Article

Developing subnational multivariate income by ethnicity statistics from administrative data, England: tax year ending 2016

Feasibility research on an initial case study producing income by ethnic group statistics for England from administrative data.

Contact:
Jo Harkrader and Michelle Bellham
Admin.Based.
Characteristics@ons.gov.uk
+44 1329 444974

Table of contents

1 . Main points

- We have produced an initial assessment of the coverage for admin-based income by ethnicity statistics (ABIES) for the tax year ending 2016 as part of our feasibility research.

- This initial case study follows on from previous feasibility research to derive admin-based income statistics, and admin-based ethnicity statistics, by linking together multiple administrative data sources.

- By combining admin-based income and ethnicity statistics, we established both an income and a stated ethnicity for 77.0% of people aged 16 years and over in the Statistical Population Dataset (SPD) for England.

- The proportion of individuals with income identified and a stated ethnicity by region in England is lowest in London (65.0%) and highest in the North West (81.6%) and Yorkshire and The Humber (81.4%).

- In future publications we will continue this feasibility research to explore how ABIES can be improved, including providing analysis of income measures by ethnic group and how these compare with official income statistics, reviewing methods to adjust for missingness in the data, and the analysis at more granular geographical levels.

These research outputs are not official statistics on income by ethnic group. Rather, they are published as outputs from research into the feasibility of producing subnational multivariate statistics using administrative data. These outputs should not be used for policy or decision-making.

2 . About our transformation research

This article presents feasibility research on the potential to produce subnational multivariate income by ethnicity statistics, to demonstrate our progress towards producing more frequent subnational multivariate statistics on population characteristics. We will refer to these as admin-based income by ethnicity statistics (ABIES). We explore the coverage of the ABIES dataset across the income and ethnicity variables, by age, sex, region and local authority.

To produce ABIES, this research combines the admin-based income statistics (ABIS) dataset with v2.0 of the admin-based ethnicity statistics (ABES) dataset. These datasets are both derived from multiple administrative data sources, which have been linked together to produce statistics about a single topic. Both datasets use the Statistical Population Dataset (SPD) v3.0 as a population spine. This initial case study explores how derived administrative data-based outputs for single topics might be combined to produce multivariate outputs.

We have used the ethnicity variable from ABES, which makes use of 2011 Census data in addition to multiple administrative data sources to derive an individual's ethnic group.

While our ABIS provide income data for England and Wales, this ABIES research has been conducted for England only based on the England only coverage of ABES. We are working with the Welsh Government to acquire additional ethnicity data for Wales, that will provide a time series that will enable us to calculate similar estimates for Wales.

These initial findings show promise. We will continue to explore data sources and methods to account for missingness in the data, and to make improvements in the categorical and geographical granularity of ABIES in the future. We will also provide analysis of income measures by ethnic group and how these compare with official income statistics, working with colleagues producing official income outputs.

