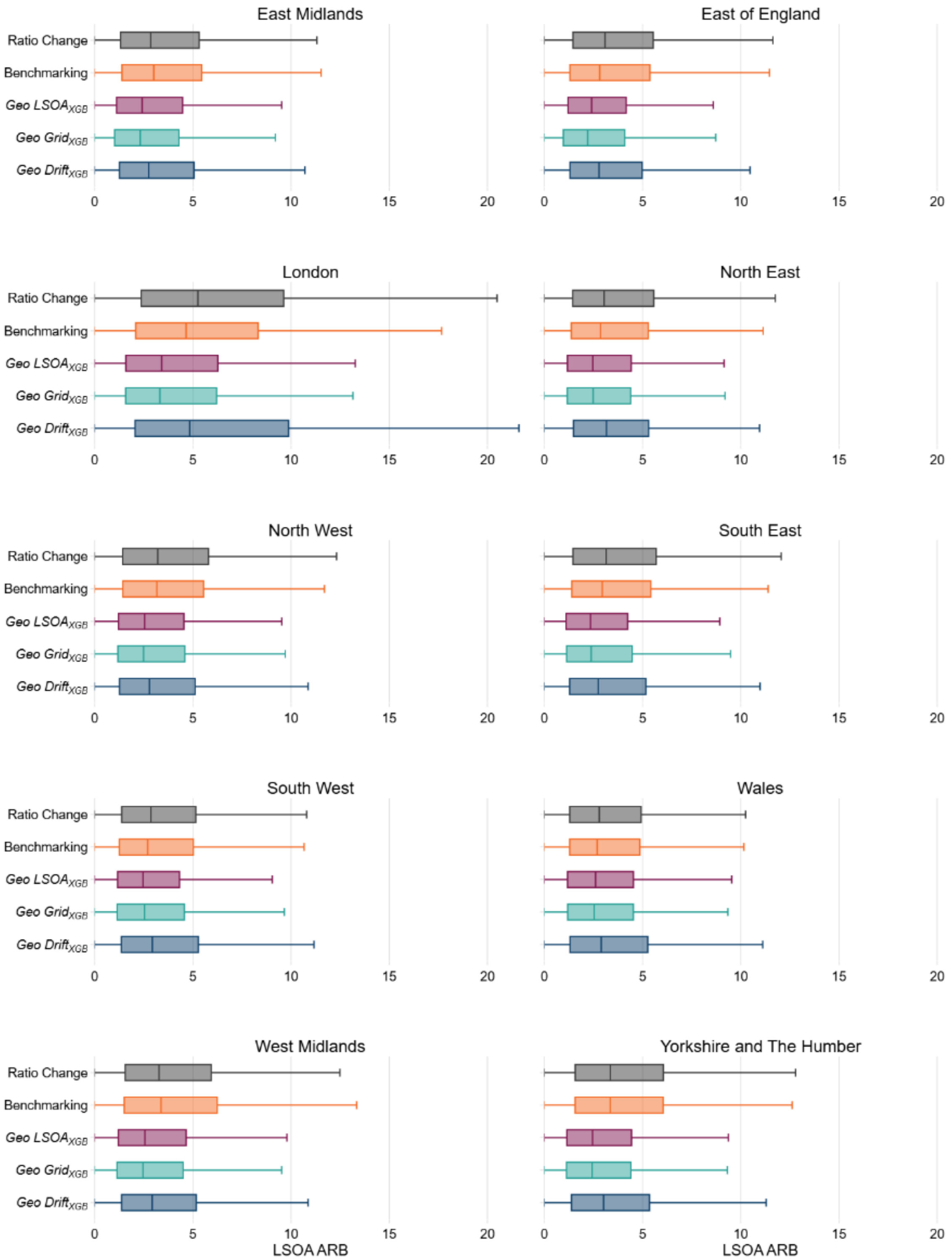

***Figure** 4 ARB distributions of LSOA estimates for all methods as measured at mid-year 2021 against Census-based mid-year estimates, broken down by region. Box and whiskers exclude LSOAs more than 1.5*IQR above the third quartile*

