Office for National Statistics

Article

Developing our approach for producing admin-based population estimates, subnational analysis for England and Wales: 2011

A comparison of admin-based population estimates with the 2011 Census at small area level. The data contained in this report are Research Outputs.

Contact:
Ann Blake
pop.info@ons.gov.uk
+44 (0)1329 444661

Release date:
27 July 2020

Next release:
To be announced

Table of contents

1 . Disclaimer

These Research Outputs are not official statistics on the population, nor are they used in the underlying methods or assumptions in the production of official statistics. Rather, they are published as outputs from research into a methodology different to that currently used in the production of population and migration statistics. These outputs should not be used for policy- or decision-making.

2 . Main points

Our latest version of the admin-based population estimates (ABPE) generally showed lower estimates than the 2011 Census, suggesting that the design had been successful in its main aim of reducing patterns of over-coverage seen in earlier versions. It is possible however, that over-coverage still contributes to this net pattern, and this will have an impact on the reliability of estimates produced using a coverage adjustment process similar to the Census.

We have undertaken further analysis of the ABPE by comparing these with 2011 Census estimates and analysing the patterns in the differences.

As expected, most local authorities (LAs) follow the national pattern of lower ABPE populations than official estimates. Our analysis shows that some people may be missing or not captured in the right place. This is particularly prevalent in small areas where there are communal establishments and/or special population groups such as armed forces or students.

When comparing the 2011 ABPE V3.0 with Census estimates, we find the coverage profiles of LAs falls into four main groups and these are:

1. four LAs with much lower ABPEs (10% to 20%) than Census estimates across most age and sex combinations
2. 94 LAs with much higher ABPEs (5% to 15%) than Census estimates for 5- to -15-year-old males and females, lower ABPEs (5% to 30%) for 18- to 21-year-old males and females
3. 87 LAs with much lower ABPEs than Census estimates in almost all age groups, concentrated in working ages
4. 161 LAs with lower ABPEs than Census estimates in almost all age groups, but with less extreme values than (3)

These patterns support the idea that we are missing some population groups (underestimating) as well as allocating them to incorrect areas (both under- and overestimating). We attempt to unpick some of the reasons for these differences.

Areas with communal establishments (CEs) are harder to estimate in the ABPE. At the lowest level of geography, output areas (OAs), some OA ABPEs are very different from Census estimates. Our analysis shows these areas are likely to contain communal establishments such as military bases, prisons, or student halls. However, these groups are particularly difficult to enumerate in the Census.

The current ABPE method may not be capturing population groups living in CEs effectively; measuring population in CEs is challenging because of higher population churn and associated lags in updating administrative data sources. Our early research suggests we are missing some groups (foreign armed forces and their dependents), and capturing others but allocating them to a different address than our official estimates (students, home armed forces, care home residents and prison population). Other data sources might help us understand these differences and how best to adapt our rules to provide better quality ABPEs.

There is an increase in the difference between the ABPE and official estimates over time. This may be because of intercensal drift in the mid-year estimate (MYE), or underlying issues with data sources. We are working on a sources of error framework to further understand individual data sources, including how they change over time, to improve our inclusion rules.

3 . Things you need to know about this release

We are transforming the way we produce population and migration statistics to better meet the needs of our users, and to produce the best statistics from the best-available data. For information on this transformation see our overview of the transformation of the population and migration statistics system .

The analysis in this report advances the previous research we have undertaken to produce estimates on the size of the population using administrative data, previously known as a Statistical Population Dataset (SPD). Throughout this report we will refer to this approach as admin-based population estimates (ABPE). The previous iteration of the research will be denoted ABPE V2.0, while our most recent methodology will be referred to as ABPE V3.0.

This report shares initial results for subnational geographies based on the ABPE V3.0 approach for producing admin-based population estimates using activity-based rules. We recognise that more work is required to refine and develop the methodology for allocating addresses based on our understanding of the data sources.

This report is published alongside a further assessment of the national level coverage patterns for the ABPE V3.0 and a report to understand what measures of statistical uncertainty tell us about the ABPE V3.0 at local authority (LA) level. The analysis published will help us focus our next steps in refining our rules and addressing coverage patterns in our ABPEs.

As in our previous research, our methodology links anonymised person records on administrative datasets to construct admin-based population estimates. For information about our previous methodologies, please see SPD V1.0, SPD V2.0 and ABPE V3.0 methodology reports.

For further information on the data sources included in these Research Outputs see the data source overviews.

We welcome feedback on the quality, value or impact of using these figures if they were used in place of existing official statistics. Please contact us at pop.info@ons.gov.uk.

4 . Introduction

In June 2019, we published the results of a new admin-based population estimates (ABPE) activity-based [1] approach. We aimed to reduce over-coverage seen in previous ABPE versions, an artefact of people no longer being present in the population but having records remaining on admin data. We found ABPE V3.0 generally has net under-coverage compared with official population estimates as expected, as we had designed the method to only include those records with activity.

We previously outlined our intention to combine the ABPE with a population coverage survey (PCS) and an estimation method to correct for coverage errors and produce high-quality estimates. One option for estimation is the existing dual-system estimation (DSE) method used for the Census. In our last publication, we described how at a national level, the results appeared similar to 2011 Census counts before estimation. This would require an ABPE containing minimal over-coverage. Please see the national level coverage article.

We have since carried out analysis to understand whether the patterns seen at the national level apply across lower-level geographies, including local authority (LA) and output area (OA) levels. Our analyses compare the 2011 Census estimates with the 2011 ABPE V3.0, identifying patterns of difference by age and sex. Published alongside this article, the uncertainty measures for the ABPEs will help users further understand the differences between the outputs, taking into account the uncertainty inherent in the ABPE method. This analysis is important to help us identify improvements to our rules and methods for allocating records to the correct location.

