

Data Explained

Migrant Workers Scan and linked datasets

Author: Felix Ritchie, Van Phan, Damian Whittard, Alex Bryson, John Forth, Carl Singleton
Date: December 2025

This Data Explained summarises experiences and learning from working with the Migrant Workers Scan and its links to other datasets, the Annual Survey of Hours and Earnings (ASHE) and the HMRC Paye As You Earn (PAYE) and Self-Assessment (SA) datasets. This publication is intended to help guide future researchers using this data and to provide feedback into dataset development and documentation.

The administrative data discussed in this Data Explained was made securely available through Wage and Employment Dynamics Phase II (WEDII), funded by ADR UK. The data used in this research project comes from HM Revenue and Customs (HMRC) and the Office for National Statistics (ONS) and was accessed through the ONS Secure Research Service. The data was not originally collected for research, and it is expected that there are gaps and inconsistencies in its recording, a number of which are detailed in the following.

Introduction

In this Data Explained we consider some analytical characteristics of the Migrant Workers Scan (MWS)¹ and three other datasets: The Annual Survey of Hours and Earnings (ASHE), collected by ONS from employers; the PAYE payslip data reported to HM Revenue and Customs by employers as part of their statutory tax recording; and the Self-Assessment (SA) returns reported to HMRC at the year of tax year. The MWS data described here is for the 1% of the population covered in ASHE (all persons with a National Insurance number (NINo) ending in a specified two digits). We refer to the data hereafter as WED MWS. It can be linked to ASHE through the 'piden' field based on the lookup file provided by ONS while it is linked directly to HMRC PAYE/SA by the variable 'hmrc_id'.

How is the data collected?

The WED MWS data is extracted from HMRC administrative records. Since 2002 the MWS has been recording all non-UK residents who apply for a NINo. Generally, these are adults who have arrived in the UK, or children who arrived with parents and who need to apply for a NINo as they turn 16. However, changes in child benefit rules mean that some children of parents applying for child benefit could be given an NINo before they reach 16. The dataset supplied to the WED team only contains those aged 16 or over.

Although the current MWS system was implemented in 2002, the WED MWS dataset contains individuals who were recorded as arriving and registering for National Insurance back to 1975. While the quality of that data looks comparable to the post-2002 data (in terms of missing values and plausible variable distributions), it is not clear how or from where this data was collected, and there currently appears to be no information publicly available.

The 'PAYE' files are the submissions made by employers on payments made to employees through HMRC's Real-Time Information (RTI) system. Each submission is identified by a pay period, either week 1-52 or month 1-12, with the tax year starting on 6 April. These could involve multiple resubmissions for the same pay period as an employer updates its records. The WED files take the last RTI submission in any pay period as the definitive one, so that any employee only has one payment record, per PAYE scheme, per pay period.

The SA files come from the annual submissions made by individuals with non-employment income, and the variables correspond closely to the fields on the self-assessment form. The data originally comes in multiple files, each one corresponding to a subsection of the tax declaration form; we have combined them all into one record per individual per year.²

1

<https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/migrantworkerscanqualityassuranceofadministrativedatausedinpopulationstatisticsjan2017>

² For more details of the HMRC PAYE and SA datasets, see the WED *HMRC Quick Guide for users*

ASHE is collected in April each year, so 'ASHE year 2016' (collected in April 2016) will correspond to 'tax year 2017' (April 2016-March 2017). ASHE is fully described in the ONS metadata catalogue (ONS, 2024).

The PAYE data is available from tax years 2015-2019. The SA data is available from tax years 2011-2018. ASHE data is available from 1997-2022 (tax years 1998-2023).

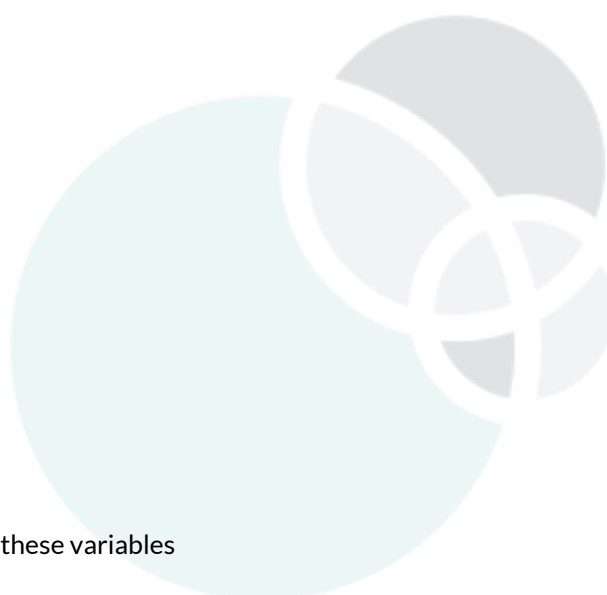
The number of distinct individuals in the WED MWS dataset are detailed in Table 1³, which also breaks this down by arrival year and sex. Note that in the MWS, sex is recorded as male, female (married) and female (unmarried). The information is self-reported, but DWP does carry out registration checks via interview to confirm the data.