This research forms part of our population and social statistics transformation programme, which aims to provide the best insights on population, migration and society using a range of data sources. Producing income by ethnicity statistics for smaller geographical areas would fill a current gap in the evidence base available to inform areas such as levelling up policy and inequalities research. It also aligns with Ambition 1.2 of the Government Statistical Service (GSS) subnational data strategy and recommendations of the Inclusive Data Taskforce about improving disaggregation across groups and at differing levels of geography.
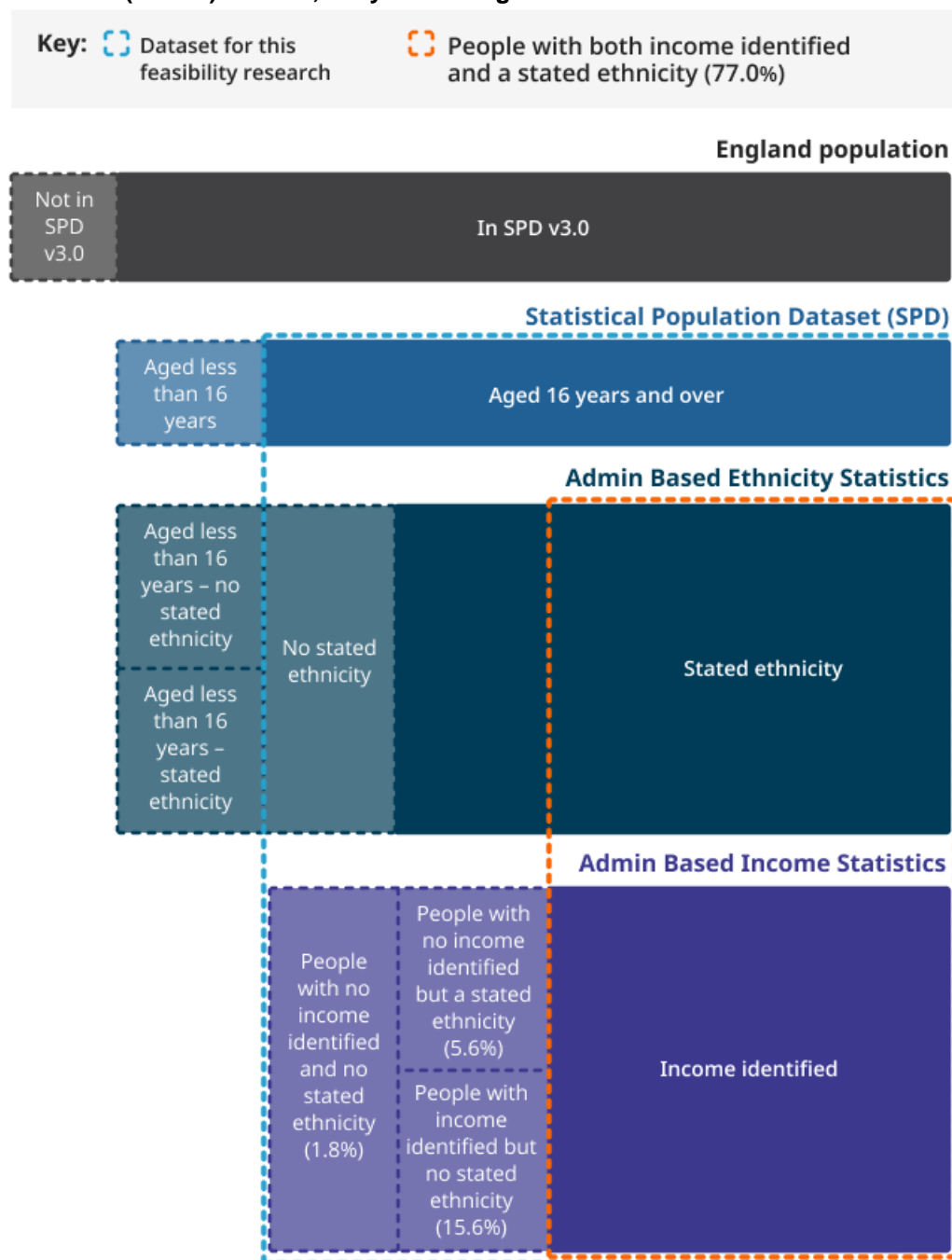
3 . Population coverage for admin-based income by ethnicity statistics (ABIES)

Both the admin-based income statistics (ABIS) and admin-based ethnicity statistics (ABES) datasets use the 2016 Statistical Population Dataset (SPD) v3.0 as the population base, and are joined together based on a unique identifier to create the admin-based income by ethnicity statistics (ABIES). Figure 1 provides a visual representation of how the SPD is linked to the ABIS and ABES.

The SPD v3.0 for England in 2016 contains around 43.9 million individuals aged 16 years and over. Previous analysis of the coverage of the population base has shown that even after accounting for possible linkage error, there are still some gaps in SPD coverage, particularly for age groups where we do not yet have access to sources of activity data, for example, the self-employed or the working-age population that is neither working nor receiving a benefit. These coverage issues will also have an impact on the coverage of the ABIES data.

Of the 43.9 million individuals aged 16 years and over in the SPD v3.0, 92.6% have income data from at least one source in the ABIS dataset. The ABES dataset provided a stated ethnicity for 82.6% of individuals in the SPD.