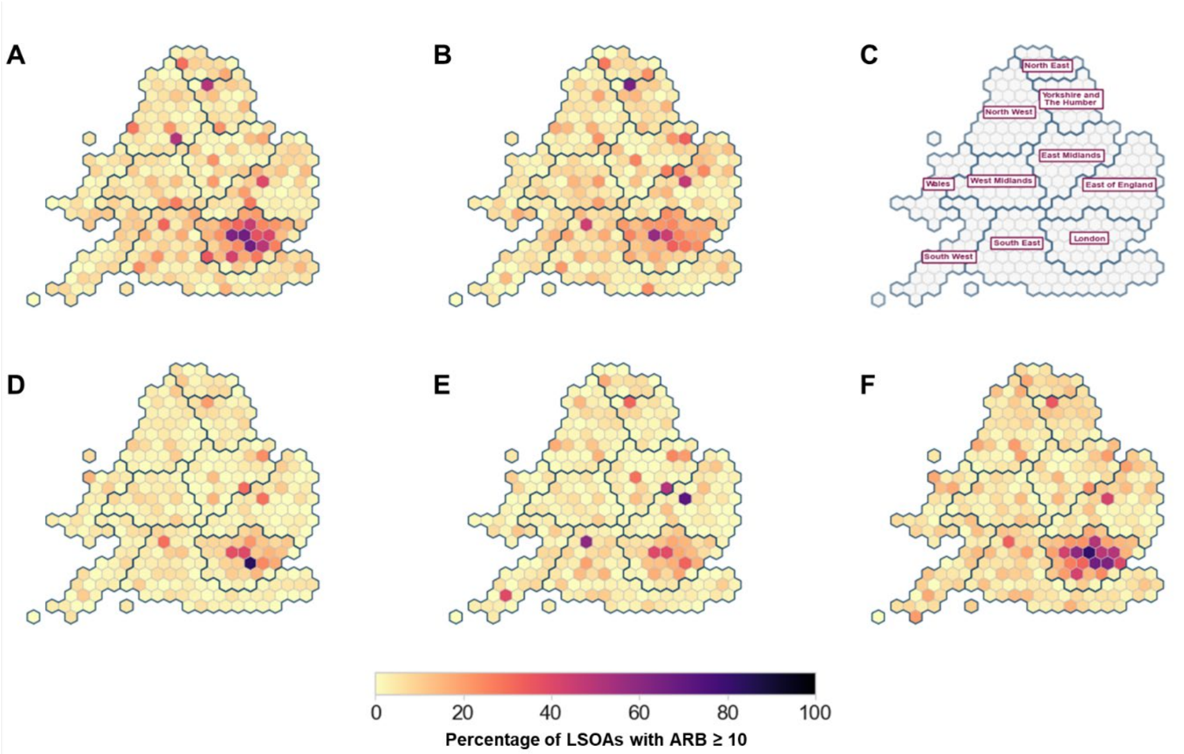

***Figure** 5 Hex map of England and Wales where each hex represents a LA, colour coded by the percentage of problem areas in each LA (defined as LSOAs with an ARB above 10). (A) ratio change, (B) benchmarking, (C) region guide, (D) geospatial LSOA$_{XGB}$, (E) geospatial Grid$_{XGB}$, (F) geospatial Drift$_{XGB}$. Hex layout produced by Open Innovations (MIT).*

There is substantial overlap in the top 10 most problematic LAs (defined as those with the highest proportion of LSOAs with ARB≥10) across all methods, and they are predominantly in London (Figure 9). However, even in the worst performing LAs most LSOAs are still in the less problematic categories for all methods, with only a minority of LSOAs in the worst ARB categories (Figure 9).

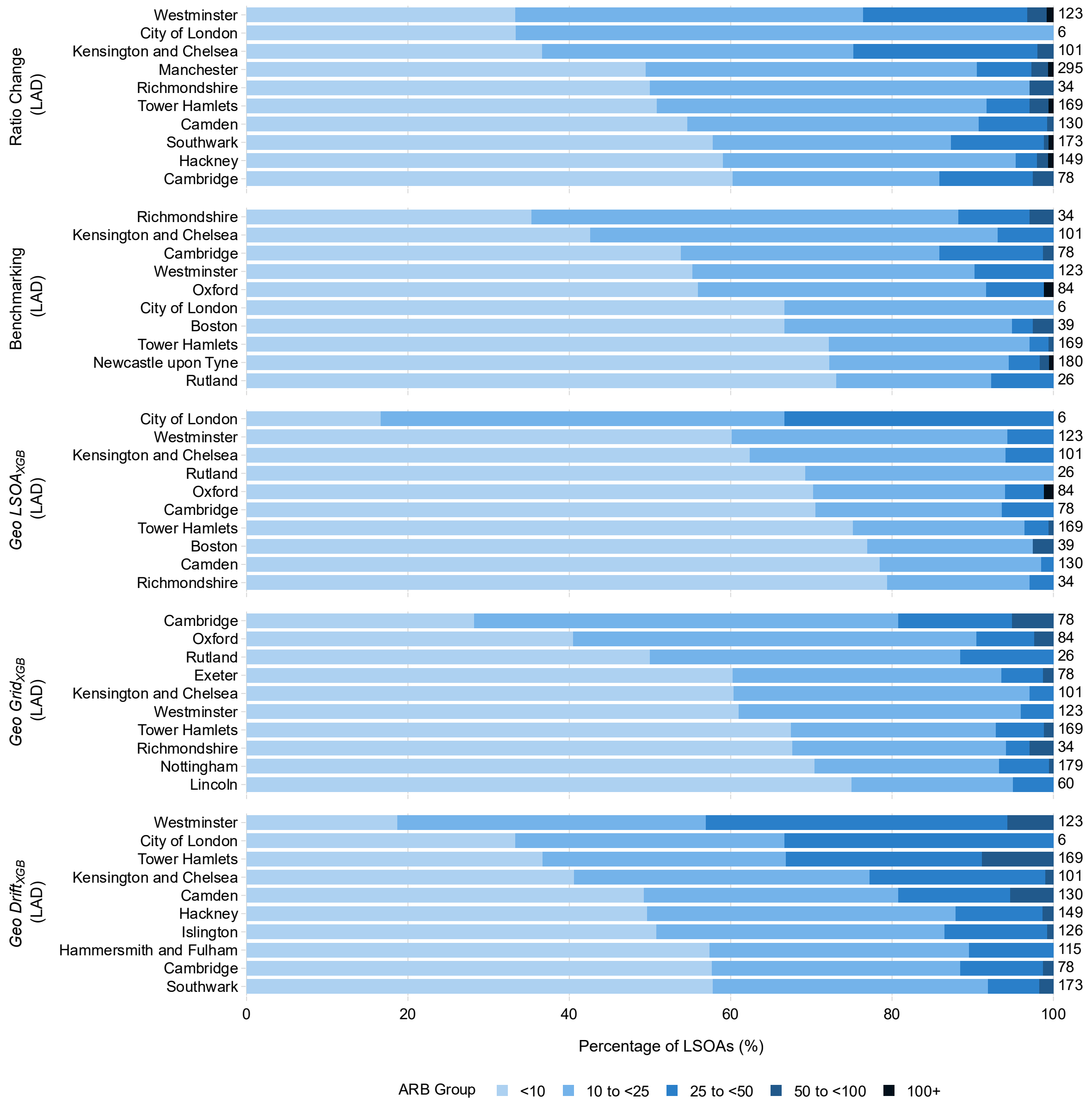

***Figure** 6 The top 10 most problematic LAs for each method (defined as LAs with the highest proportion of LSOAs with an ARB≥10), broken down by the percentage of LSOAs in each ARB problem group for each LA for each method. Numbers to the right of the bar display the number of LSOAs in each LA.*

5.3.3 Population

Assessing relative bias as a function of population density shows slightly different patterns of over- and under-estimation compared to nominal population (Figure 10). Specifically, all the geospatial models appear to underestimate high-density LSOAs but have more mixed results at lower and intermediate population densities. The LSOA model is relatively symmetrical for over- and under-estimation; the grid model tends to underestimate at lower and intermediate densities; and the drift model is more likely to overestimate at the lowest densities and gradually become more likely to underestimate as population density increases. In contrast, the traditional methods do not appear to have any tendency to primarily over- or underestimate at high densities, but both have more over-estimates than under-estimates in areas at lower and intermediate population densities.

