Geospatial methods for Small Area Population Estimates: proof of concept

Key asks

1. Does the strategy for developing geospatial approaches for small area population estimation seem viable?
2. Does the panel have any suggestions on other geospatial research/methods/data that we should be aware of?
    a. Does the panel have any thoughts on how we can best incorporate address occupancy into our model covariates?
3. Are there any other applications the panel think may be suitable for the geospatial applications outlined in this paper?

1. Introduction to Small Area Population Estimation

This paper reports on an initial proof of concept exercise on the use of geospatial approaches for producing small area estimates of the household population in England and Wales. The Office for National Statistics (ONS) has well developed and documented methods for producing population estimates in England and Wales at local authority district (LAD) level (see below).

There is also a strong user need for robust population estimates below LAD level, for instance Middle layer Super Output Areas (MSOAs), Lower layer Super Output Areas (LSOAs) and Output Areas (OAs). For example, local authorities and councils need population statistics at a local level so that they can understand their local communities and changes in those communities. This will enable them to make well informed decisions about local service provision and to target resources and interventions effectively. We at the ONS are investigating options for improving methods for small area population estimates (see our methodology article Small Area Population Estimates in the transformed population estimation system).

One potential approach is to make use of geospatial methods and data sources. The essence of the geospatial approaches is to use data sources of a very high spatial resolution, capturing geographical, demographic and socio-economic data, that may indicate population size at small area level. For instance, use of satellite imagery to provide detailed classifications of land use and cover, or to produce fine-grained maps of night-time light radiance.

In addition, record-level data sources can be used to identify “where things are”, for instance pin-pointing the locations of buildings to create a detailed picture of an areas building stock and footprint. This type of data can also pinpoint amenities that the population interact with, such as hospitals, schools, shops, and so on. The richness of this anonymised information may offer a unique insight into resolving some challenges for estimating small area populations in England and Wales.

Transformed population system for England and Wales

Geospatial approaches for small area population estimation contributes to the wider on-going ONS transformation of the population and migration statistics system for England and Wales (see our article on How population and migration estimates are evolving for more information). The transformation aims to make the most of several available data sources, focusing on administrative data to produce high quality population statistics.

Methods such as the Dynamic Population Model (DPM) have been developed as a means to provide more frequent and timely population statistics that meets user needs. The DPM uses a model-based cohort component method incorporating multiple administrative and survey data sources, including Statistical Population Datasets (SPDs) as an admin-based population stock, information on births, deaths and migration, as well as census data (see our Dynamic population model, improvements to data sources and methodology article from December 2023).

The DPM has produced high quality Admin-Based Population Estimates (ABPEs) at LAD level and we aim for these to become the official mid-year population estimates in 2025. We will gather feedback from users, including local authorities on the new approach in autumn 2024, so we can draw on local insight as we improve the estimates. This user feedback will form part of the criteria to support the decision on when the ABPEs will become the official mid-year population estimates. The DPM is not currently designed to produce estimates below LAD level, however.

Current population system for England and Wales

Currently, a decennial census provides the most accurate population statistics at national and sub-national geographies in England and Wales including LAs, MSOAs, LSOAs and OAs. Between census years, mid-year estimates are produced using ratio change methods for MSOAs and LSOAs and apportionment methods for OAs.

These methods roll forward census estimates using administrative data as a proxy for changes in the population (see our Methodology note on production of population estimates for details on the ratio change and apportionment methods). While these methods produce, on average, respectable measures of bias against known population counts, there are issues with several small areas having much larger bias (see our Small Area Population Estimates in the transformed population estimation system methodology article). Estimates from ratio change and apportionment are also prone to “drift” over the decade between census years, meaning small area population estimates become less accurate the further away from the previous census year. This drift is highlighted in our Rebasing of mid-year population estimates following Census 2021, England and Wales bulletin.

We think the geospatial approach shows good potential as an alternative as it relates information about the infrastructure on the ground to population density and does not

necessarily rely on the census. Geospatial information is also available at more frequent time points across the decade, potentially helping to minimise the “drift” of population estimates the further away from the census year.

2. Geospatial methods and data

Geospatial approaches for small area estimation make use of rich, spatially refined data sources that potentially offer a unique insight into understanding the make-up and dynamics of the population. For example, approaches such as Population 24/7 demonstrate how geospatial data can be used to understand how the population of an area changes throughout a day, for more details see the Developing a flexible framework for spatiotemporal population modelling article, published in the Taylor and Francis online journal. In a similar manner, this paper considers using geospatial information to produce small area population counts of the population, as shown in our United Kingdom population mid-year estimate time series.

Typically, geospatial data are used in two types of methodology to produce population estimates for small areas: “bottom-up” modelling and “top-down” disaggregation. A detailed review of these methods can be found in the Spatially disaggregated population estimates in the absence of national population and housing census data article, published in the PNAS online journal.

Top-down approaches, or disaggregation methods, take known population totals at higher levels of geography and disaggregate these to more granular levels of geography. Geospatial information captured at the small area level is used to inform this disaggregation. Several disaggregation methods exist (see the Disaggregating population data article, published in the Taylor and Francis online journal for a detailed overview). One approach is to use a random forest dasymetric mapping method as developed by WorldPop. The method is built from random forest algorithms that model the relationship between the population and geospatial information at a level of geography where the population estimate is of sufficient quality (for example, LAD level).

The algorithm, which offers relatively good predictive performance from minimal model tuning, as explained in the Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data article published in the Plos One online journal, then uses geospatial information at small area geographies to make a population prediction for these small areas. The predictions at small area level are then used as a weighting layer to redistribute the higher-level (LAD) population totals. The outcome is a set of small area population counts that are calibrated to a high-quality population benchmark at higher levels of geography.

The random forest disaggregation approach assumes that the relationships between population and geospatial data at the aggregated (LAD) level are similar to the relationships at the disaggregated (LSOA) level. Furthermore, the random forest method is not able to extrapolate, or predict, beyond the limits of the data set in the training model. Consequently, we are likely to expect poorer predictions at the

extreme ends of the population density distribution at LSOA level, owing to the vastly different spatial scales from LAD to LSOA level.

In contrast, bottom-up methods focus on the relationships between population and geospatial information at or below the level of geography of interest. Of course, the challenge with this approach is deriving population data at the small area of interest to use in fitting, or training, the models. Conventional bottom-up approaches tend to make use of data sources capturing the population for a smaller number of areas, typically measured from a survey. The relationships between the small area population and geospatial information for the sampled areas is used to then make a population prediction for the "out-of-sample" areas.