Figure 1: Visual representation of the population coverage of the admin-based income by ethnicity statistics (ABIES) dataset, tax year ending 2016

Source: Office for National Statistics

Notes:

1. Please note this diagram is not to scale.

2. "Stated ethnicity" refers to those with a stated ethnicity and no refusal on their most recent administrative data record in 2016, in line with the methods used to derive an individual's ethnic group in the ABES dataset.

3. "No stated ethnicity" refers to the ethnicity being recorded as refused or unknown, in line with the methods used to derive an individual's ethnic group in the ABES dataset. No stated ethnicity also includes individuals who are in the SPD v3.0 but have not been linked to any sources of ethnicity data.

4. "Income identified" refers to those identified with income data from at least one source in the ABIS dataset. Please see the glossary for more information about the sources used to derive income in the ABIS dataset.

As shown in Figure 2, the ABIES linked dataset identified that of people in the SPD v3.0:
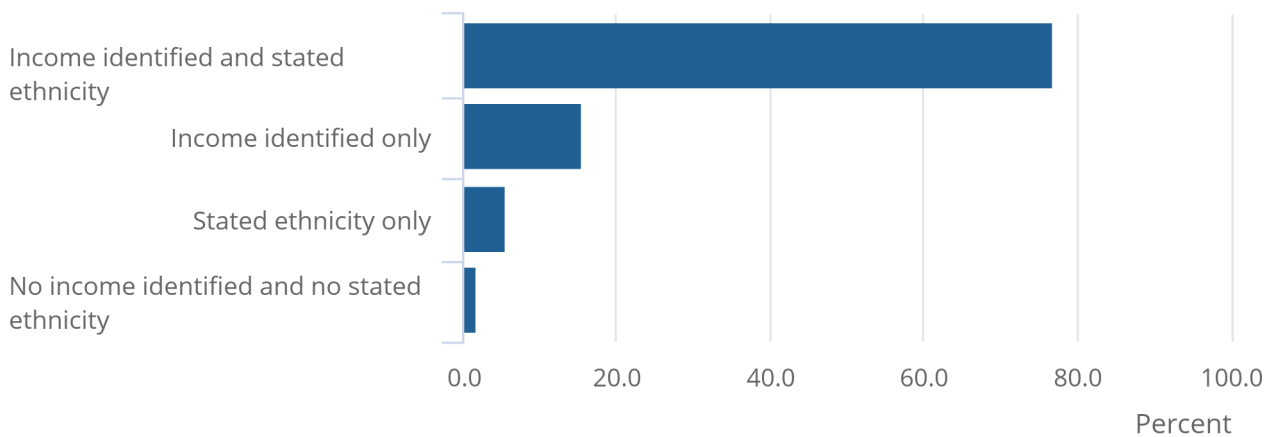
- 77.0% have income identified and a stated ethnicity

- 15.6% have income identified but no stated ethnicity

- 5.6% have no income identified but have a stated ethnicity

- 1.8% have no income identified and no stated ethnicity

Figure 2: Just over three-quarters of the Statistical Population Dataset (SPD) have both a source of income identified and a stated ethnicity

Proportion of the population aged 16 years and over in the SPD v3.0 with identified income, a stated ethnicity, both or neither, England, tax year ending 2016

Figure 2: Just over three-quarters of the Statistical Population Dataset (SPD) have both a source of income identified and a stated ethnicity

Proportion of the population aged 16 years and over in the SPD v3.0 with identified income, a stated ethnicity, both or neither, England, tax year ending 2016

Source: Office for National Statistics

Notes:

1. Proportions may not sum to 100.0% owing to rounding.

2. "Stated ethnicity" refers to those with a stated ethnicity and no refusal on their most recent administrative data record, in line with the methods used to derive an individual's ethnic group in the ABES dataset.

3. "No stated ethnicity" refers to the ethnicity being recorded as refused or unknown, in line with the methods used to [derive an individual's ethnic group in the ABES dataset](#). No stated ethnicity also includes individuals who are in the SPD v3.0 but have not been linked to any sources of ethnicity data.

4. "Income identified" refers to those identified with income data from at least one source in the ABIS dataset. Please see the glossary for more information about the sources used to derive income in the ABIS dataset.