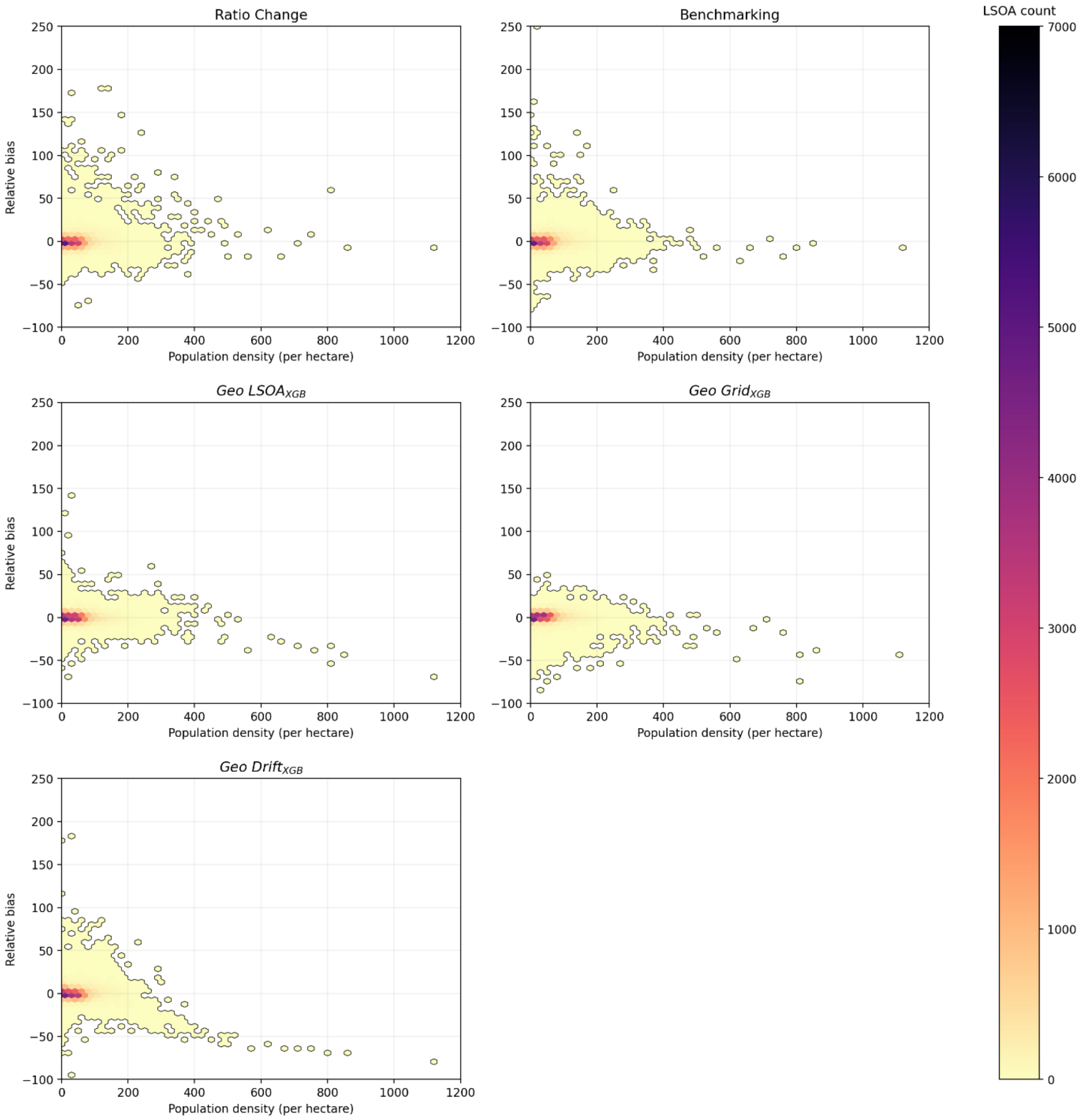

Figure 70 *Relationship between population density and relative bias for all methods. Population density measured as the Census-based mid-year population estimate per hectare for each LSOA. Hex points are bins of width 20 people per hectare by 17.3 points of bias (bias bin width defined as function of the density bin width), coloured by the number of LSOAs in the bin.*

Problem LSOAs tend to be more likely to contain Communal Establishments (CEs), with a higher proportion of areas containing at least one CE as ARB group increases (Figure 11). However, the inverse is not true and most LSOAs that contain at least one CE are in the lowest ARB group.