The critical challenge for the bottom-up methods is access to a quality, small scale survey from which population estimates can be derived for a selection of areas across England and Wales. An alternative would be to consider the most recent census to provide the population data at small area level. In this approach, geospatial data are modelled against the census data at the census reference point (March 2021). Geospatial data from later time points can then be used to predict population for time points in non-census years. This approach assumes that the relationships between the population and geospatial data at the census reference point are maintained over time. In general, this is more likely to hold true when estimating closer to the census. However, the 2021 Census was taken during the coronavirus (COVID-19) pandemic lockdown conditions, so the relationships identified at that time may not be fully representative of those when lockdown conditions had ended.

A final consideration for the bottom-up modelling approach is the geographical level to model at. One option is to use geospatial data derived directly at the small area level of interest, such as LSOAs. However, an alternative option explored in this paper was to make use of geospatial information at very fine-grained levels of geography, specifically grid squares. In the geospatial literature, data are summarised at grid squares, typically between 100m to 1km squared, with all grids being of equal size. Gridding geospatial information is expected to better capture the properties of geospatial data, allowing for more robust relationships between geospatial information and population to be established (see the Disaggregating population data for assessing progress of SDGs: methods and applications article, published in the Taylor and Francis online journal for more detail).

In contrast, administrative/statistical boundaries, for example OAs, LSOAs and MSOAs, are based on population size rather than physical size (see our Census 2021 geographies methodology). At these boundaries, the physical size of areas will vary and are much larger than the typical grid area used for population estimation. Summarising geospatial data at these levels of geography may not capture the detail in the geospatial data and risk the granular nuances being averaged out.

In this paper we explore three approaches to estimating the household population – one top-down approach and two bottom-up approaches:

1. A top-down approach to disaggregate the LAD census population to LSOA level.

2. Use geospatial data at LSOA level to directly estimate at this level

3. Use gridded geospatial information to produce 100m square population estimates, which are aggregated up to LSOA level.

We outline our approach in more detail in Section 3: Strategy. In Section 4: Methods and models, we describe the different top-down and bottom-up approaches in more detail. Section 5: Quality Measurement and validation, outlines our approach for measuring the bias of the population estimates as compared with the 2021 Census. Results from each method are presented in sections 6: Top-down results, 7: Bottom-up results at LSOA level, and 8: Bottom-up results at Grid level. Section 9: Discussion, and Section 10: Future recommendations, overviews the paper and provides an outline of next steps.

3. Strategy

Census data as a "source of truth"

For this proof-of-concept exercise, we make use of the 2021 Census population data to provide known population totals at LAD level to constrain estimates to. We also use the known census population estimates at LSOA level as a comparator for the small area estimates obtained from the geospatial models to calculate bias and assess the performance of the methods. The 2021 Census is used as a measure of the "true" population totals.

Crucially, we only counted the census population that, as of the census reference date, were living in households and not communal establishments. We did this because the geospatial covariates, outlined later, largely capture the population living in households. Consequently, attempting to estimate the communal establishment population with the currently available covariates will likely lead to less accurate estimates. Consequently, this paper focuses on modelling the population living in households only. Future work will consider approaches for estimating both the household and communal establishment population separately, in a similar manner to how the mid-year population statistics are estimated. For more details, see our Population estimates for England and Wales, mid-2023: methods guide.

Geospatial and covariate data sources

A variety of geospatial and administrative data sources were acquired to provide covariate information at LAD, LSOA and grid level. The following geospatial data sources were considered:

- ESRI's 10m land cover classification from sentinel-2 satellite imagery, measured at the mid-year for 2021
- ESRI's 500m night-time lights radiation from sentinel-2 satellite imagery, measured at the mid-year for 2021
- The Met Office's 1km climate variables, including sunlight, rainfall, wind speed, humidity, plus others, annual measures for 2021

- DEFRA's 1km air quality variables, including particulate matter, sulphur dioxide, nitrous (di)oxide, plus others, annual measures for 2021
- DEFRA's 50m risk of flooding from rivers and seas for England and Wales, measured as of November 2023
- Ordnance Survey's 50m terrain height points, measured as of November 2023

The quality of these data sources was assessed by acquiring metadata for each source, to understand how the data were collected and processed. Assessment of metadata and inspecting the data allowed us to confirm these sources were of sufficient quality for inclusion in this paper. Other sources, where we were not satisfied in either the way data were collected or processed, that compromised the accuracy and intent of what these sources were meant to measure, were not included.

We also examined the data sources with regards to how we expected the data to correspond to population in England and Wales. For instance, land cover classification contains 10-metre grids that identify "built-up" areas. This includes residential properties, but also non-residential properties, travel networks, industrial sites, and so on. Similarly, night-time lights radiance does not solely reflect the residential population at night, but will capture non-residential areas including non-residential buildings, air and ferry ports, and so on. Using only open-source data to predict the residential population across and England and Wales would likely to lead to very inaccurate estimates.

We also considered several other geospatial data sources, a mixture of open-source and those available to ONS, derived only at LAD and LSOA level, which we believe relate to the residential population at these levels:

- Anonymised and aggregated mobile network operator travel and location dataset – providing hourly population estimates used to estimate residential population and changes in population over time, at high spatial and temporal resolution
- the Open Street Map (OSM) – an open dataset used to derive measures of street network design that provide the framework for residential development; this provides counts of network nodes (intersections) and edges (streets) for both the entire network and residential streets, and could be substituted with similar data from Ordnance Survey
- the Rural-Urban Classification dataset – providing an LSOA level categorisation of rural and urban areas used to test whether different relationships exist between population and input covariates in different area types

The approach we took in this paper was not to rely just on typical geospatial data, but to also make use of administrative data sources available in the ONS that provide more direct counts of the population and housing than open-source geospatial data would, but do not provide sufficiently accurate distributions at small

area level alone. For example, administrative datasets are designed for administrative purposes rather than population estimation. However, the approach outlined in this paper aims to combine the strengths of various types of data to produce small area population estimates.