ABPE V3.0 uses a variety of "activity" sources including health, tax, benefits and education data to determine whether a record should be included in the estimate for the population. The type of activity differs between sources and we want to understand how this might have an impact on ABPE V3.0 results. For example, a pseudonymised record may have activity in the Higher Education Statistics Agency (HESA), Patient Register (PR) and Department for Work and Pensions (DWP) Customer Information System (CIS), but different address information associated with each source, or the address may be out of date. This could lead to the placement of a record into a geography which does not accurately represent residency for that record. By comparing official and research estimates with one another at subnational geographies, we identified patterns in the differences, and our future work will use this information to further refine the ABPE V3.0 method.

This report outlines our subnational analysis starting by comparing 2011 ABPE V3.0 by LA and single year of age and sex with the 2011 Census estimates and looking at the common patterns across the LAs. We then look at what we have learned moving from ABPE V2.0 to ABPE V3.0 and comparing our 2016 ABPE V3.0 with mid-2016 population estimates. For the very small areas, we have used outlier analysis to understand what is driving the larger differences observed between the 2011 ABPE V3.0 and 2011 Census estimates. We then use this understanding to test our assumptions on how effectively the current ABPEs capture population groups living in communal establishments (CEs).

Notes for: Introduction

1. "Activity" can be defined as an individual interacting with an administrative system, for example, for National Insurance or tax purposes, when claiming a benefit, attending hospital or updating information on government systems in some other way. Only demographic information (such as name, date of birth and address) and dates of interaction are needed from such data sources to improve the coverage of our population estimates.

5 . Local authority (LA) level findings

Several patterns emerge when comparing ABPE V3.0 with Census estimates. For the majority of local authorities (LAs), a similar pattern to the national level analysis is seen, with lower estimates for working ages compared with Census estimates. This suggests that we are broadly meeting our objectives for removing over-coverage in our ABPE, as our estimation framework works to adjust for under-coverage.

LAs can be broadly categorised based on the profile of differences that occur. For example, areas may have similar amounts of difference across the same age and sex groupings. Overall, four groupings were identified (after removal of Isles of Scilly and City of London, which were outliers because of their small populations). These were:

1. four LAs with much lower ABPEs (10% to 20%) than Census estimates across most age and sex combinations – this is limited to three central London Boroughs and Forest Heath (Figures 1a and 1b)

2. 94 LAs with much higher ABPEs (5% to 15%) than Census estimates for 5- to 15-year-old males and females, lower ABPEs (5% to 30%) for 18- to 21-year-old males and females, and variable age groups containing higher ABPEs (Figures 1c and 1d)

3. 87 LAs with much lower ABPEs than Census estimates in almost all age groups, concentrated in working ages (Figures 1e and 1f)

4. 161 LAs with lower ABPEs than Census estimates in almost all age groups, concentrated in working ages, but with less extreme values than (3) (Figures 1g and 1h)

Figure 1: Coverage patterns vary by local authority (LA), age and sex and can be grouped into four clusters

Mean percentage difference between ABPE and Census by cluster, age and sex, England and Wales, 2011

Data download

We see differences in coverage patterns across younger working-age males and females. In the 20 to 30 years age groups, most LA ABPEs are between 0% and 20% lower than Census estimates for males. However, for some LAs the ABPEs are higher than Census estimates for young working-age females; this pattern appears concentrated in LAs with large universities. This may be a result of lag in administrative sources, or incorrect resolution of records with multiple addresses. We know from our previous research that females in their 20s and early 30s are more likely to reregister with a GP than males, and as a result we are more likely to include female than male records in the ABPE. Our record level coverage analysis suggests that for this age group, we may include more short-term residents from recent registrations on the Patient Register (PR) and in higher education.

Group 1 (Figures 1a and 1b) LAs have much lower estimates across multiple age and sex combinations compared with LAs in other groups; these are Kensington and Chelsea, Westminster, Camden and Forest Heath. Populations in these London areas may not be captured in ABPE V3.0 because they may have higher proportions of their populations attending private schools, receiving private healthcare, are self-employed or are not eligible to receive benefits, all of which are unlikely to appear in our administrative data sources. Alternatively, lack of timely interaction with admin data sources because of high rates of migration to an area may also explain the lower ABPEs.

The ABPEs also appear to be very low in Forest Heath across all ages, with the largest differences in males of working ages. Forest Heath is an area with a very large foreign armed forces (FAF) population, concentrated in two large military bases. Our further analysis suggests that we may be missing this population group (see Section 6).

In Group 2, the pattern is generally higher ABPE than the Census estimates for children, combined with lower estimates of 18-to 21-year-olds (Figures 1c and 1d). For children, the main sources of activity are school Census and Child Benefit data, along with small amounts of PR activity as children are reregistered at GPs after a move. Higher estimates for this age group may occur through incorrect address allocation. If a record has multiple sources, the ABPE may assign the record to the wrong address. This could lead to a record being counted (incorrectly) in one geography (higher ABPE estimates) and missing a record in another (lower ABPE estimates). Lower ABPEs for the 18- to 21-year-olds suggest that we may be incorrectly allocating some of these young adults to alternative geographies, or missing them from data sources altogether (if they are not studying or in work).

Groups 3 and 4 both behave as expected, with lower estimates than Census estimates across working ages (Figures 1e, 1f, 1g and 1h). The main difference between the two groups is the level of differences between the two estimates, with Group 4 displaying substantially lower ABPEs than Group 3, and (on average) more prevalence of higher ABPEs for children in Group 4 than Group 3.