Figures 3 and 4 show that the proportion of people with both income identified and stated ethnicity in the ABIES dataset is lowest at 16 years of age but rises substantially until the age of 22 years for both males (from 16.0% to 71.3%) and females (from 15.8 % to 73.8%). This is because of an increase in the availability of income data, which may correspond to individuals within this age range leaving full-time education and entering the labour market, and therefore appearing in the ABIS dataset.

For females, there is a decrease in the proportion of people with both income identified and stated ethnicity between the ages of 45 years and 62 years, from 82.7% to 78.4%. This is driven by a decrease in the proportion of people with an income recorded in the ABIES dataset.

For both males and females there is an increase in the proportions of individuals who have income identified and stated ethnicity from the age of 63 years for females and from the age of 65 years for males. This corresponds with State Pension age, at which point income would be recorded owing to the receipt of State Pension. In the tax year ending 2016, the State Pension age was 63 years for women and 65 years for men.
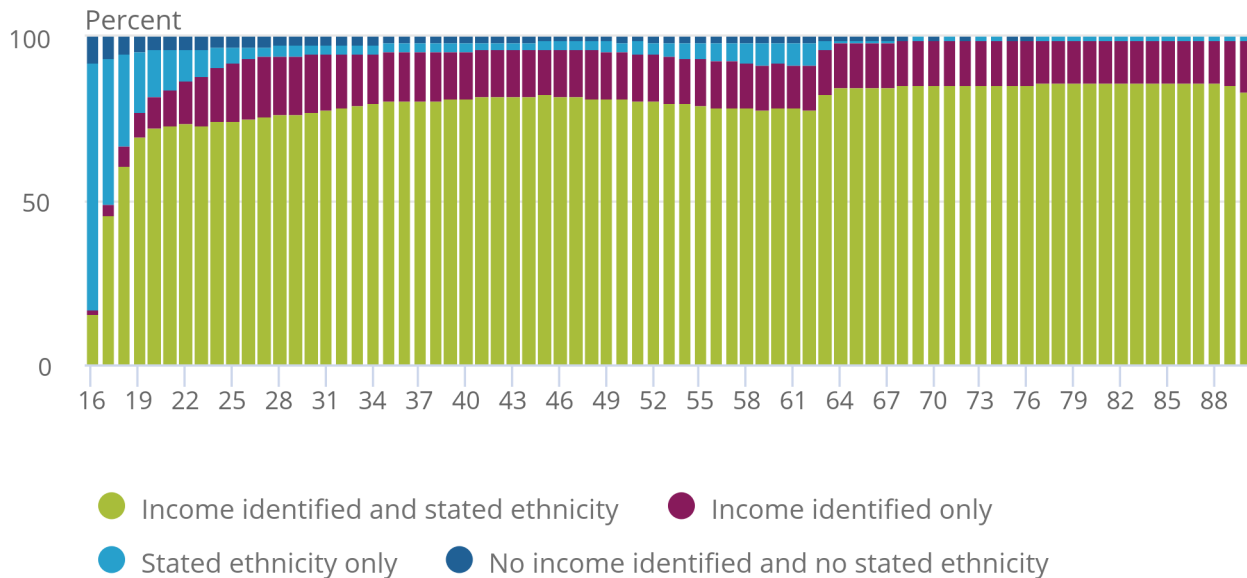
The proportion of males and females with no income identified and no stated ethnicity is small for all ages but is slightly higher for 16- to 20-year-olds. It is at its lowest (less than 0.5%) from the age of 66 years onwards for males and from the age of 67 years onwards for females.

Figure 3: Females aged 16 to 19 years are least likely to have both a source of income identified and a stated ethnicity

Proportion of females aged 16 years and over in the SPD v3.0 with income identified, a stated ethnicity, both or neither, England, tax year ending 2016

Legend:
- Income identified and stated ethnicity
- Income identified only
- Stated ethnicity only
- No income identified and no stated ethnicity