The administrative data sources provide records of information alongside the coordinates of where those records are located, meaning we could summarise these data sources at grid, LSOA and LAD levels of geography. The following data sources were considered:

- the ONS-derived dataset, Address Index, which integrates Ordnance Survey AddressBase Premium, Royal Mail Postcode Address File, and LAD gazetteers to provide a list of addressable objects across the country, which is used to provide residential addresses, points of interest (hospitals, schools, etc) and types of land (crops, parks, water, etc)
- the Valuation Office Agency (VOA), which provides characteristics about residential addresses, including total floor area, number of rooms, number of bedrooms

The Address Index aims to provide an accurate and comprehensive list of addressable objects across the country, using reliable data sources from the Ordnance Survey AddressBase, the Royal Mail Postcode Address File, and LAD level address information. As a single framework to capture addresses across the UK, Address Index can be used for many applications requiring accurate address data, including being used as an address framework for the 2021 Census. The quality of the Address Index is maintained with regular data supplies every 6 weeks, providing a timely source of data from which to base address data.

The comprehensive nature of the Address Index means we can link in other information about addresses using Unique Property Reference Numbers (UPRNs). One data source considered in this paper is data on address characteristics provided by the Valuation Office Agency (VOA). VOA was used in the 2021 Census to provide information on the number of rooms for households. However, VOA data does have some missing values after linking to Address Index data, around 5 to 6% for the number of rooms, number of bedrooms and floor area variables. In the census, missing values were addressed with robust imputation methods. Future work could consider similar methods to ensure these missing values are handled when linked to the Address Index.

Future work can consider using an SPD data source, in combination with address data that will give us information on the number of occupied addresses across England and Wales. Similar outputs on occupied addresses have been produced before with our Admin-Based Housing Stock (ABHS) dataset, though these data have only been produced for 2021 using older versions of the SPD. Providing such data on occupied addresses will, however, likely provide a useful variable for estimating the residential populations across England and Wales.

4. Methods and models

Top-down LAD to LSOA disaggregation

For this application, we modelled the relationship between covariate data and LAD level population living in households to estimate LSOA level population, using a random forest model. The models were trained on all LADs in England and Wales using 2021 Census population and covariate data. The coefficients from the model were used to predict population for all LSOAs in England and Wales, using corresponding covariate data. This approach is relevant for prediction of small area population, where population estimates for non-census years are only available at higher levels of geography.

Similar to the WorldPop approach, model estimates were in the form of the natural log population density at LSOA level, as population density is more consistent across different spatial scales than population counts. The logarithm transformation reshapes the response variable as a Gaussian distribution, which matches better with the distributions of covariates. The distributions of count-based input covariates, such as total residential address density, were also non-Gaussian. Therefore, we used the natural log of count-based input covariates for consistency.

The population density estimates at LSOA level were used to estimate an indicative population count. A weighting factor for each LSOA was derived from the ratio of this count to the sum of these counts across each LAD. The known LAD level population counts were then disaggregated using the weights to provide the final constrained population count estimates at the LSOA level. These estimates were then compared with the 2021 Census population living in households estimate as a “source of truth” (see Section 5: Quality measurement and validation).

We first developed a model using covariates detailed in the random forest dasymetric mapping method as developed by WorldPop. This provided a baseline for comparison against models developed using other geospatial covariates listed above. For consistency, between the three modelling approaches outlined in this paper, a subset of geospatial covariates was selected and used for each model approach, based on the strength of relationship measured by Pearson’s correlation (see Table 1). Further discussion on the findings from model exploration using alternative covariates is included in the discussion section.

For each model, we used the following covariates:

- Address Index residential address density
- Total Valuation Office Agency (VOA) floor area
- VOA number of beds density
- VOA number of rooms density
- ESRI night-time lights (VIIRS) intensity
- UK AIR particular matter (2.5g) concentration
- UK AIR number of days that maximum 8-hr ozone concentration is greater than 120 micrograms per cubic metre
- UK AIR nitrous oxide concentration micrograms per cubic metre

These covariates were shown to have the strongest relationships to the census population living in households, of the covariates that were available at LAD, LSOA and grid level. Although other covariates such as from anonymised and aggregated mobile network operator data were also strongly correlated with residential population, as this covariate was not available at grid level, this was not used in the models reported here.

Bottom-up model at LSOA level

For this application, we modelled the relationship between covariate data and LSOA level population living in households, using a random forest model. The models were trained on a randomly selected sample of LSOAs, with this model predicting populations for the remaining "out-of-sample" LSOAs. We only predict population for these "out-of-sample" LSOAs to avoid over-fitting of the model for validation purposes.

In this application, we randomly sampled 30% of LSOAs for the training data. We experimented with smaller samples, which may be more representative of typical survey sample sizes. Further discussion on the effects of sampling size is provided in the discussion section.

The covariates were modelled against the log population density at LSOA level (the same as for the top-down approach). To derive LSOA level estimates for the "out-of-sample" LSOAs, the population density estimates were used to estimate an indicative population count. Because of withholding the training data from the predictions, weighting factors were then derived from the ratio of these counts to their sum across the remaining "out-of-sample" LSOAs within each LAD, to constrain the known LAD level population counts to component LSOAs.

Bottom-up model at grid level

The appeal of using gridded data is that the properties of the geospatial information may be better captured at fine-grained levels of geography, for instance 100m square grids. Table 1 shows the strength of relationships between population and geospatial data are quite similar across LAD, LSOA and grid level (particularly when using population density at LAD and LSOA level). However, the potential benefit of modelling at grid level may come from nuances in the data being better captured, rather than at aggregated levels where such nuances may be lost in the aggregation process.

Table 1: Correlation coefficients between geospatial covariates and 2021 Census population at LAD, LSOA and grid level

Geography	Address Index address density	Address Index floor area density	Address Index bedroom density	Address Index room density	UKAIR pm2 concentration	UKAIR nitrous oxide concentration	UKAIR ozone concentration	ESRI night-time lights radiance

Grid population density	0.89	0.96	0.96	0.94	0.37	0.46	0.29	0.61
LSOA population density	0.92	0.92	0.95	0.95	0.95	0.86	0.67	0.89
LAD population density	0.98	0.98	0.98	0.99	0.46	0.32	0.32	0.39

Using gridded geospatial data, we modelled the relationship between the same covariates used in the other models with census population living in households, using a random forest model. As the amount of data fed into these models was substantially increased because we were modelling at 100m grid level, for computational reasons we ran the same random forest model but for each of the 9 English regions and Wales separately.