What do we know about the quality of the ABPE over time?

These differences are exaggerated as we move further away from the Census and compare with mid-year estimates (MYEs) particularly in areas with universities, large cities, and several areas with high migration, where the ABPEs are higher. This may reflect the compound impact of lag in the administrative sources and high population churn in these areas. This may be a result of the ABPEs including records in the wrong places (as a result of lag and churn) and incorrectly allocating addresses.

Other reasons for the increase in the differences observed could be because of drift in the MYEs between the Censuses, making comparisons between the ABPEs and MYEs difficult to unpick. We have produced uncertainty measures for both the official MYEs and our ABPEs, and compared the two by LA, age and sex.

By combining our analysis in this report with our analysis based on our measures of uncertainty, we learn more about which LA, age and sex combinations are likely to represent differences because of uncertainty in the MYE, and which may be a result of error introduced into the ABPE.

Understanding statistical uncertainty and the admin-based population estimates has also been published today. Further research is needed to fully understand which components of the methods may lead to the difference. This will help us understand whether the differences that we have detected are a result of drift in the MYEs or issues with the ABPE V3.0 methodology and underlying data sources.

Understanding more about the quality of the underlying administrative data sources we use in the ABPEs and how the quality changes over time is also an important factor in improving our methods. Our analysis shows that changes in underlying data sources may have an impact on the quality of the ABPEs at the local level. For example, Enfield and Haringey have higher estimates across all ages by 1% to 3% compared with the official estimates as a result of increased Personal Demographics Service (PDS) activity. Further analysis shows that this is a result of using system updates as a sign of activity rather than a genuine registration, which results in erroneous record inclusion in the ABPEs. This reinforces the need to understand how real activity is reflected in administrative sources to be able to effectively develop inclusion rules. We have developed quality frameworks which will help our understanding of the underlying data sources – these are described in Section 6.

When comparing ABPE V3.0 with ABPE V2.0, we can see the impact using an activity-based approach has on reducing the over-coverage observed in V2.0 but with an increase in under-coverage as a result. This is favourable in the context of applying a duel system estimation (DSE) approach to adjusting the estimates for under-coverage, however, an important part of producing good quality subnational estimates is ensuring the placement of records to the correct geography. This is reliant on good quality address information being collected, and where there are multiple addresses for the same records across sources, a suitable method for resolving these conflicts. For ABPE V3.0 the allocation of records is based on the address information on the activity sources where there are conflicts, we use the address from the activity source, or the data source we believe to be more accurate. For some address types, and some population groups, this is more of a challenge. The next section looks at how well the ABPE V3.0 performs at output area level.

6 . Outlier detection to understand output areas (OA)

In our previous research on admin-based population estimates (ABPE) V2.0, we found that populations in some small areas behaved differently to the general trends seen at the local authority (LA) or national levels. These were areas that contained special populations[1] or communal establishments[2] (CEs). Both special populations and CEs represent large, regularly changing populations concentrated in small areas, which can lead to the build-up of error in administrative sources.

To understand the impact of measuring these populations using activity-based rules in ABPE V3.0, we analysed ABPE data for 181,408 OAs by quinary age group and sex. This analysis helps us understand more about what is contributing to the differences we see at the LA level. For example, we can determine areas with substantially lower ABPE and see where they co-locate with armed forces bases. It also allows us look further into the impact of missing records and how inaccurate allocation of addresses plays out at the small area level.

This dataset is extremely large, so we have used outlier detection (as we did for our ABPE V2.0 OA analysis) to identify the most extreme differences. Our analysis helps us to understand how many of our outliers contain a CE or special population, and whether there are different patterns in the differences we observe in ABPE V3.0 as a result. We will use our findings to further refine our methods to include special population groups and people living in CEs. The reduced bias at OA level should also result in improvements to the ABPEs at aggregate, local authority levels.

There are 591 OA (0.33%) extreme outliers for differences in total population estimates, while 8,970 (0.26%) age group and OA combinations were detected as outliers. These represent OAs or combinations of age groups and OAs that exhibit extreme differences between official and ABPE estimates (more than five standard deviations from the average difference). We have combined this outlier information with the location of CEs and special populations at the OA level, to help us understand where both occur together. Patterns of error associated with specific CE or special population types will help us to further refine the ABPE method.

Who are our special population groups?

The current mid-year estimates (MYEs) have inbuilt adjustments for three types of population that are not captured by our usual methods accounting for internal and international migration: home armed forces (HAF), foreign armed forces (FAF) and prisons. These are known as "special populations". In each case, we make an adjustment to our mid-year estimates, as we may fail to include them in the correct location, or in the case of FAF, at all.

The student population is included in the MYEs through their interactions with health providers and the Higher Education Statistics Agency (HESA), but we know that greater uncertainty is present in areas with large student populations. This is because many younger people (especially young males), do not regularly interact with health services. As a result, error accumulates in areas with high churn, where people regularly move into and out of the area but are not measured.

At present, we have made no adjustments in the ABPE V3.0 for these populations, as we have assumed that the majority would be included in our core data sources and therefore picked up using our activity-based rules at the national level. However, we are aware that the current method may place records into incorrect geographies when multiple addresses are present, or when address data is not up to date. The following analysis tests these assumptions using auxiliary data from care homes and boarding schools.

Notes for: Outlier detection to understand output areas (OA)

1. Special populations are specific population groups which are harder to measure as they do not appear on our administrative sources and if they do, they may appear at a different address.

2. Communal establishments are residential addresses that are not considered private households.

7 . How are special populations captured in our ABPEs?