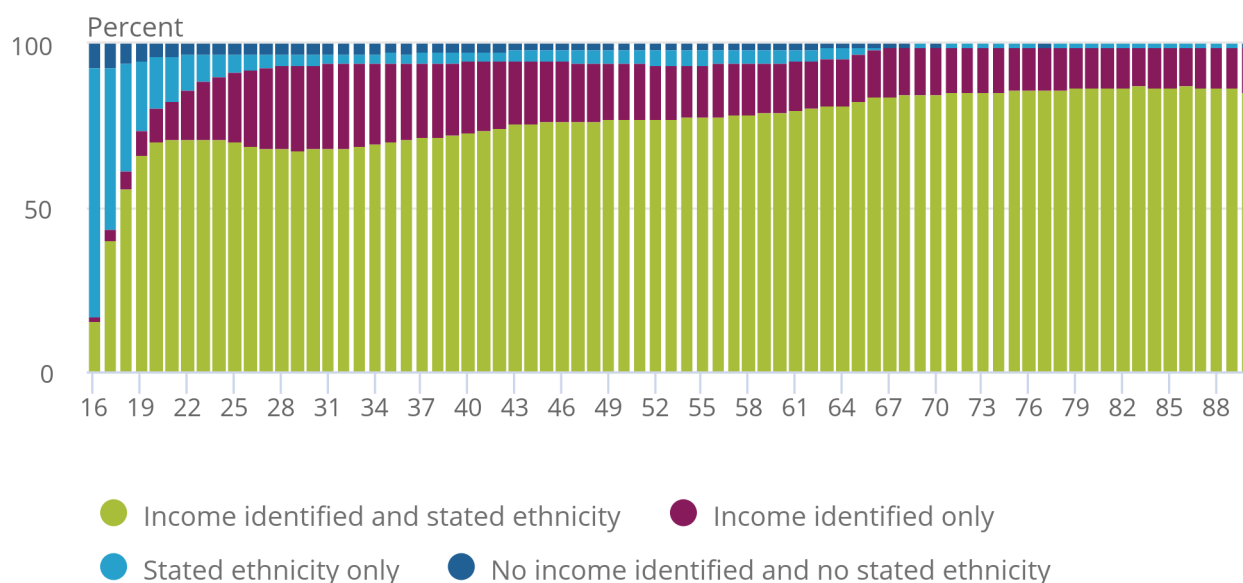**Source: Office for National Statistics**

Notes:

1. "Stated ethnicity" refers to those with a stated ethnicity and no refusal on their most recent administrative data record, in line with the methods used to derive an individual's ethnic group in the ABES dataset.

2. "No stated ethnicity" refers to the ethnicity being recorded as refused or unknown, in line with the methods used to derive an individual's ethnic group in the ABES dataset. No stated ethnicity also includes individuals who are in the SPD v3.0 but have not been linked to any sources of ethnicity data.

3. "Income identified" refers to those identified with income data from at least one source in the ABIS dataset. Please see the glossary for more information about the sources used to derive income in the ABIS dataset.

4. All individuals on the SPD v3.0 have their sex recorded as either male or female.

Figure 4: Males aged 16 to 20 years are least likely to have both a source of income identified and a stated ethnicity

Proportion of males aged 16 years and over in the SPD v3.0 with income identified, a stated ethnicity, both or neither, England, tax year ending 2016

Figure 4: Males aged 16 to 20 years are least likely to have both a source of income identified and a stated ethnicity

Proportion of males aged 16 years and over in the SPD v3.0 with income identified, a stated ethnicity, both or neither, England, tax year ending 2016

Percent

● Income identified and stated ethnicity   ● Income identified only

● Stated ethnicity only   ● No income identified and no stated ethnicity

Source: Office for National Statistics

Notes:

1. "Stated ethnicity" refers to those with a stated ethnicity and no refusal on their most recent administrative data record, in line with the methods used to derive an individual's ethnic group in the ABES dataset.

2. "No stated ethnicity" refers to the ethnicity being recorded as refused or unknown, in line with the methods used to derive an individual's ethnic group in the ABES dataset. No stated ethnicity also includes individuals who are in the SPD v3.0 but have not been linked to any sources of ethnicity data.

3. "Income identified" refers to those identified with income data from at least one source in the ABIS dataset. Please see the glossary for more information about the sources used to derive income in the ABIS dataset.