For each model, we again took a random 30% sample of grid cells to train the random forest model on, with predictions made for the remaining 70% of grids. For almost all regions, every LSOA was represented by having at least one grid square in the training and test data. For London, only one LSOA (Croydon 046A) did not have a grid square in the 70% test data, meaning no estimate was produced for this LSOA.

As 100m grid squares are equivalent to 1 hectare, population count is equivalent to population density, as used in the other model applications. The models used the log of the census population count as our dependent variable, with the exponential of the predictions giving us a population count estimate at 100m grid level. Like the other modelling approaches, grid estimates were used as weighting factors, derived from the sum of component 100m grid census population within each LAD, to disaggregate the known LAD level population counts to component 100m grids. Because the training data is withheld from the predictions, the disaggregation of the total population counts is from the sum of the remaining 100m grid census population counts (rather than the LAD total population).

We then aggregated the 100m grid cell population predictions to LSOA level. In most cases, 100m grids fitted wholly within an LSOA boundary. However, a small number of grids overlapped at least two LSOA boundaries. For this paper, we decided to split the 100m grid estimate into overlapping LSOAs based on the amount of space that each grid lay within the respective LSOA boundaries. For example, if a grid had 60% of its area in one LSOA and 40% in another, then the grid level estimates was split 60/40, respectively. This method produced a set of LSOA estimates that were compared with the 2021 Census population living in households estimates at LSOA level.

5. Quality measurement and validation

To assess the quality of LSOA level modelled estimates for each of the three approaches outlined above, we compared LSOA count estimates with Census 2021 count of residents living in households by calculating the Absolute Relative Bias (ARB) for each LSOA in England and Wales, which is defined as the absolute value of:

$$100*[(\text{estimate}-\text{true value})/\text{true value}]$$

From the LSOA level ARB measures, we take the median, 25th and 75th quantiles, and minimum and maximum ARB value to summarise overall performance of each method. We also conducted a deep dive into the twenty LSOAs with highest bias to understand characteristics of these areas to provide an understanding as to why these areas had extreme estimates. Table 2 (shown in Section 8: Results: Bottom-up model at Grid level) summarises the ARB measures across all four methods presented in this paper.

6. Results from top-down LAD to LSOA disaggregation

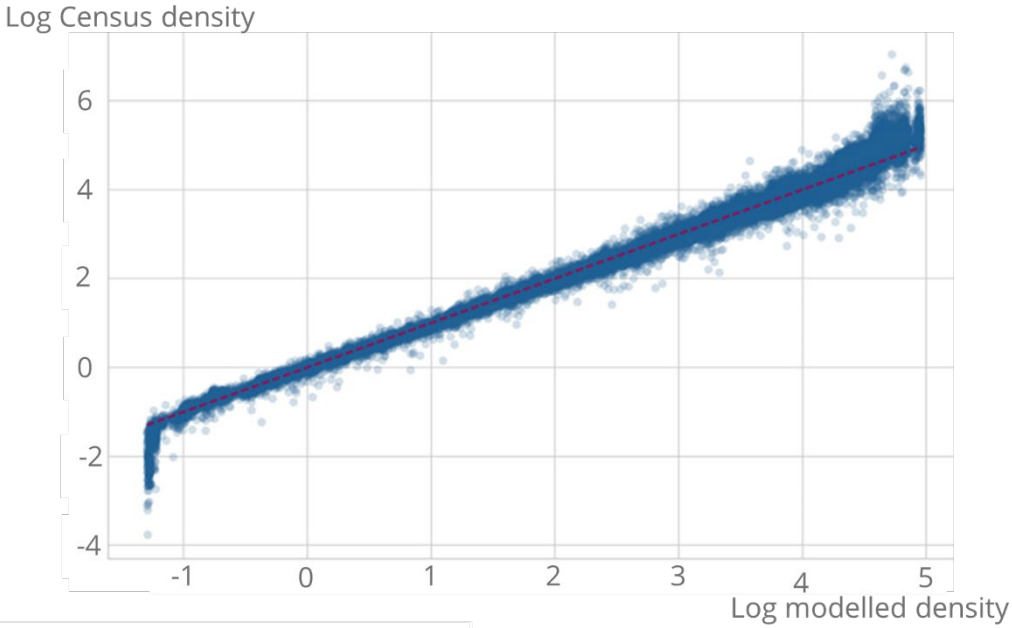
Random forest models can provide a measure of the relative importance of each input variable, or "feature", to the construction of the model. This "feature importance" (measured by the mean decrease in impurity) enables an approximate ranking of which covariates are given the most importance during the training of the model. Of the geospatial covariates used in this modelling application, covariates relating to residential buildings were by far the most important. Bedroom density had the highest feature importance, followed by residential address density, floor area and room density, respectively. Night-time lights and covariates relating to pollution, had only marginal importance.

When plotting model predictions against the Census 2021 population (see Figure 1), extrapolation issues inherent in random forest models were evident. Within the range of LAD level population density, predictions aligned reasonably well with 2021 Census population density. However, at the extreme ends of the population density distribution, low densities were overestimated, and high densities underestimated – predictions were stacked at the upper and lower limit of LAD level population density values. There was also a general pattern of increased residuals at higher population densities.

Figure 1

LAD to LSOA top-down method shows extrapolation issues at extreme ends of LSOA distribution

Census 2021 population density against model predicted population density on a natural log scale



Source: ONS

The median ARB for the top-down method was 8.21%, the ARB at the 25th quantile was 3.80% and at the 75th quantile was 14.90%. The maximum ARB was 781.62%, with 126 LSOAs having an ARB of over 100%. The majority of LSOAs (34,643 out of 35,672) had an ARB value less than 40%. The 20 LSOAs with the highest bias all had positive bias values, meaning that they were overestimates of the LSOA population.

There are differences in ARB values by region and area types. London was an outlier with a median ARB of 14.63%. Urban areas had higher median ARB values, but this is likely distorted by London LSOAs.

The very large ARB measures were likely because of the extrapolation issues inherent in the random forest model for this application. To overcome this, we experimented with a hybrid top-down modelling approach where 32 LSOAs, with population densities outside of the dynamic range of that from the LAs, were combined with the LAD data to form the training dataset.