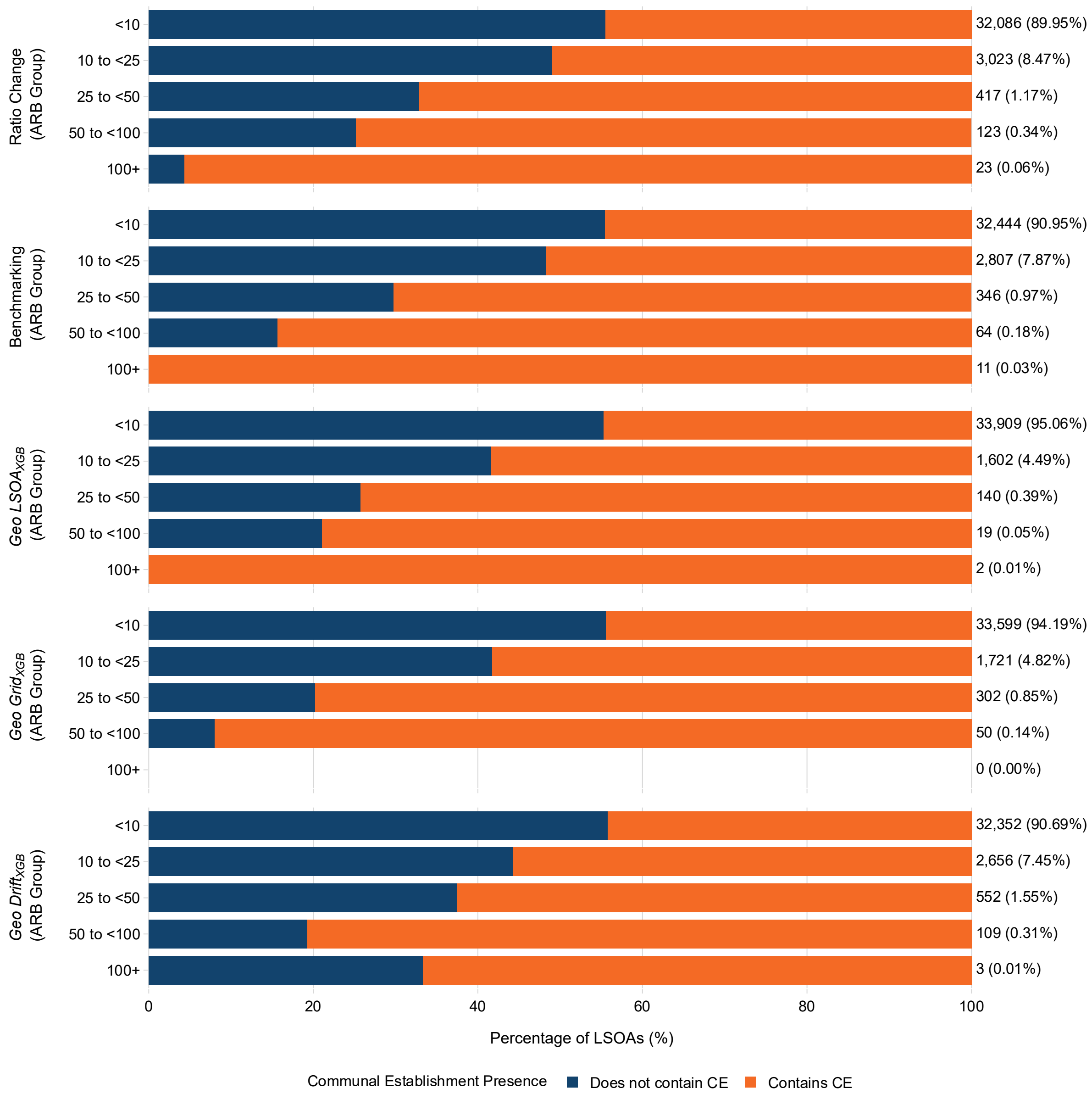

Figure 11 *Percentages of LSOAs in each ARB group for each method broken down by the percentage of LSOAs that contain at least one communal establishment. Numbers to the right of each bar display the number of LSOAs in each group for each method (percentage in parentheses).*

5.4 Temporal stability (2011–2021)

We explore the temporal stability of the methods between Census day 2011 and Census day 2021 to give a more complete picture of intercensal SAPE quality. Note that while we have made efforts to make our reduced drift model comparable to our full LSOA model, it does not use all the features available in 2021 and so has slightly higher ARB than the full model (Annex Table A2).

5.4.1 Bias drift

The change in median ARB between 2011 and 2021 fits the expected patterns for the traditional methods; ratio change starts the decade from Census estimates with very low bias but ends the decade with the highest median ARB, and benchmarking is relatively more stable starting the decade worse than ratio change but slightly outperforming it by the end of the decade (Figure 12). The median ARB of benchmarking is slightly higher in 2021 than 2011 due to input data change (switching from PR in 2011 to PDS in 2021) rather than a general temporal trend that we might expect to continue as with ratio change. The geospatial drift model follows the same pattern as benchmarking, albeit with slightly lower median ARB (Figure 12). In contrast to benchmarking, we would expect the median ARB of the geospatial model to continue to rise over time due to model drift unless the model was retrained with subsequent Census data. The exact year of the cross-over when ratio change falls behind other methods cannot be precisely determined from the linear interpolations we present in Figure 12, but it is likely to be in the final quarter of the decade.

The rank order of LSOA ARB achieved by the drift model is only weakly correlated between 2011 and 2021 ($r_S(33{,}645)=0.25$, $n=33{,}647$, $p<0.001$), and the ARB of some areas actually decreases between 2011 and 2021 (Table 5). While the increase in median ARB shows that model drift results in worse performance overall, it also results in changes to the distribution of bias across LSOAs rather than simply increasing bias by a uniform factor across all LSOAs.

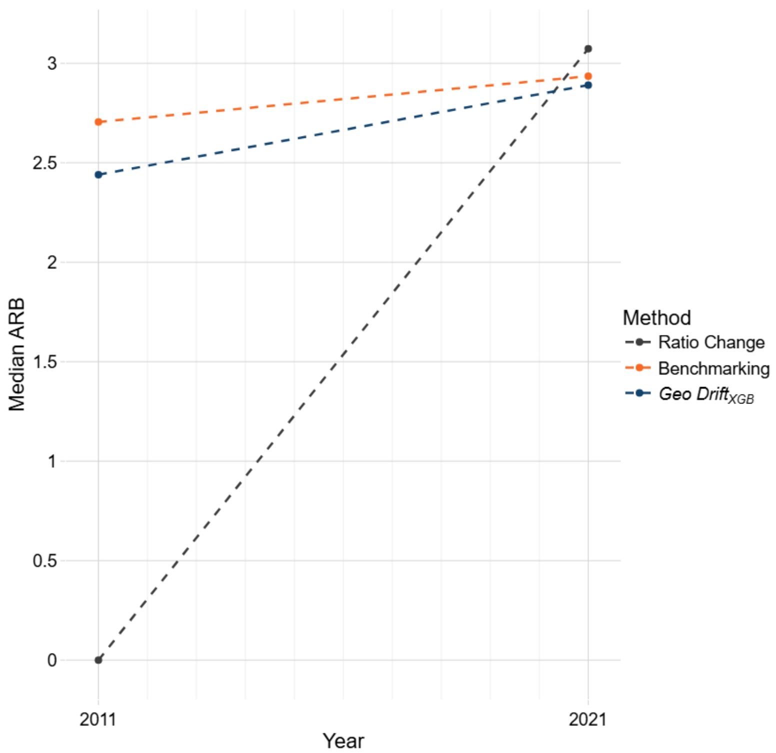

***Figure** 812 Median ARB of the traditional methods compared with the geospatial drift model in 2011 and 2021. Dashed lines are linear interpolations.*