Arrival year	Number	% Male	% Female	
			Married	Unmarried
75-81	6,127	53.5%	43.0%	3.4%
82-86	4,707	52.0%	44.0%	4.0%
87-91	8,296	52.1%	36.0%	11.9%
92-96	7,694	52.2%	32.7%	15.1%
97-01	10,123	50.0%	42.6%	7.4%
02-06	25,130	53.2%	39.7%	7.1%
07-11	30,801	53.6%	35.8%	10.6%
12-16	35,147	53.6%	15.4%	30.9%
2017 on	28,175	51.5%	7.7%	40.8%
Total	156,200	52.7%	27.6%	19.7%

Table 1 Observations in the WED MWS by arrival data and sex

The Wage and Employment Dynamics website <https://www.wagedynamics.com/hmrc/> provides more information in the WED MWS quick user guide.

³ See the MWS User guide for details on the construction of these variables



Key variables

The WED MWS contains a small number of variables. The key ones are:

Original variable	Values	Quality
Age_at_q1_2000	Used to derive a birth year, from which age can be derived	Verified by WED against ASHE age data – values from individuals found in both datasets are highly consistent (i.e. recorded age is one dataset is within a year of the other).
Sex_status	0 Male 1 Female married, 2 Female unmarried	Verified by WED against ASHE binary marker for sex – values from individuals found in both datasets are almost always the same.
Country_code	3-digit code for nationality	Always filled in, but codes are sometimes inconsistent, perhaps reflecting changes in reference numbers over time (for example, the same country may have different code numbers at different times); about 1.5% are ‘unknown’; about 0.5% have British nationality. The country-code table is provided by ONS.
Prev_residence	3-digit code for country of residence immediately prior before arriving in the UK	Supplied in less than 25% of cases. Of those, region of previous residency and region of nationality are usually the same.
Date of arrival	Date of arrival	Always provided. Dates look plausible (eg lots of arrival from Eastern Europe 2004 onwards).
Date of registration	Date of registration	Missing in about 6% of cases; a small number arrive after their date of registration; this is feasible as it is possible to apply for NINo before arriving in the country (for example, if right to work or study has been granted).
Postcode area	Address at point of registration for a NINo: first set of letters for a postcode, e.g. AB for Aberdeen	Missing in about 15% of cases; distribution broadly reflects population, although with some skewing towards London/SE.

Table 2 Original variables in WED MWS

The small number of sex/age inconsistencies (under 1%) found between the ASHE and WED MWS datasets are a similar proportion to inconsistencies found within the ASHE dataset itself when comparing across years, indicating data collection errors. We have not identified any pattern in the errors to suggest they are non-random.

From these original variables, several more practical variables have been derived by the WED team, including a binary marker for sex, and groupings of countries. See the MWS User Guide for details.

There are three key linking variables in the data:

Variable	Present in	Purpose
hmrc_id	All MWS, PAYE and SA source files	Unique person identifier across all the HMRC data
piden	ASHE source files Added to HMRC files	Unique person identifier for ASHE, linkable to hmrc_id via a lookup table
index_scheme	PAYE files	Identifies PAYE scheme (can be used as a proxy for employer)

Table 3 Linking variables

What can the data be used for?

The MWS is of limited value when used in isolation, as it contains very little analytical information. However, when combined with other datasets it can provide useful insights into the impact of migration on labour market data. In this section we explore how well the WED MWS data links to other relevant datasets. Figure 1 illustrates how MWS links to other datasets, and Tables 4-6 show the extent of these linkages.