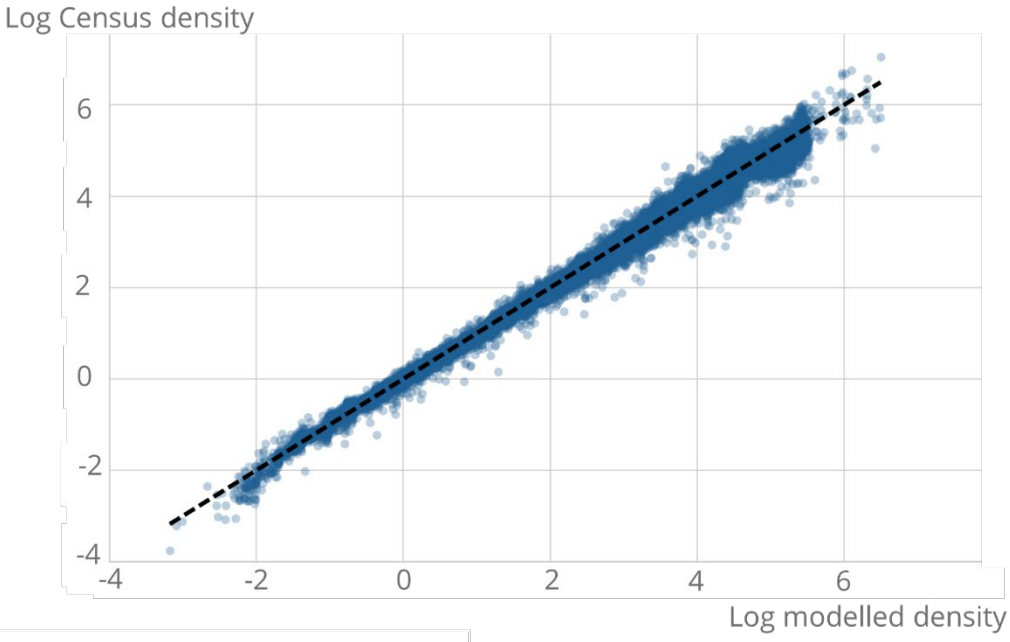
With the hybrid approach, there was a reordering of feature importance, although address-based measures remain dominant. There was also a small increase in the

relative importance of the night-time light covariate. The stacking of population density predictions at the upper and lower limit of LAD level population density values was removed by adding a small sample of LSOAs (see Figure 2), but the general pattern of increased residuals at higher population densities remained.

Figure 2

Top-down hybrid method demonstrates no extrapolation issues

Census 2021 population density against model predicted population density on a natural log scale



Source: ONS

The median ARB for the random forest top-down hybrid method was 7.84%, the ARB at the 25th quantile was 3.63% and the 75th quantile was 14.01%. The maximum ARB was 251.81%, with 48 LSOAs having an ARB of over 100%. The majority of LSOAs (35,036 of 35,672) had an ARB value less than 40%. The 20 LSOAs with the highest bias all had positive bias values, meaning they overestimated the LSOA population.

The hybrid method produced similar results to the LAD top-down method as LSOAs in major cities in England and Wales, including London, were the LSOAs with the highest bias. Rural LSOAs were estimated with smaller levels of bias, with a median ARB for rural LSOAs of 6.90%, compared with an urban median ARB of 8.50%. As in

the LAD trained model, London LSOAs were again outliers, but the median ARB dropped to 12.78%. The characteristics of the LSOAs with the highest bias include built-up areas with large numbers of shared houses and flats, with most households containing 1 to 2 individuals.

7. Results from bottom-up model at LSOA level

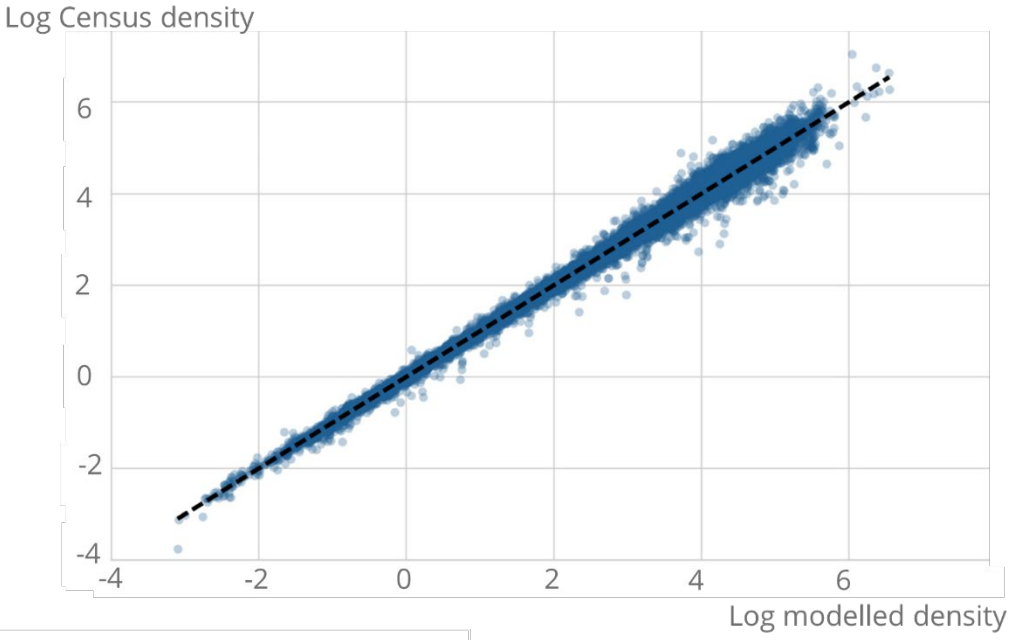
Of the geospatial covariates used in this modelling application, covariates relating to residential buildings were by far the most important. Bedroom density had by far the highest feature importance, followed by room density and address density. Floor area, night-time lights and covariates relating to pollution, had only marginal importance.

When plotting model LSOA counts against the Census 2021 density (see Figure 3), we see the model predicts the Census 2021 population generally well. The general pattern of increased residuals at higher population densities remains but was reduced.

Figure 3

Bottom-up LSOA model shows improved estimates from top-down models

Census 2021 population density against model predicted population density on a natural log scale



Source: ONS

The median ARB for the LSOA level method trained on 30% of the LSOAs was 6.24%, the ARB at the 25th quantile was 2.89% and at the 75th quantile was 11.20%. The maximum ARB was 230.13%, with 28 LSOAs having an ARB of over 100%. The majority of LSOAs (24,716 of 24,971) had an ARB value less than 40%. The 20 LSOAs with the highest bias all have positive bias values, meaning they overestimated the LSOA population.