__Table 5__ Number of LSOAs that increase or decrease in ARB group between 2011 to 2021 (drift model). Shaded cells on the diagonal are counts of LSOAs that remained in the same ARB group in both years, cells above the diagonal count LSOAs whose ARB moved into higher bias groups between 2011 and 2021 (i.e. bias got worse), and cells below the diagonal count LSOAs whose ARB moved into lower bias groups between 2011 and 2021 (i.e. bias got better).

		Drift$_{XGB}$ (2021)				
	ARB Group	**<10**	**10 to <25**	**25 to <50**	**50 to <100**	**>100**
Drift$_{XGB}$ (2011)	**<10**	30,090	1694	207	27	0
	10 to <25	804	473	134	25	1
	25 to <50	54	48	51	12	0
	50 to <100	2	4	8	8	2
	>100	0	1	0	1	1

5.4.2 Model drift feature sensitivity

The factor by which the median ARB of the geospatial drift model increases at a Census lag of 10 years vs. a lag of 0 years was impacted by the choice of features. Removing any theme of features increases the model drift factor relative to including all feature themes, but loss of Administrative Population or Building/Property has the largest impact (Figure 13).

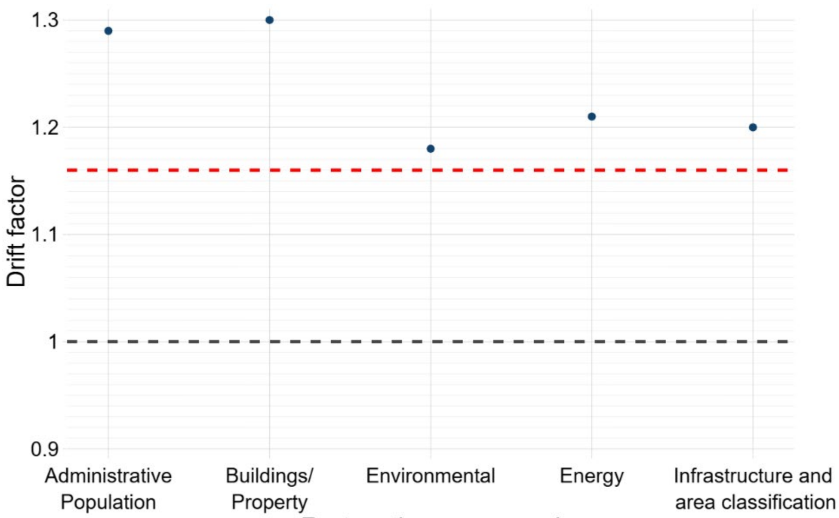

***Figure** 913 Drift factors (the scaling factor between the median ARB of the drift model in 2021 compared to 2011) for the drift model when different feature themes are removed. Dashed red line indicates baseline drift factor when all feature themes are included, dashed black line indicates no drift.*

6. Discussion

We have shown that bottom-up geospatial ML models suffer less extreme bias than ratio change by the end of the intercensal decade, and generally outperform benchmarking. However, they do not perform well enough across the full intercensal decade to recommend as a replacement production method for intercensal SAPE. We have also shown that grid-based geospatial models make very few extreme errors, although can be prone to under-estimation in populous areas, and their performance across the intercensal decade is not assessed.

While we cannot recommend replacing ratio change with geospatial models, the two approaches have an apparent complementarity in their average bias profiles that may make a combined approach advantageous. Ratio change utilises Census estimates to yield extremely low levels of bias for most of the decade but has trouble with extreme errors (particularly towards the end of the decade). Geospatial models have higher bias across most of the decade but have much more consistent performance across LSOAs with much fewer extreme errors (particularly the grid models), and LSOA models outperform ratio change towards the end of the decade. Research is needed to explore approaches for combining the methods, but our initial thoughts are related to using one method as an input to the other. For example, model-based SAPE could be used to calculate annual population change scaling factors to be used as inputs to ratio change in place of the scaling factors currently calculated from admin data stocks, or geospatial models could be specified to use annual scaling factors (and other types of longitudinal features) and/or ratio change estimates as features. A hybrid approach may be possible to achieve consistently lower average bias across the decade than either method can achieve independently if the annual accumulation of bias in model-based LSOA population size across the decade is smaller than that of the admin data currently used by ratio change. However, depending on the underlying causes of this apparent complementarity, it could be the case that a hybrid approach may inherit the weaknesses of both and result in higher average bias across the decade, or more extreme bias in some areas.

There are also some modelling options we could explore that may improve the performance of the geospatial models, including some where we haven't yet explored alternatives to the original WorldPop approach. These include, but are not limited to, predicting linear population density rather than log density (tree-based algorithms should be robust to non-normal data), predicting population size rather than density, exploring how we might incorporate spatial autocorrelation into the models, considering approaches that can meaningfully extrapolate beyond observed data, and exploring unexplained heterogeneity in the estimates. Alternatively, as we are ultimately using model predictions to produce LSOA-level disaggregation weights for LA estimates we may want to consider predicting these weights directly. There are also alternative methods for producing intercensal SAPE that have been beyond the scope of this work e.g. Bayesian Hierarchical Models or linear mixed-effects models, as well as alternative ML algorithms e.g. neural networks or mixed-effects random forests.