4. All individuals on the SPD v3.0 have their sex recorded as either male or female.

4 . Subnational coverage for admin-based income by ethnicity statistics (ABIES)

For this initial case study, we have analysed the proportion of people with income identified and a stated ethnicity in the admin-based income by ethnicity statistics (ABIES) dataset at regional and local authority (LA) levels (for more information about how region is assigned, see our Statistical Population Dataset (SPD) subnational analysis).

Regional coverage

The level of coverage for those individuals who have income and a stated ethnicity varies between regions in England. The proportion of individuals with income identified and a stated ethnicity is lowest in London (65.0%) and highest in the North West (81.6%) and Yorkshire and The Humber (81.4%).

London has the highest proportion of individuals on the SPD v3.0 with income identified but no stated ethnicity (23.3%); this is notably higher than the other eight regions. London also has the highest proportion of individuals with a stated ethnicity only (7.7%) and the highest proportion of individuals with no income identified and no stated ethnicity (3.9%). The higher level of coverage seen in the North West and Yorkshire and The Humber is largely the result of the high level of coverage for stated ethnicity in the Admin-based Ethnicity Statistics (ABES).

Table 1: Coverage of income and ethnicity in the admin-based income by ethnicity statistics (ABIES) dataset varies by region
Proportion of individuals aged 16 years and over in the SPD v3.0 by region with a source of income identified, a stated ethnicity, both or neither, England, tax year ending 2016

Region	Income identified and stated ethnicity	Income identified only	Stated ethnicity only	No income identified and no stated ethnicity
England	77.0	15.6	5.6	1.8
North East	79.7	13.8	5.4	1.0
North West	81.6	11.7	5.3	1.4
Yorkshire and the Humber	81.4	11.9	5.4	1.3
East Midlands	79.8	13.8	5.2	1.2
West Midlands	79.9	13.0	5.7	1.4
East of England	80.2	13.3	5.2	1.3
London	65.0	23.3	7.7	3.9
South East	73.8	19.3	5.2	1.7
South West	80.3	13.9	4.7	1.1

Source: Office for National Statistics

Notes

1. Proportions may not sum to 100.0% owing to rounding.

2. "Stated ethnicity" refers to those with a stated ethnicity and no refusal on their most recent administrative data record in 2016, in line with the methods used to derive an individual's ethnic group in the ABES dataset.

3. "No stated ethnicity" refers to the ethnicity being recorded as refused or unknown, in line with the methods used to derive an individual's ethnic group in the ABES dataset. No stated ethnicity also includes individuals who are in the SPD v3.0 but have not been linked to any sources of ethnicity data.

4. "Income identified" refers to those identified with income information from at least one source in the ABIS dataset. Please see the glossary for more information about the sources used to derive income in the ABIS dataset.

Local authority (LA) coverage

Coverage at the LA level of geography varies widely. Corresponding with the analysis for the regional coverage, the proportion of people for whom we identified income and a stated ethnicity within the ABIES linked dataset by LA was lowest in City of London (44.5%) and highest in St. Helens (89.3%). Figure 5 shows that just over four-fifths of LAs (n=258) have both income identified and a stated ethnicity for at least 70% of individuals aged 16 years and over.

Figure 5: Coverage of income and ethnicity varies by local authority

Number of local authorities by proportion of individuals aged 16 years and over in the SPD v3.0 with both a source of income identified and a stated ethnicity, England, tax year ending 2016

Notes:

1. There are 309 local authorities in this analysis overall. The SPD v3.0 uses the 2021 local authority geographies.

2. "Stated ethnicity" refers to those with a stated ethnicity and no refusal on their most recent administrative data record in 2016, in line with the methods used to derive an individual's ethnic group in the ABES dataset.

3. "Income identified" refers to those identified with income information from at least one source in the ABIS dataset. Please see the glossary for more information about the sources used to derive income in the ABIS dataset.

Download the data

.xlsx

Caution should be exercised when interpreting any conclusions drawn from admin-based income by ethnicity figures. Population coverage of the SPD v3.0 will be affected by areas with high population churn owing to the time lags in production of up-to-date data; some geographical areas will be affected by this more than others. Furthermore, we have not yet been able to analyse how representative the individuals with income identified and a stated ethnicity are of the total population by LA. Lower proportions of individuals with both income identified and a stated ethnicity would not automatically equate to less reliable analysis, if the proportion of people for whom we do have income identified and a stated ethnicity share the same characteristics as those for whom we do not have data.