As with the previous model application presented, there were differences in ARB values by region and area types. London was again an outlier, but the median ARB dropped further to 8.86%, while other regions were in the range of 5.38% to 6.42%.

We again see the pattern of the LSOAs with the highest bias located in London or major cities in England and Wales with high population densities, with rural LSOAs being estimated more accurately than urban LSOAs. Again, LSOAs with larger bias were characterised by large numbers of flats and shared houses, with most households containing 1 to 2 individuals in these LSOAs.

8. Results from bottom-up model at grid level

Population predictions at grid level, generally, were close to the grid-level census population living in households. Of the 1,285,306 grids with a modelled estimate, 96% (1,236,695) were within 20 people of the known census grid count living in households (see Figure 4, note that for illustrative purposes, any grid with a difference greater than -20 or 20 were grouped into the -20/20 band respectively). Only 6,469 grids had a modelled estimate that was greater than 50 people different from the true census grid count. It is important to acknowledge that some grid estimates were extremely different from the known census count, the maximum difference being 786 (grid underestimating the known population count). Generally, the more extreme grid-level estimates tended to underestimate the known population count, 1,038 grids had an estimate that was at least 100 people lower than the census grid count, with some grids predicting extremely small (fewer than 5 people) in areas where there were hundreds of usual residents living in households, according to the census. On the other hand, 321 grids had an estimate that was at least 100 people more than the census grid count.

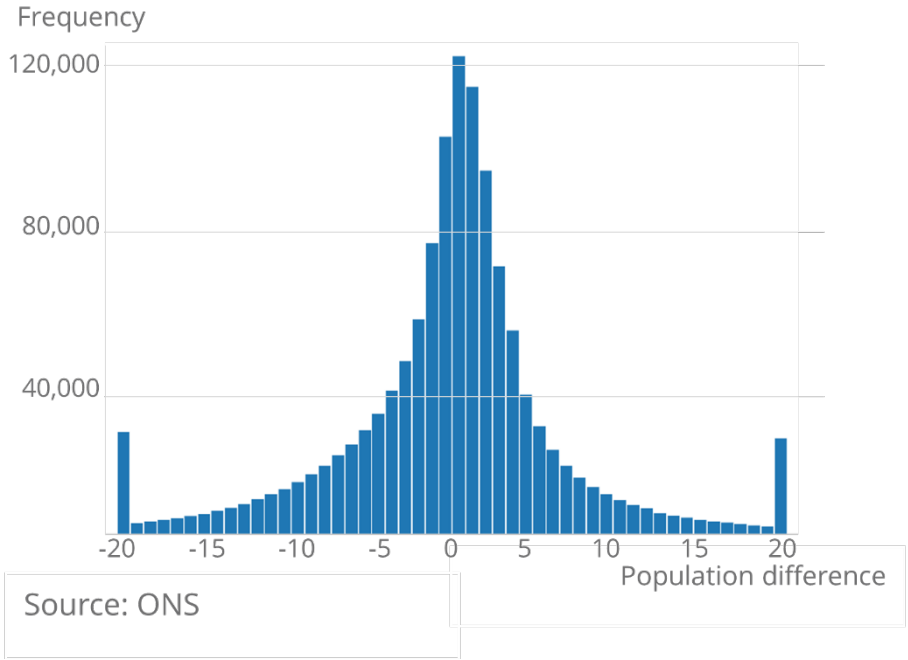
Of the 20 grids with the largest underestimation from the known census population, most grids were in London LSOAs, with one grid in Manchester. All grids show that the known census resident count outnumbered the number of addresses, implying these addresses were, as of the 2021 Census reference data, over-occupied at grid level.

For the 20 grids with the largest overestimation from the known census population, most of these were in London LSOAs, although some grids were located across the country, including Basildon, Leeds, Nottingham, and Ipswich. A common feature was that grids contained tower blocks of flats, with the census data indicating these grids had lower levels of address occupancy than models predicted.

Figure 4

Majority of 100m grid estimates within 20 of the Census 2021 population count living in households

Histogram of the difference between the Census 2021 population living in households to the modelled estimates, 100m grid, 2021



Source: ONS

As mentioned, random forest models were run at 100m grid level, separately for each English region and Wales. For six of the models (North East, North West, East Midlands, West Midlands, London and South East), bedroom and room density were the most important covariates, surprisingly total address and floor area had minimal importance alongside the night-time lights and air quality covariates. The other regions (Yorkshire, East of England, South West and Wales) showed only the room density variable having strongest importance.

At the LSOA level, the median ARB was 6.24%, the ARB at the 25th quantile was 2.89% and at the 75th quantile was 11.23%. The maximum ARB was 189.2%, with 8 LSOAs having an ARB of over 100%. The majority of LSOAs (35,421 out of 35,671) had an ARB value less than 40%. The 20 LSOAs with the highest bias are a mix of positive and negative bias values, meaning some LSOAs are overestimated while others are underestimated.

The estimates produced by this model follow the pattern of the LSOAs with the highest bias located in major cities with high population densities in England and Wales. London continued to be an outlier with a median ARB value of 8.26% compared with other regions ranging between 5.18% and 6.69%. Rural LSOAs are generally estimated more accurately than urban LSOAs. The LSOAs with high bias were characterised as built-up areas, with most households containing 1 to 2 individuals and large numbers of dwellings classified as flats or shared houses.

Table 2: Comparison of LSOA ARB measures against the 2021 Census for all 4 models

Model	Minimum	25th quantile	Median	75th quantile	Maximum
Top-down original	0	3.80	8.21	14.90	781.62
Top-down hybrid	0	3.63	7.84	14.01	251.81
Bottom-up LSOA	0	2.89	6.24	11.22	230.13
Bottom-up grid	0	2.88	6.24	11.23	189.24

9. Summary and discussion of methods

This paper has presented an initial proof of concept into the use of geospatial data and methods to produce small area population estimates for England and Wales. We have presented three models.

1. A top-down model to disaggregate LAD 2021 Census population living in households to LSOA level.
2. A bottom-up model predicting population living in households directly at LSOA level.
3. A bottom-up model predicting population living in households at grid level, then aggregated to LSOA level.