There is a good deal of overlap in the types of problem areas with the highest ARB across methods (e.g. high-density populations), suggesting that some areas have intrinsic qualities that make them difficult to estimate. This is primarily the case for several LAs in London, and some LSOAs that contain residents in communal establishments e.g. students. While we have found some evidence that the areas with the worst bias are more likely to contain CEs, further work would need to confirm which kinds of CEs make LSOAs difficult to estimate. However, there are some differences in problem areas between the methods. The geospatial models have fewer problem areas concentrated in individual LAs outside of London compared to ratio change, but struggle more with smaller rural areas. Conversely, ratio change has more difficulty within urban areas. In future it would be valuable to develop quality standards for SAPE ARB informed by user engagement so we can better define problem areas and success criteria for assessed methods.

LSOA models and grid models perform broadly similarly for most LSOAs when assessed within 2021 (i.e. 0-year lag), with two key differences. Grid models result in fewer problem LSOAs with high bias and yield a lower maximum bias, but generally tend to be more likely to underestimate the population. This underestimation issue may have a disproportionate impact on users because it seems to be more common in highly populated areas. This may be related to the presence of high-density population groups (e.g. students) that make a small number of highly populated grid cells difficult to predict because of their extreme difference from the average cell. The two most problematic areas for the grid model are two high population LSOAs at the heart of the Cambridge and Oxford University historic colleges. It could also be related to back-transforming log population density by exponentiation without a bias adjustment when converting model density predictions to nominal population totals, which can lead to underestimation (16). Addressing this issue is a prerequisite to recommending the use of grid models for the production of SAPE on Census geographies.

A drawback to geospatial models is their complexity relative to traditional methods. While the process for training and assessing the models is complex, they are not so complicated that they are problematic for statisticians to implement or use. However, the methods can be difficult to explain to non-technical users. All else being equal, we would generally prefer a simpler method, and so any increase in methodological complexity must be justified by improvements in the quality or utility of the statistics. Given that ONS already publish model-based statistics we believe that geospatial models are not too complex, but this would need to be explored via user engagement.

All the methods we have compared are reliant on input data where quality or availability may change over time beyond the control of ONS. Ratio change and benchmarking use administrative population counts produced from NHS patient data to calculate their scaling factors (although they could use alternative sources of population counts), while the geospatial models use features derived from multiple domains e.g. population, properties, energy, and infrastructure. Geospatial models, while still impacted by changes in data quality and availability, are more robust to changes in individual sources as the methods make use of multiple features that can be substituted. This is demonstrated when assessing the impact of removing the two most important feature themes from the drift model. We see larger increases in model drift when removing building/property features than when removing administrative population counts, despite the latter having a greater feature importance in the reduced model. We believe this is largely a result of improvements in quality of the former and a reduction in that of the latter over the decade i.e. when administrative population counts are removed models compensate and rely more heavily on buildings/property features, which are of higher quality in 2021. This is not observed to the same extent when removing building/property data due to the lower quality of administrative population counts in 2021.

There are some caveats with our assessment approach to consider when interpreting our results. For all methods, all LSOA estimates we assess are constrained to standard LA totals in the assessment period before ARB is calculated. This omits a key source of error in published SAPE (bias in the LA estimates) but allows us to standardise LSOA estimates across methods, and base our conclusions only on differences in bias that can be attributed to the small area methods rather than the LA estimation method. Our bias metrics measure how accurately each method reproduces the true LSOA population distribution (i.e. the percentage of the LA population in each LSOA), but we would expect to see more problem areas in published estimates as these would incorporate additional bias from LA estimates. Another key limitation is uncertainty about how grid model bias might change across the intercensal period. Our results show that grid models have different ARB patterns to LSOA models, and utilise features in different ways (likely due in part to differences in the quality of some sources at different spatial scales). This suggests that their patterns of model drift are likely quantitatively different from the LSOA models, but it is difficult to predict whether they would be qualitatively different in a way that would lead us to make different conclusions about intercensal performance. Validating the temporal stability of the grid models is a prerequisite to recommending their use in production, but it is currently unclear if we can find a solution to the lack of 2011 grid resolution data required to fit a grid drift model.

In addition to the data described in this paper we have also begun investigating data from the Valuation Office Agency (VOA) and the Higher Education Statistics Agency (HESA), but our understanding of these data is not yet sufficient to include in our models. The former contains additional building attribute data that may give more informative measures of property occupancy e.g. number of bedrooms, council tax records. The latter provides counts of students that may help with predicting CE populations, and serve as proof of concept for seeking similar admin-based population counts for other CE types. We also plan to investigate Council Tax data to explore how we might develop better measures of property occupancy rates.

We have focussed on assessing the bias in estimates of the total LSOA population, but the statistical system also requires estimates for OAs and MSOAs, estimates broken down by age and sex, and uncertainty measures. In future it will be necessary to investigate the capability of geospatial models to address these requirements. Our results suggest geospatial models can produce estimates at a range of geography sizes (from 100m x 100x grid cells to the span of variable sizes of LSOAs), but it is difficult to predict whether results at the OA and MSOA level would lead to the same conclusions. It is unclear whether geospatial models will be capable of producing breakdowns by age and sex, and additional methods may be needed. Quantifying uncertainty may also be challenging, as while there are approaches to quantify the uncertainty of model predictions (e.g. quantile regression) we would then have to propagate that uncertainty into the subsequent SAPE, likely through some sort of bootstrapping or simulation approach.

Additionally, ONS produces experimental estimates for some non-Census geographies in response to user need e.g. National Parks. Grid-based geospatial models may be a general method to produce SAPE for these bespoke geographies (again, if the underestimation issue can be solved and their temporal stability can be validated). A single set of grid estimates can in principle be aggregated to any arbitrary higher-level geography where boundary geometry is available. There are open questions regarding the validity of this approach, the most important of which are (i) whether we can constrain grid cell predictions to LA estimates before aggregating to arbitrary geographies without introducing additional bias, (ii) whether the resolution of the grid cells and our approach to splitting cell predictions when cells are intersected by a boundary places a lower limit on the size of bespoke geographies (where the ratio of boundary cells to non-boundary cells becomes too high), and (iii) whether any potential combination of ratio change and geospatial models developed for Census geographies would also be applicable in a grid-based approach for bespoke geographies.