Testing our assumptions

In designing the ABPE V3.0, we considered how each of these groups would be captured by the method. The assumptions for the inclusion of each record are shown in the table below:

Table 1: Shows our starting assumptions for the coverage of special population groups in the ABPEs England and Wales, 2011

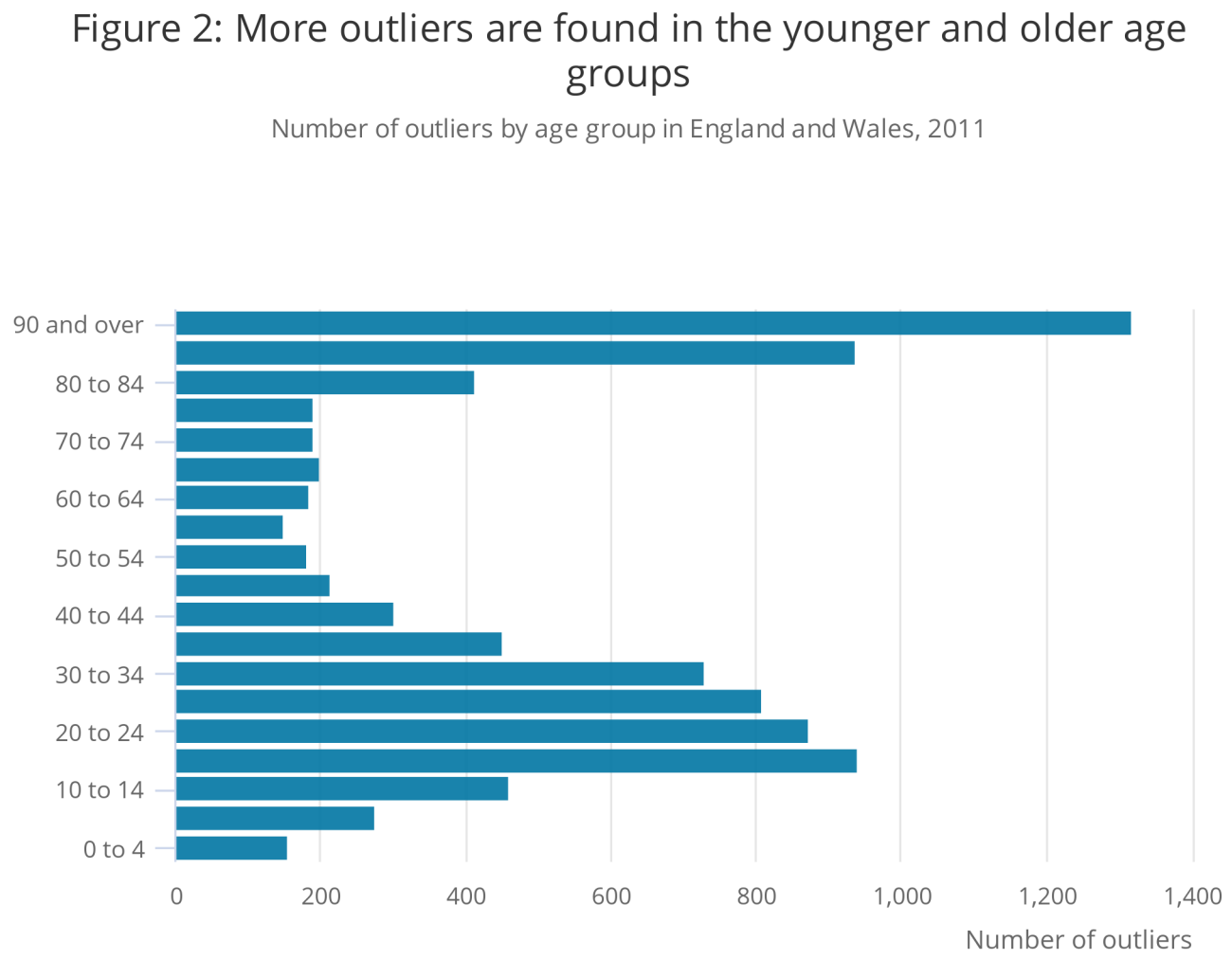
Population of interest	Assumptions for inclusion
Foreign armed forces (FAF)	Not included
FAF dependents	Not included
Home armed forces (HAF)	Address on base by Pay As You Earn (PAYE) and/or Personal Demographics Service (PDS) in later years
HAF dependents	Will be included at PAYE, school Census or Patient Register (PR) and/or PDS address
Students	Included through Higher Education Statistics Agency (HESA)
Prisoners	Included through PR and/or PDS
School boarders	Included through PR and/or PDS
Care home residents	Included through PR and/or PDS

Source: Office for National Statistics

What did we find?