Across the three methods, median ARB measures ranged from 6% to 8%, which, at present, are not obviously better than those achieved using our current "baseline" methods (ratio change and benchmarking) described in our methodology article Small Area Population Estimates in the transformed population estimation system). It is important to stress that direct comparisons of bias are complicated by the different nature of the applications.

Baseline methods, such as the ratio change method, take the census as a base and roll the estimates forward each year, using the change in the population recorded in administrative sources for consecutive years as an indicator of change in the true population. On the other hand, the geospatial approaches outlined in this paper have not used a direct admin count of the population in the initial application: the focus here has been to assess the potential of the geospatial data in informing the small area distribution of the population.

This paper has demonstrated the potential for using geospatial sources to produce small area population estimates. There is potential to improve the geospatial models in future applications by improving the covariate data and trying alternative modelling approaches (please find more detail in Section 10: Future direction). An important part of this is to optimise the covariate selection and model fit and to include covariate data that represents the population in communal establishments in addition to the household populations.

We can also investigate whether the geospatial data can be used to supplement the distributions from the SPD, particularly in areas where the SPD data are not robust. A first step will be to do a more direct comparison of the estimates obtained using geospatial and baseline methods, particularly comparing the characteristics of the outlying areas with higher bias. In the longer term, a future aim could be to investigate the use of geospatial models for population estimates by age and sex and possibly other characteristics.

The top-down approach taken in this paper was inspired by the approach taken by WorldPop; a random forest model trained at LAD level was used to predict the population at LSOA level. We initially replicated the WorldPop method directly, using geospatial covariates used in this WorldPop top-down demonstration.

Disaggregating Census 2021 population data using these covariates produced poorer quality population estimates at LSOA level, with a median ARB of 40% and a maximum ARB of 1,980%. Using covariates developed for this paper improved the models, however we highlighted an important issue in that the random forest model could not extrapolate predicted population density values at the LSOA level beyond the range between the minimum and maximum population density observed at the LAD level.

This becomes an issue where there are differences in the range and distribution of input data. In this implementation of the modelling approach, this results in poor predictions at the upper and lower limits of population density because of the differences in the response variable and input covariates at different spatial scales, that is, i.e., population density and factors of influence of population density will be expected to be observed at much higher values at LSOA level than at LAD level. To better understand the impact of this issue and how it could be addressed, we have demonstrated a hybrid training approach with a small sample of LSOAs added to the training set, which removes this extrapolation issue and improves ARB values, particularly by reducing maximum ARBs.

However, this basic solution would not be advised in practice, because of leakage between the training data and the model evaluation predicted populations. Alternatively, a better approach would be to include synthetic data in the training data that represent the relationship between covariates and population density beyond that possible at LAD, but without including actual LSOA data points.

The top-down model as presented uses a common set of geospatial covariates that are available for each of the three different modelling approaches. We experimented with additional covariates, including average night-time population from anonymised and aggregated mobile network operator data, street network design derived from Open Street Map, and area classifications. Combinations of these covariates improve the model performance over the common set presented in this paper. Using the LAD trained approach, median ARB values are reduced to 6.9%, although maximum ARB values are not improved because of the previously mentioned issues with extrapolation.

Using the hybrid approach alongside these additional covariates, median ARB values are reduced to around 6.5% and maximum ARB values to as low as below 150%, although these values were achieved using different combinations of covariates. In models trained with these additional covariates, those derived from anonymised and aggregated mobile network operator data have the highest feature importance, however, these data come with certain caveats. Firstly, the data are weighted to census population to account for market share of the data provider.

This process aims to produce a dataset that is representative of the population, rather than of the data providers customer base. In addition, as the data are weighted to the Census population, there are circularity issues of using population data to estimate the population. Secondly, the data are adjusted at source to ensure issues of disclosure control are removed from the data and no individuals can be identified. As the data are provided at fine spatial granularity, this results in issues with aggregation, i.e., where records have been removed because of disclosure control, aggregated totals may contain discrepancies. A better understanding of the methodology used in the production of the data is required, working with the data providers to assess its suitability for generating population statistics. When anonymised and aggregated mobile network operator data is removed from models, residential address density measures become dominant.

As an alternative approach, we considered ‘bottom-up’ models where we model the LSOA population directly at this level. The LSOA level model application shows improvements over the LAD trained model. This is to be expected as because the models are trained on LSOAs, issues of extrapolation outside the range of LAD population density are removed. It is important to reiterate that these different applications are designed for different situations of known population statistics.

The LAD trained model is suitable where population is only known or estimated at the LAD level, while the LSOA trained model is more suited where there is a sample survey of population. In the latter example, it is more likely that a smaller than 30% sample would be available. We therefore also experimented with smaller samples which may be more realistic to a potential future data collection designed to sample a smaller number of areas.

The distribution of predicted population density is very similar whether trained using 5 (median equals 6.68%, maximum equals 237%), 10 (median equals 6.45%, maximum equals 197%) or 30% (median equals 6.24%, maximum equals 230%) of LSOAs and median ARB values are increased by only small amounts.

The third approach taken in this paper was a similar bottom-up method but to model estimates at 100m grid level. Academic work has outlined the potential benefits of using geospatial data at much more fine-grained levels of geography. The results from this approach, in terms of bias measures at LSOA level, are like the bottom-up LSOA approach, most likely because the strength of correlations between geospatial data and population at grid and LSOA level are similar.

At grid level, the modelled estimates were generally of a high accuracy when compared with the grid-level known census population. However, several grids

showed very extreme differences between modelled and known population estimates. Some of these differences were from estimates that were predicting incredibly low (fewer than 5 people) and high (more than double the known census estimate) estimates. These extreme differences likely caused some of the more extreme ARB measures at LSOA level. Future work will consider the covariate data sources and model set-up at the grid level to understand why the models were predicting such extreme estimates, given most grids were predicted to a high level of accuracy.

The grid-level models were somewhat consistent with the LSOA level models, where feature importance was typically higher for the bedroom and room densities. However, total address density and total floor area had minimal importance across the models. In addition, for some regions, covariates such as night-time lights radiance and nitrous oxide concentration had more importance (though not as strong as bedroom and room density). For this paper, we made a conscious effort to align the covariates used across the three models. However, it may be that each model will perform better with unique covariates. This will be considered in the next phases of this work.