7. Conclusion

Geospatial ML models do not achieve a level of bias sufficient to justify replacing ratio change in the production of intercensal SAPE at the LSOA level. However, LSOA geospatial models achieve lower bias than ratio change 10 years after a Census, and have much fewer issues with extreme errors; achieving more consistent bias across LSOAs over time. Grid-based geospatial models have even fewer problems with extreme errors when assessed within 2021, and may be the preferable geospatial method if issues with under-estimating high-density areas can be overcome, and if their performance 10 years after a Census can be validated. An approach that combines the ratio change and LSOA geospatial models may be able to achieve lower average bias and less extreme bias than either method independently and should be tested.

8. Future Steps

We recommend the following future research (in suggested priority order):

- Explore methodological options to combine ratio change and geospatial modelling to produce LSOA SAPE.
- Explore options for assessing model drift in the grid models.
- Explore potential improvements to geospatial models, particularly addressing underestimation in high-density areas.
- Test the capability of geospatial models to produce estimates at other Census geographies (OA and MSOA).

- Develop a method to quantify the uncertainty of geospatial model-based SAPE.
- Test the capability of geospatial models to produce estimates broken down by age and sex.
- Explore the capability of grid-based models to act as a generalised approach for bespoke geographies.

In addition to this research we recommend that Demography pursue user engagement to develop prioritised SAPE use cases and user quality requirements. Quantitative user quality requirements would allow us to more robustly define and prioritise problems with current estimates, and establish quality targets/success criteria for experimental estimates.

An earlier version of this paper has been reviewed by members of the WorldPop group at the University of Southampton, and we will seek consultation and where possible collaboration with them on future work.

9. References

1. **Office for National Statistics.** The plan for ONS economic statistics. [Online] 26 June 2025. https://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/economicstatisticstransformation/theplanforonseconomicstatistics.

2. —. Population estimates for England and Wales, mid-2024: methods guide. [Online] 25 July 2025. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/populationestimatesforenglandandwalesmid2022methodsguide.

3. —. Methodology note on production of population estimates by output areas, electoral, health and other geographies, England and Wales. [Online] 25 November 2024. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/methodologynoteonproductionofpopulationestimatesbyoutputareaselectoralhealthandothergeographiesenglandandwales.

4. —. Small Area Population Estimates in the transformed population estimation system: methods development. [Online] 12 2023. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/smallareapopulationestimatesinthetransformedpopulationestimationsystemmethodsdevelopment.

5. *Small Area Estimation in the Era of Machine Learning and Alternative Data Sources: Opportunities, Challenges, and Outlook.* **Tzavidis, N.** 2025, Journal of Official Statistics, pp. 921-929.

6. *EAP210 Geospatial methods for Small Area Population Estimates: proof of concept.* **Office for National Statistics.** 2024. Office for National Statistics Methodological Assurance Review Panel.

7. **Office for National Statistics.** Geospatial methods for Small Area Population Estimates: proof of concept. [Online] 30 July 2024. https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/geospatialmethodsforsmallareapopulationestimatesproofofconcept.

8. *Random Forests.* **Breiman, L.** 2001, Machine Learning, pp. 5-32.

9. *XGBoost: A Scalable Tree Boosting System.* **Chen, T. and Guestrin, C.** 2016. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785-794.

10. *Mapping the denominator: spatial demography in the measurement of progress.* **Tatem, A.J.** 2014, International Health, pp. 153-155.

11. **Tatem, A.J. and Bird, T.J.** *High resolution 'bottom-up' population mapping.* [Online] 2026. https://www.worldpop.org/current-projects/completed-projects/high-resolution-bottom-up-population-mapping/.

12. *Spatially disaggregated population estimates in the absence of national population and housing census data.* **Wardrop, N.A., et al.** 14, 2018, Proceedings of the National Academy of Sciences of the United States of America, Vol. 115.

13. *On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation.* **Cawley, G.C. and Talbot, N.L.C.** 2010, Journal of Machine Learning Research, Vol. 11, pp. 2079-2107.

14. *A study of cross-validation and bootstrap for accuracy estimation and model selection.* **Kohavi, R.** 1995. IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence. Vol. 2, pp. 1137-1143.

15. *Optuna: A Next-generation Hyperparameter Optimization Framework.* **Akiba, T., et al.** 2019. Proceedings of the 25th International Conference on Knowledge Discovery and Data Mining.

16. *On the distribution of a variate whose logarithm is normally distributed.* **Finney, D. J.** 2, s.l. : Supplement to the Journal of the Royal Statistical Society, 1941, Vol. 7.

17. *Scikit-learn: Machine Learning in Python.* **Pedregosa, F., et al.** 2011, Journal of Machine Learning Research, pp. 2825-2830.

18. **Office for National Statistics.** Measuring uncertainty in international migration estimates. [Online] 2023. https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/measuringuncertaintyininternationalmigrationestimates.

19. **Apache.** Machine Learning Library (MLlib) Guide. [Online] 2026. https://spark.apache.org/docs/latest/ml-guide.html.

20. **Office for National Statistics.** Small Area Population Estimates (SAPE) Evaluation: Report on Accuracy Compared to Results of the 2011 Census. [Online] 2016.

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/smallareapopulationestimatessapeevaluationreportonaccuracycomparedtoresultsofthe2011census.