Figure 2: More outliers are found in the younger and older age groups

Number of outliers by age group in England and Wales, 2011



Source: Office for National Statistics

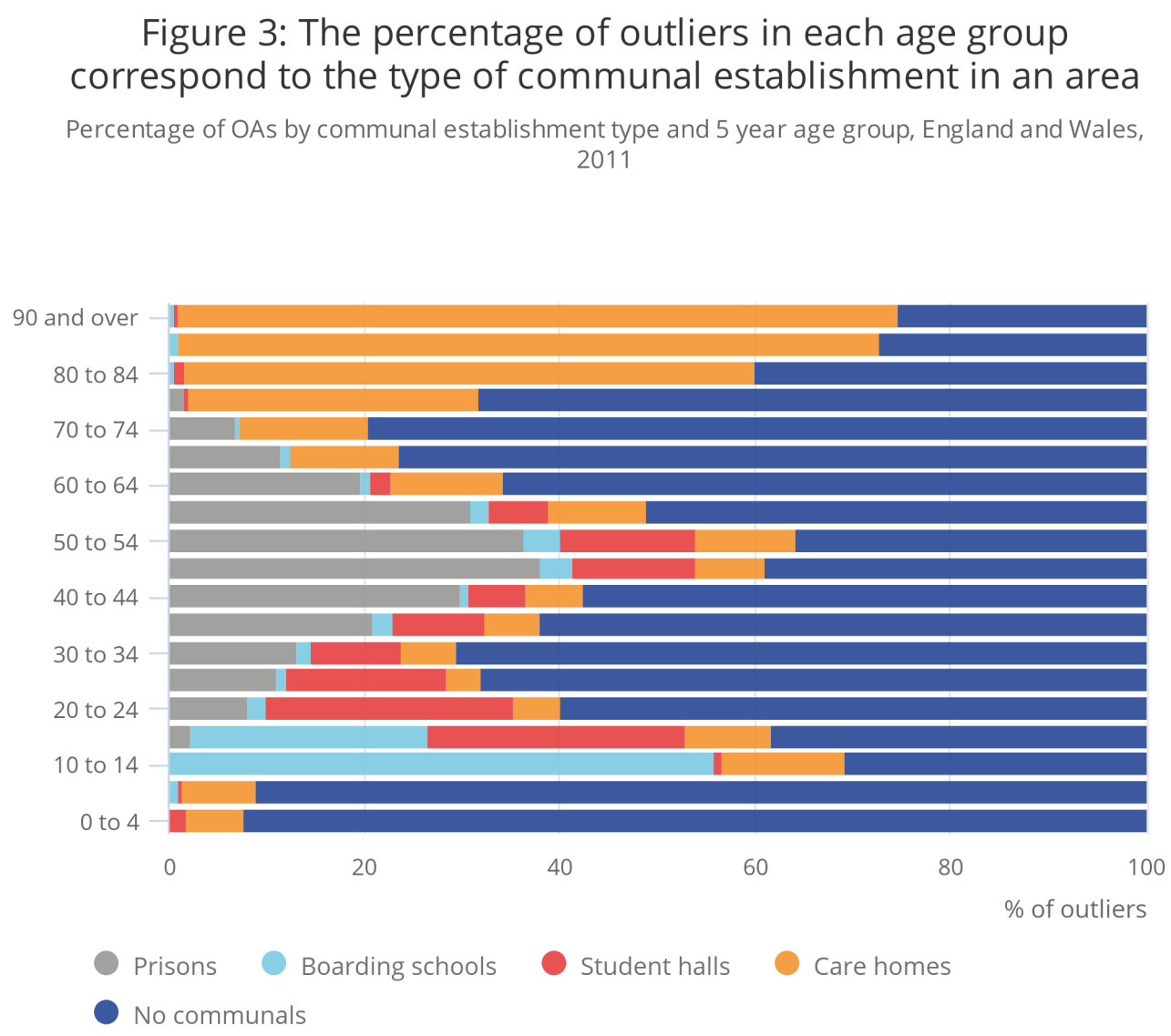
Notes:

1. Compares the number of OAs as outliers using 2011 admin-based population estimates (ABPE) and 2011 Census data.

Figure 2 shows the number of outlier output areas (OAs) by age group, with higher numbers at younger ages (15- to 19-year-olds) and especially higher at the older age groups, peaking for age 90 years and over with 1,318 outlier OAs. This section investigates whether the presence of different types of communal establishment (CE) may explain why these OAs are outliers.

Figure 3: The percentage of outliers in each age group correspond to the type of communal establishment in an area

Percentage of OAs by communal establishment type and 5 year age group, England and Wales, 2011



Source: Office for National Statistics

Figure 3 show the proportion of outliers for each age group that contains a specific type of CE (prisons, boarding schools, university halls or care homes). The important thing here is the age distribution of these differences is as we would expect, given the types of CEs observed.

Almost 60% of outliers in the school age group occur in OAs containing boarding schools, and approximately 25% of 15 to 19 years and 20 to 24 years age groups (roughly covering student ages) outliers occur in OAs containing student halls. We also note that OAs containing prisons co-locate with outliers at a range of working ages, peaking at 40 to 45 years, while care homes are in most outliers for age groups 80 to 84 years, 85 to 89 years and 90 years and over. Overall, this suggests that the ABPE performs less well in areas containing CEs and that we may be including CE residents at alternate addresses or not at all.

It is possible that some co-occurrences could be coincidental, but the patterns revealed are useful in identifying general trends. For example, for the 90 years and over age group, more than 70% of outliers contain a care home, 1% contain a boarding school and less than 1% contain a prison or a student hall. This indicates that for this age group, almost all extreme outliers are in the same geography as a care home. As presence of CEs is not mutually exclusive, the low numbers in the other CE categories may be included because some OAs contain both a care home and another CE type.

The following sections look at these CE types in more detail.

Prisons

Prisons are overrepresented in outliers for total population, with 69% of OAs containing prisons included. Age-group outlier analysis (Figure 3) shows a gradual increase in the proportion of outliers containing a prison for each working age group through to age 45 to 49 years. We note that for most working age groups at the national and LA level, the ABPE performs well, with small differences from the Census estimates.

The results shown in Figure 3 suggest that those OAs that do not perform as well for these age groups often contain a prison, and that these outliers are a result of substantially lower estimates in ABPE V3.0. There are three reasons why ABPEs may be underestimating the population in prisons: lack of up-to-date administrative data for prisoners (leading to their inclusion at an alternative address), over-counting in the Census, or no admin data footprint for prisoners at all.