Ongoing development of our admin-based income and ethnicity statistics, as well as work to develop methods to account for missingness in our individual and multivariate measures, will help to improve the representativeness and completeness of the data we are producing.

5 . Developing subnational multivariate income by ethnicity statistics from administrative data, England: tax year ending 2016 data

Developing subnational multivariate income by ethnicity statistics from administrative data: England, tax year ending 2016 data
Dataset | Released 31 August 2022
Data for feasibility research on an initial case study producing income by ethnicity statistics for England from administrative data.

6 . Glossary

Ethnic group

The self-reported ethnic group of the individual, according to their own perceived ethnic group and cultural background.

Five categories are presented in this article. This is a list of the ethnic groups included in each category:

- Asian ethnic group: Bangladeshi, Chinese, Indian, Pakistani, Asian Other

- Black ethnic group: African, Caribbean, Black Other

- Mixed ethnic group: White and Asian, White and Black African, White and Black Caribbean, Mixed Other

- White ethnic group: British, Gypsy, Roma or Irish Traveller [note 1], Irish, White Other, White not specified [note 2]

- Other ethnic group: Arab, Any other ethnic group

Notes for Ethnic group

1.The Gypsy, Roma and Irish Traveller ethnic groups have been aggregated because of differences in response options across data sources meaning that it is not possible to separate them. Hospital Episode Statistics (HES) and Improving Access to Psychological Therapies (IAPT) do not include any Gypsy, Roma or Irish Traveller response options.

2.The Higher Education Statistics Agency (HESA) data for England and Wales only have categories for White and Gypsy or Traveller within the higher-level White ethnic group. Those with a sub-category ethnicity of White in HESA were re-coded as White not specified.

Ethnicity stated or stated ethnicity

Ethnicity stated refers to the ethnicity being recorded as a specific ethnic group and not refused or unknown on their most recent administrative data record in 2016, in line with the methods used to derive an individual's ethnic group in the Admin-based Ethnicity Statistics (ABES) dataset.

Income identified

"Income identified" refers to those identified with income data from at least one source in the ABIS dataset. For more information about the income measures used in admin-based income statistics (ABIS) and admin-based income by ethnicity statistics (ABIES), please refer to the ABIS glossary. For more information about income measures, sources and definitions, please refer to our income and earnings statistics guide.

No stated ethnicity

No stated ethnicity refers to the ethnicity being recorded as refused or unknown, in line with the methods used to derive an individual's ethnic group in the ABES dataset. No stated ethnicity also includes individuals who are in the Statistical Population Dataset (SPD) v3.0 but have not been linked to any sources of ethnicity data.

7 . Data sources and quality

This research forms part of our population and social statistics transformation programme, which aims to provide the best insights on population, migration and society using a range of data sources. The findings will form part of the evidence base for the 2023 National Statistician's Recommendation on the future of population and social statistics.

The research has been conducted for England only. We have income data for England and Wales. We are working with the Welsh Government to acquire additional ethnicity data for Wales, that will provide a time series that will enable us to calculate similar estimates for Wales.

Population base

The 2016 Statistical Population Dataset (SPD) v3.0 was used as the population base for the admin-based ethnicity and admin-based income statistics datasets. It aims to approximate the usually resident population as at 30 June 2016. The quality of the population base will have an impact on the quality of the admin-based income and ethnicity statistics. More information about the coverage of the population base can be found in a previous report.

The SPD was previously called the admin-based population estimates (ABPE), and is one of the inputs to the Dynamic Population Model (DPM), our statistical modelling approach to produce more timely estimates of the population that are able to better respond to user needs.

Admin-based income statistics

The admin-based income statistics (ABIS) linked dataset was produced using the following administrative data sources:

- Department for Work and Pensions' (DWP) National Benefits Dataset (NBD) data

- DWP's Single Housing Benefit Extract (SHBE) data

- HM Revenue and Customs' (HMRC) Pay As You Earn (PAYE) P14 data

- HMRC's Tax Credits data

- HMRC's Child Benefits data

- Winter Fuel Payment (ONS-derived) and Christmas Bonus (ONS-derived) data

Income information from these data sources were combined using a unique identifier to produce a net annual income amount for each individual. This was then linked to the SPD v3.0 dataset using a unique identifier to retain income information only for the usually resident population. Income Tax and National Insurance amounts were estimated using available data and used to produce a measure of net annual income for individuals.