Annex

A1. Features

Table A1 *Feature data input to the models. Numerical features are normalised by land area, and transformed to natural log scale when right-skewed. Source ID corresponds to Table 1. Final three columns describe which models use which features.*

Name	Description	Theme	Source ID	No. of features	LSOA	Grid	Drift
SPD population counts	Counts of SPD records	Admin population	11	1	Yes	Yes	No
PDS population counts	Counts of PDS records	Admin population	13	1	Yes	No	2021 only
PR population counts	Counts of PR records	Admin population	12	1	No	No	2011 only
Aggregate occupancy ratio (PDS)	Ratio of admin-based population counts from PDS to property counts from AddressBase	Admin population & Building/ Property	6, 13	1	Yes	No	2021 only
Aggregate occupancy ratio (SPD)	Ratio of admin-based population counts from SPD to property counts from AddressBase	Admin population & Building/ Property	6, 10	1	Yes	Yes	No
Aggregate occupancy ratio (PR)	Ratio of admin-based population counts from PR to property counts from AddressBase	Admin population & Building/ Property	6, 12	1	No	No	2011 only
AddressBase property counts and proportions	Counts and proportions of various types of residential properties and establishments as defined by the AddressBase property classification scheme	Building/ Property	6	248	Yes	Yes	Yes
Linked building features	Various summary statistics of building properties e.g. modal property construction period, mean floor area	Building/ Property	6, 7, 9	42	Yes	Yes	Yes
Electricity	Domestic electricity consumption and number of meters	Energy	3	2	Yes	Yes	Yes
Gas	Domestic electricity consumption and number of meters	Energy	3	2	Yes	Yes	Yes
Night-time light intensity	Artificial night-time light intensity	Environmental	1	2	Yes	Yes	Yes
Air pollution	Background air pollution levels	Environmental	2	7	Yes	Yes	Yes
Land cover	physical material of land surface (e.g. grassland, woodland, water, urban, etc)	Environmental	5	21	Yes	Yes	Yes
Land use classifications	Area of land associated with different land uses, i.e., water, farming, woodland	Environmental	9	9	Yes	Yes	No
Rural urban classification	A statistical classification for categorising geographies as either rural or urban (RUC2011 and RUC2021)	Infrastructure and area classification	4	10	Yes	Yes	Yes
Street network characteristics	Various summary statistics of street and road networks e.g. Road length by highway type, number of intersections	Infrastructure and area classification	8	9	Yes	Yes	Yes

A2 Traditional methods

A2.1 Ratio change

Ratio change uses scaling factors calculated from the annual change in LSOA population counts between reference years in an administrative source to roll forward Census SAPE across the intercensal period, constrained to annual LA MYE produced from the cohort component method.

We include results from one implementation of the ratio change method as a baseline reference to recreate the method currently used for production (albeit at the aggregate LSOA level rather than by age and sex). This implementation uses Patient Register (PR) data to calculate the 2011-2020 LSOA scaling factors, PDS data for 2021 LSOA scaling factors, and rolled-forward LA MYE for constraining in the intercensal period (not rebased using Census 2021).

A2.2 Benchmarking

Benchmarking uses scaling factors calculated from the ratio of LA population estimates to admin-based LA population counts within the same reference year to scale known LA estimates down to the LSOA level.

We include results from one implementation of the benchmarking method as the leading benchmarking method developed by MQD colleagues. This implementation disaggregates LA estimates using PR LSOA counts for 2011, and PDS LSOA counts for 2021.

A3. Additional processing

A3.1 Aggregation and disaggregation for grid cells

Population size predictions at the grid level are aggregated to the LSOA level before being constrained to LA totals. Most grid cells exist entirely within one LSOA and so can be simply summed by LSOA, but some cells are intersected by LSOA boundaries and so their predictions must first be disaggregated further into grid cell ‘segments’ which are then summed separately for the parent LSOAs. We disaggregate a grid cell’s prediction into its segments using the number of residential buildings in a segment as a proportion of the total number of residential buildings in the cell.

A3.2 Feature disaggregation for grid cells

Features that are only available at a resolution above the grid level (e.g. gas consumption) must either be disaggregated to grid cells or cells must inherit the less granular values from higher level geographies as-is.

When inheriting values as-is, some grid cells are intersected by the boundaries of statistical geographies and so could inherit one of two or more possible values. We choose a single value to inherit by picking the value from the geography that has the most land area in the cell.

When disaggregating values down to the grid level we use the number of residential buildings in each grid cell as a proportion of the total number of residential buildings in the higher-level geography to calculate proportions of the higher-level total value to disaggregate into each cell. This approach is currently only applied to energy consumption data published at the LSOA level.

A3.3 Processing of remote sensing imagery data

For night-time light intensity features: for the LSOA models, the satellite imagery is overlaid with the geography. Aggregate statistics of the pixel values of all pixels contained by the relevant geography are returned. For grid models, the pixel value corresponding to the grid centroid position is returned. The approach does not consider grid cells that span more than one pixel in the raster. This is because the grid cells of 100m are small relative to the resolution of the lights raster data (~300x500m). Therefore, any border effects are expected to be marginal. This assumes that the 'pixel' the grid centroid falls into represents most of a grid cell, even if not all.

For air pollution features: for LSOA models, a spatial join is performed between geography and the air quality grid centroids, with the mean air quality returned from all matched air quality grids. For grid models, a spatial join is performed between the grid centroids and the nearest air quality grid centroid, with the air quality values assigned accordingly. Some grid cells were found to have significant distances (>1 km) from their nearest air quality grid cell. These are typically found in coastal areas and likely result from the air quality grid's relatively coarse resolution and the way this maps to coastlines. In all cases the closest air quality grid is used.

A4. Comparison of full feature and reduced feature LSOA models

Table A2 *Absolute relative bias of the LSOA estimates for the full and reduced LSOA model feature sets at 0-year lag (both trained on Census 2021 and tested at mid-year 2021 against Census-based mid-year estimates).*

Method	Version	Census lag (years)	P_{25} ARB	P_{50} ARB	P_{75} ARB	Max ARB
Geospatial ML	$LSOA_{XGB}$ (full)	0	1.22	2.57	4.64	143.10
	$LSOA_{XGB}$ (reduced)	0	1.29	2.76	4.97	162.70