We could learn more about these differences by looking at Ministry of Justice (MoJ) prisoner data to understand where they are accessing services from and therefore where best to include them. We intend to undertake further research on this in future.

Boarding schools

At the total population level, boarding schools are present in 10% of outliers, and less than 1% of non-outliers. Again, there are disproportionately more boarding schools in outlier OAs, suggesting that they affect the accuracy of the ABPE method. When assessed based on outliers by age group and OA, an obvious pattern arises.

Figure 3 shows that large proportions of OA outliers for 10- to 14-year-olds, and 15- to 19-year-olds contain boarding schools. Outliers for these age groups in OAs containing boarding schools are lower than official estimates suggesting that we may not be capturing records in the correct place, or that we may be missing records for school boarders altogether. For example, a child at a boarding school may have been registered with a local GP in their first year, but not subsequently. If there are no other sources of activity for that child, we would not include them in the ABPE after their first year at the school. Alternatively, they may appear on another administrative data source, such as Child Benefit. This would be associated with their home address, and so we would include them there. This would differ from their enumeration in the official estimates, where they would be included at their term-time address, leading to differences between official and ABPE estimates.

University halls

The presence of a university hall of residence is a proxy for high student-related churn. Student halls are present in 20% of total population OA outliers, with their presence in non-outlying OAs being less than 1%.

These differences are accentuated when the analysis is carried out by age group. This shows that for all outlying OAs for the 15 to 24 years age group, 25% contain a student hall (Figure 3).

Our secondary analysis looks at OAs containing large proportions of students based on Census-estimated student populations. This differs from the analysis above as it includes students enumerated in private residences, houses of multiple occupation, private student halls and university-owned student halls. By using these additional classifications, we identify that areas containing student halls often have ABPEs that are higher than official estimates, while areas containing students living in private accommodation have lower ABPEs.

We hypothesise that this is a result of lag in student records. Students who do not update their Higher Education Statistics Agency (HESA) address information from year to year will be included at the first address provided to HESA, leading to higher estimates in first year student residences. Subsequent moves may be missed, resulting in the underestimation of student populations elsewhere.

Residential care homes

The presence of a care home in an OA had very little impact on total population outliers but has a substantial impact on outliers in older age groups (Figure 3). We expect this is a result of the relative size of the population in the age groups, compared with the general population. Most outliers for age groups 80 to 84 years, 85 to 89 years, and 90 years and over contain at least one care home. Care home outliers for those aged 75 years and over are generally lower than official estimates (81%), but there are some higher ABPEs (19%). These may be a result of similar address conflicts to those we describe in the student analysis. For example, differences could occur when someone enters a care home but does not update the address on an associated pension. This would mean that their activity would be placed at their home address rather than in their new residence. In addition, we know that there was at least one isolated issue with lack of enumeration of care homes in a small area in the 2011 Census. This may explain the higher estimates in the ABPE noted in some areas, which may suggest that it can capture portions of these populations. However, further analysis of these results is required before we can draw conclusions.

It may be that for this age group, updates to the Personal Demographics Service (PDS) and Patient Register (PR) are deemed more accurate as they may more accurately reflect a person's address, however we will need to undertake further research to understand this.

Analysing home armed forces and their dependents

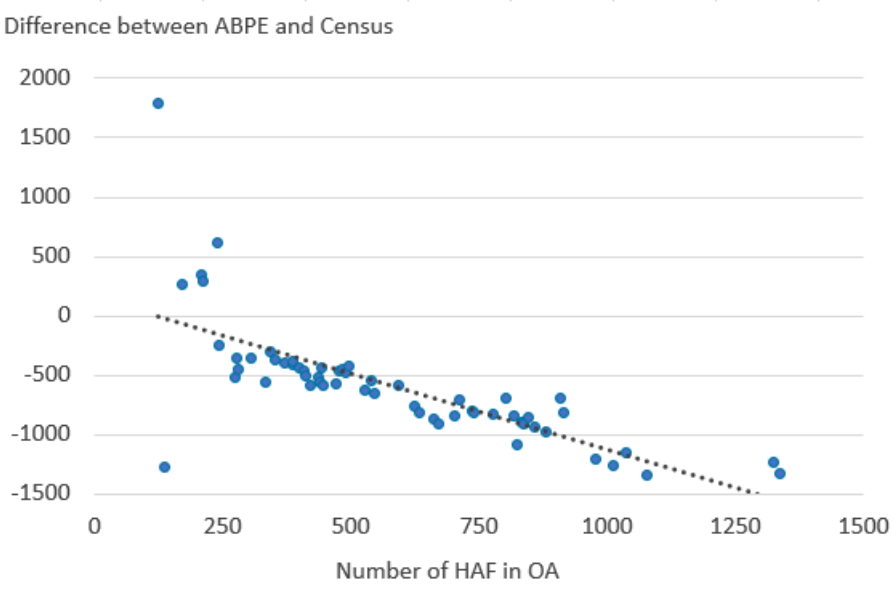
For the home armed forces (HAF), we use the size of OA HAF populations as a proxy for HAF bases in OAs. There are some differences in the way that HAF populations were enumerated during the Census compared with other special populations living in CEs. Service persons were counted at their family home unless no other residence was available, at which point they were enumerated on base. To understand how HAF populations are included in the ABPE, we correlated the size of the HAF population with the amount difference present between ABPE V3.0 and the Census estimates.