Admin-based ethnicity statistics

The Admin-based Ethnicity Dataset was produced using the following administrative data sources:

- [English School Census (ESC)](#), 2011 to 2016

- [Hospital Episode Statistics (HES)](#), 2009 to 2016

- [Improving Access to Psychological Therapies (IAPT)](#), 2012 to 2016

- [Birth Notifications](#), 2006 to 2016

- [Higher Education Statistics Agency (HESA)](#), 2010 to 2016

- [2011 Census](#) data

Ethnicity records from these data sources were linked to the 2016 statistical population dataset (SPD) v3.0 using unique identifiers. A method to select a final ethnicity per person was then implemented, as described in [Producing admin-based ethnicity statistics for England: changes to data and methods](#).

Creating the income and ethnicity joined data asset

As the admin-based income and ethnicity datasets were both created using the SPD v3.0, they could be joined together using a unique identifier. The ethnicity dataset contains records for England only (as per the local authority on the SPD) so all records for Wales in the income dataset were dropped (3.0 million). All records for those aged under 16 years (10.3 million records) were also dropped, as the population of interest for income statistics is people aged 16 years and over. This left 43.9 million records for the analysis.

8 . Future Developments

This feasibility study of producing admin-based income and ethnicity statistics (ABIES) shows early promise at this initial stage. The approach offers the potential to produce more granular outputs, in terms of geography and small population groups, than what is currently possible using survey data alone, while reducing respondent burden.

We will continue to explore how we can develop and improve upon subnational multivariate ABIES. In future publications, we aim to:

- produce income measures by ethnic group using the ABIES dataset, comparing trends with those observed in official income statistics

- produce ABIES for more recent reference periods

- continue to progress our research to improve the admin-based income statistics (ABIS) and admin-based ethnicity statistics (ABES) measures that are used to produced ABIES

- continue exploring the characteristics of the individuals in the Statistical Population Dataset for whom we do not have a stated ethnicity

- begin analysing the quality of ABES estimates through census linkage

- continue exploring how the limitations of the ABIES dataset may be influencing our findings that do not align with trends by ethnic group in comparator data

- begin exploring the potential to produce admin-based multivariate statistics for smaller geographical areas and with a more detailed breakdown of ethnic groups, subject to statistical disclosure control considerations

- begin exploring methods to adjust for missingness in the admin-based ethnicity and admin-based income statistics

Feedback

We welcome feedback on the quality, value and impact that using these figures would have, and our planned future developments. Please email your feedback to Admin.Based.Characteristics@ons.gov.uk.

9 . Related links

Developing admin-based ethnicity statistics for England: 2016
Article | Released 23 May 2022
Update on feasibility research on producing statistics on the population by ethnic group for England from administrative data, following the inclusion of additional data sources and method changes. These research outputs are not official statistics.

Producing admin-based ethnicity statistics for England: changes to data and methods
Article | Released 23 May 2022
An overview of the changes to data sources and methods in the feasibility research on producing statistics on the population by ethnic group from administrative data.

Admin-based income, England and Wales: tax year ending 2016 revised results
Article | Released 23 June 2021
Experimental small area estimates of gross and net income using data from HM Revenue and Customs' (HMRC) Pay As You Earn (PAYE), Self-Assessment and the Department for Work and Pensions' (DWP) benefit systems.

Admin-based population and migration estimates: research update
Article | 26 November 2021
Provides an overview of all the latest research into the transformation of population and migration statistics to an administrative data first methodology.

Income and earnings statistics guide
Methodology | 2 March 2022
Explains the relationship between income and earnings data and outlines the statistics produced by the Office for National Statistics, Department for Work and Pensions and HM Revenue and Customs.

10 . Cite this article

Office for National Statistics (ONS), released 31 August 2022, ONS website, Article, Developing subnational multivariate income by ethnicity statistics from administrative data, England: tax year ending 2016