We may expect the grid-based approach to have outperformed the LSOA level bottom-up approach given that, in the grid approach, we trained and predicted grid-level population estimates for each region separately. The LSOA model, however, did not have information about what region the LSOA was from. For completeness we re-ran the bottom-up LSOA level approach by running the models for each region separately, which led to similar performance in terms of ARB measures to the original LSOA bottom-up model.

Some error will be introduced in the process of aggregating from grid to LSOA, as we have split some grids that straddle LSOA boundaries proportionally based on the amount of area in each grid “segment”. It is unlikely that this is the most accurate approach for partitioning grids as it assumes that the population is evenly distributed within the grid, and in some cases, there may be more people living in a smaller section of the grid. Alternative approaches to splitting grid estimates could consider using address information, splitting grid estimates based on the number of addresses, property size, property type, and so on.

As expected, for both the LSOA and grid bottom-up models, the more training data fed into the models, the better performance. For example, when the data are trained on 70% of areas, median ARBs were around 4% with a maximum ARB less than 100. A larger training sample would be justifiable in a scenario where we model the relationship between geospatial and population data using Census 2021 data, and then use that model to predict populations for subsequent years. This approach, however, relies on the relationships established during 2021 to maintain over time. Should this assumption fail, we will likely see patterns of estimates drifting away from the known estimate, outlined in our Rebasing of mid-year population estimates following Census 2021, England and Wales bulletin.

10. Future direction

Through this work, we have identified several avenues of further work to continue developing the geospatial approach. An important step will be to continue the development of geospatial covariate data sources. In this paper, we have outlined the value in using address data to predict populations. However, more information about these addresses could be included. For instance, we could include the type of building that each address is (detached, semi-detached, block of flats, etc).

We could also look to include information about address occupancy, which we believe may address some of the issues outlined in Section 8: Results from Bottom-up model at Grid level, where for some grids address data and usual residence occupancy are not correlated. Using additional data sources such as the SPD, or utility data (gas, electric, water consumption) may highlight which addresses are occupied, and whether those occupants are part of the usually resident population.

Other geospatial data sources could be considered as useful predictors of small area populations. In addition to residential building density, other land use types are likely to be strong predictors of population density. For example, convenience retail, leisure venues, schools only exist where there is sufficient population to support them. Such data can be acquired through the Address Index, which we can use in a similar way to how we have mapped residential addresses in this proof of concept. We will also consider data such as transport accessibility, which will likely be a strong predictor of travel demand and therefore population density.

In the previous section of the paper, anonymised and aggregated mobile network operator data and associated caveats are discussed. While these data may not be appropriate for estimating population counts, they may be more suited to measuring population change over time. In addition, other datasets can be used to assess change at fine spatial granularity, such as the Land Registry transaction data, which can provide monthly information of new build residential sales by postcode. These types of frequently updated, high spatial granularity data could be used in several ways. One, as a sense check between two sets of population estimates to validate changes at the LSOA level over time. Another, as an input to measure month-on-month change from a baseline such as the Census. Further investigation and quality assurance of these datasets are required to assess their usage in producing population statistics.

In this paper we outlined that the top-down WorldPop method may not work sufficiently well for Office for National Statistics (ONS) contexts as the model cannot extrapolate beyond the training data. Instead, we could consider alternate modelling approaches that are better suited for ONS contexts. One possibility is to use agent-based modelling approaches. Agent Based Models are widespread in the field of transport and mobility modelling. These models estimate individual agents' behaviour, and their interactions with other agents and their built environment, such as resident and workplace location and travel between activities, such as shopping and leisure.

The principles of how agents are assigned to locations, such as home, workplace and activities, often referred to as facility sampling, could be applied to small area

population estimates. This could follow top-down disaggregation approaches, for example, assigning a known population at LAD level to LSOAs, grids, or even individual buildings, based on demographic characteristics of the population at LAD level and the characteristics of the lower-level geography or features.

Alternatively, a bottom-up approach could be applied, where population is estimated based on the characteristics of land use, such as building type, and known demographic composition of individual areas or area types. This approach may produce predictions with lower errors, as the relationships between population and geospatial data are modelled at finer spatial granularity, rather than assuming relationships are similar across England and Wales. This approach also has the benefit of being used for estimating hourly based population estimates, but this would require substantial further development.

The models presented show signs of spatial instability. When calculating ARB values by region, for LAD, LSOA and grid-level trained models, we tend to see higher ARB values in regions with metropolitan areas, such as London, the North West, and the West Midlands, and in areas with high population density more generally. These areas are characterised by large numbers of flats and shared houses, with most households having low occupancy, i.e., containing 1 to 2 individuals. Model predictions are distorted by these outliers, suggesting that a focus of improvement should be on how to better estimate population in higher density areas.

By its nature, a random forest model output will differ on each model run. We have tested the stability of models by running multiple models using the same input covariates. It is worth noting that median ARB values typically change at the second decimal place when re-run, but maximum ARB values can change more substantially, so these should be used with caution. Further work should consider how more stable model estimates are generated using the model applications presented in this paper.

As discussed, a substantial limitation with the adoption of a random forest model is that these models cannot extrapolate predictions beyond that of their training data. Therefore, exploration of alternative machine learning models should be considered. This could include experimenting with alternative regression models or, instead, ensemble models. Ensemble models combine multiple learning algorithms to make their predictions. Since any given learning model has strengths and weaknesses, using a combination mitigates the limiting factors of any given one (e.g., an inability to extrapolate). Such an approach may also potentially improve on the observed spatial instability (discussed above), as some constituent models may perform better in areas of high population density, while others perform better in areas of low population density (relative to our initial results using random forest).

To assess the suitability of these modelling applications to future population estimates, further work is required to understand the temporal stability of model predictions. For example, can a model trained on 2021 data, predict 2011 Census population to similar levels of accuracy. To achieve this, input covariates need to be available or approximated for 2011. This is feasible with the set of covariates presented in this paper, as underlying data used for covariates derived from the Address Index are available within the ONS going back to 2013 (and alternatives

available for earlier years from Ordnance Survey). Further work should focus on the availability of these covariates and the assessment of temporal stability of modelling approaches.

Future applications of this modelling approach can be used to predict the full sample of LSOAs where their covariate information has not been used to train the model, i.e., a situation where the model is trained using 2021 population and covariate data, to predict population for future years, using updated covariate data.