Our HAF analysis is focused on OA-level estimates of HAF population, rather than presence of military CEs or locations of military bases. For the majority of large HAF populations, the larger the HAF population, the more the ABPEs underestimate the OA population size when compared with Census estimates. The relationship between lower ABPEs and HAF population size is shown in Figure 4.

The Census was designed to count HAF at their family home if they had one (adjustments were made to the MYEs to account for this enumeration), which suggests that there is a real under-enumeration of HAF in the ABPE V3.0 at the base areas. The ABPE V3.0 method assumes that HAF records would be identified through Pay As You Earn (PAYE) data, placing them in their on base address where applicable. However, it appears that this may not be the case, and that HAF records are placed at alternative PAYE addresses (for example, private home addresses of the HAF members). We hope to carry out further research in future, to better understand how we can improve the estimation of the HAF population and their dependents in future iterations of the ABPE.

Figure 4: Higher HAF population is linked with the ABPE being lower than the Census

Outlier OAs by HAF population and difference between ABPE and Census in England and Wales, 2011



Source: Office for National Statistics

Analysing foreign armed forces and their dependents

The United States service persons and their families living on base have access to self-contained schools and healthcare and are not required to pay taxes in the UK. They are unlikely to leave an admin data footprint, and as such are unlikely to be included in ABPE V3.0. We may be able to use a similar aggregate adjustment approach to our existing MYE method for this population group.

Table 2 shows our main conclusions for our starting assumptions, based on our analyses above.

Table 2: Shows our findings against initial assumptions in Table 1 for the coverage of special population groups in the ABPEs
England and Wales, 2011

Population of interest	Assumptions for inclusion	Our findings
Foreign armed forces (FAF)	Not included	Evidence suggests true ABPEs are likely to exclude this group.
FAF dependents	Not included	Evidence suggests true ABPEs are likely to exclude this group.
Home armed forces (HAF)	Address on base by PAYE (and/or PDS in later years)	Evidence PAYE address is unlikely to be base address. ABPEs likely to underestimate this group.
HAF dependents	Will be included at PAYE, school Census or PR and/or PDS address	Less clear, further research required.
Students	Included through HESA	Some evidence students update activity in Year 1 and not after. ABPEs may not capture in the right place or pick up moves after study.
Prisoners	Included through PR or PDS	Some evidence prisoners may not be included and if they are, unlikely to be in right place. ABPEs likely to underestimate this group.
School boarders	Included through PR or PDS	If included, likely to be at home rather than term time address, unlike census.
Care home residents	Included through PR or PDS	Some evidence might remain registered at previous address rather than care home. ABPEs may not include residents in the right place.

Source: Office for National Statistics

Notes

1. Please note: The first column of this table was missing when published, it has now been added (27 July 2020, 15:29). Back to table

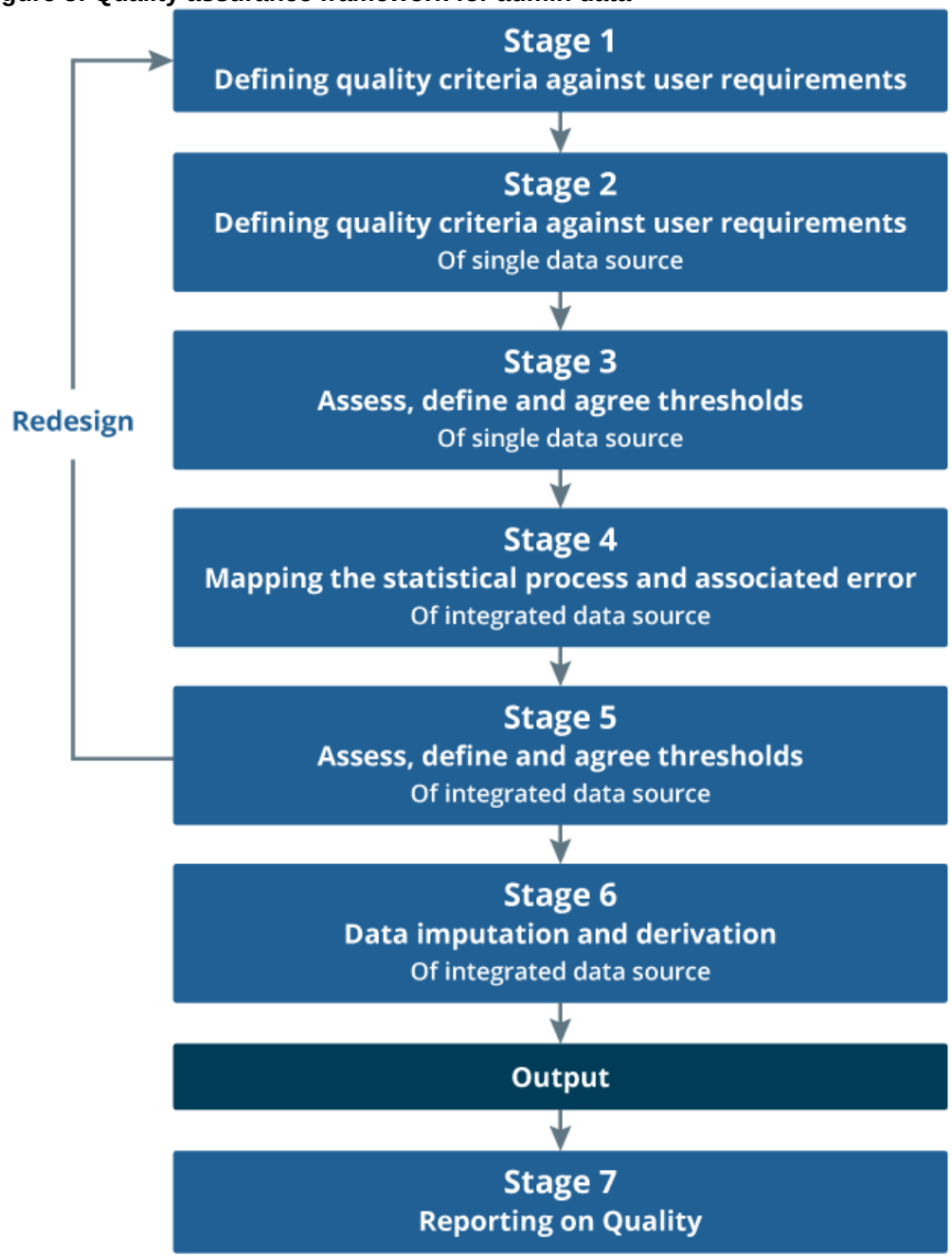
Understanding how these population groups interact with administrative systems is important for ensuring we are capturing them effectively. We are looking to acquire further demographic data specifically on these populations to support Census 2021, and we will look to explore how this information can be used to improve how we measure these groups in our future ABPEs.

8 . Developing a framework to understand the quality of the ABPEs

We are aware that some of the differences we have discussed are a result of varying quality in the underlying data sources that we use, lags in the timeliness of some data sources, lack of interaction with data sources by some demographic groups, and artefacts introduced during creation and processing of data. This may be because of data collection practices, cleaning of the data, processing of the data or through our methods.

In order to better understand this, we have developed a quality framework (Figure 5) to help the quality assessment of the administrative data sources feeding into the ABPE. The framework is underpinned by the Generic Statistical Business Process Model (GSBPM)[1], which ensures the individual data sources are assessed throughout their journey.

Figure 5: Quality assurance framework for admin data



Source: Office for National Statistics

Figure 5 outlines the quality framework using seven stages to ensure that administrative data are being quality assured against the user expectations in line with the quality dimensions: relevance, coherence, accuracy, interpretability, accessibility and timeliness.

In order to assure users of the quality of the ABPEs, it is paramount that we independently assess the quality of the individual sources against the design purpose. Each of these data sources have been assessed throughout their journey to determine what factors within the process may affect the quality of the outputs.

Errors, or potential errors, have been identified at each GSBPM phase and assessed against the design purpose to identify how it may impact the ABPE outputs.

Associated quality metrics have been developed and an action plan for further investigation and research has been outlined.

Quality metrics and investigation will initially be run against ABPE single-source historical data in order to ensure these measures are fit for purpose and further the ongoing development of the ABPE. We are developing ways to assess the quality of the ABPE output when all the data sources are combined, independent to comparing with official statistics (for example comparing admin data combined outputs with MYEs). Future consideration will also be given to how best to summarise and publish information on the quality of the ABPEs alongside future ABPE outputs.

As outlined in our next steps in the national level coverage paper, we will be looking at how we can use the longitudinal view of the data to understand how the population interacts with admin sources over time and what that can tell us about move into and out of the resident population and between areas. Alongside the quality framework described, the Error framework for longitudinally linked administrative sources will be particularly useful for understanding the accumulation of error in our linked data sources across time.

Notes for: Developing a framework to understand the quality of the ABPEs

1. The Generic Statistical Business Process Model (GSBPM) is a means to describe statistics production in a general and process-oriented way. It is used both within and between statistical offices as a common basis for work with statistics production in different ways, such as quality, efficiency, standardisation, and process orientation. It is used for all types of surveys, and "business" is not related to "business statistics" but refers to the statistical office, simply expressed. Taken from: https://ec.europa.eu/eurostat/cros/content/gsbpm-generic-statistical-business-process-model-theme_en

9 . Conclusions and future work

Our analysis shows that subnational patterns broadly follow those discussed in the national analysis, but that there is variation in the level of higher and lower ABPEs across age, sex and geography. For 2011, four main types of pattern were detected, as discussed in Section 5. We believe differences between Census estimates and ABPEs are likely to be a result of including records erroneously, not at all, or at an incorrect address, leading to higher and lower estimates in the ABPE compared with our official estimates. This is concentrated in areas with high inward and outward migration and may lead to the build-up of multiple records in a geography as a result of lag.

At the output area (OA) level, the impact of these problems is more evident, particularly for certain population groups and those people living in communal establishments. Several population groups may currently be omitted, or more likely counted at an alternative address to that in the Census. To better understand these differences, and to further improve the ABPE method, we will undertake more detailed analysis of records with multiple addresses.

In addition to our plans outlined in national level coverage (to improve our inclusion rules, improve our linkage and address our coverage gaps), to improve our subnational ABPEs there are two areas we need to focus on.

Firstly, allocating records to the correct address; our future work will look at improving the accuracy of address information on administrative sources and further developing the method for allocating records with multiple addresses to the most likely address. To develop these methods further we will:

- improve our address frames, including the classification of address types
- develop our tools to allocate addresses to administrative records

And secondly, ensuring special population groups are included at the correct address; alongside the work to improve the address allocation methods, we will also develop our approach for ensuring our ABPEs include the special population groups identified in this report. We will be looking at these groups and their interactions with data sources over time to better understand if and how they might already be included in our ABPEs. We will also research how data sources specific to special population groups can improve our estimates. This will inform any special adjustments we might need to make to improve the quality of our ABPEs.

We will also continue to progress our quality frameworks and, where possible, produce quality measures around our ABPEs. The uncertainty measures represent our first attempt at producing quality measures around our estimates. We will develop these measures further by better understanding the sources of errors across our statistical processes from the data collection journeys through to our estimation methods.