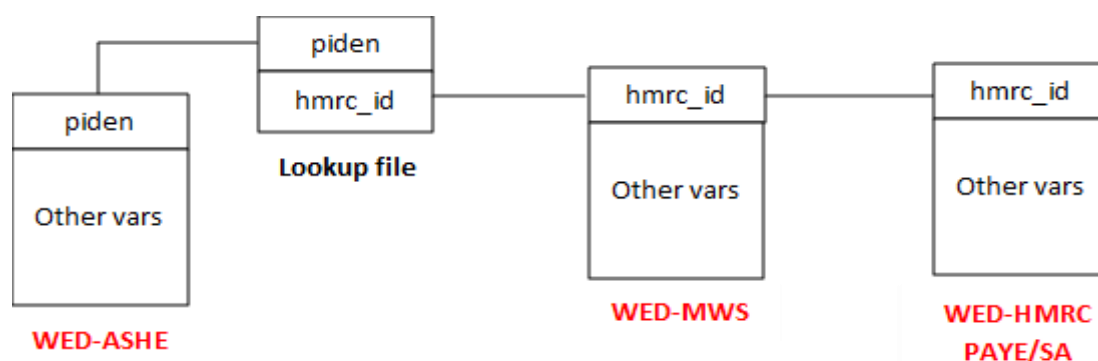


Figure 1: Overview of linking datasets together

		Total
Does the person have a link to the ASHE piden?	No	87,857
	Yes	68,343
Total		156,200

Table 4 Linkage to ASHE and HMRC index variables

All MWS respondents should have a valid NINo (as that is the purpose of registering), and should therefore appear in the lookup file which links hmrc_id to piden. Once linked to the piden, linkage to ASHE becomes feasible. However, there are significant numbers of records that are found in the lookup table but not found in ASHE. 56% of those that could be linked to ASHE are not linked. One possible explanation is that ASHE only measures workers at a single point in the year (April),

and so seasonal workers, for example, are quite likely to be missed. Some immigrants may also have left the country permanently before they become eligible for the ASHE survey. Some young migrants may not have started working yet. Finally, there are some problems with recording NINOs in ASHE, and some non-response in ASHE which may be concentrated in firms employing migrants (small, private sector). Table 5 shows that individuals arriving before 2000 and after 2016 are least likely to be found in ASHE.

Arrival year	Found in lookup table but not found in ASHE	Appears in both MWS and ASHE
75-81	4,836 78.93%	1,291 21.07%
82-86	3,363 71.45%	1,344 28.55%
87-91	5,781 69.68%	2,515 30.32%
92-96	4,440 57.71%	3,254 42.29%
97-01	4,278 42.26%	5,845 57.74%
02-06	9,756 38.82%	15,374 61.18%
07-11	14,770 47.95%	16,031 52.05%
12-16	18,622 52.98%	16,525 47.02%
2017 on	22,011 78.12%	6,164 21.88%
Total	87,857 56.25%	68,343 43.75%

Table 5 Linkage rates to ASHE

For the PAYE and SA data, linkage rates are much better. We link directly via the hmrc_id variable, which in theory should give 100% match if migrants are still in the country and working. However, because the HMRC data is only available for a few years, comparing how many of those appear in the HMRC datasets is of limited use. Instead Table 6 shows the number of matches for the years for which we have HMRC data. We can see that, of the 156,000 MWS respondents (Table 1), about 60% appear in one or more of the HMRC datasets each year.

	Numbers matching to PAYE records			Numbers matching to SA
	Any	With employment	Pensions only	
2011				9,409
2012				9,997
2013				11,011
2014				12,175
2015	81,743	43,758	37,985	13,520
2016	81,217	47,555	33,662	14,966
2017	81,219	50,185	31,034	16,106
2018	81,197	51,826	29,371	16,213
2019	56,901	53,209	3,692	
2020	17,475	0	17,475	
Total	399,752	246,533	153,219	103,397

Table 6 Match rates with HMRC data

Most migrants observed in the HMRC data period have payslip data, either from occupational pensions or employment. Smaller numbers link to the self-assessment data but this reflects the smaller numbers in self-assessment: only 100,000 per year compared to 450,000 per year receiving payslips.

It is not easy to tell if the links are biased, as there are good reasons for thinking that the ones linking are different to those that don't link. For example, migrants recorded in the MWS as arriving before 1980 are less likely to be matched, and those matched are more likely to be from Europe (see MWS User Guide for tables). These together reflect the large increase in European migration 2004-2016. As another example, higher match rates with the SA data for migrants arriving earlier could simply reflect that older workers have more time to establish a new business. Moreover, the migration statistics only tell us who has arrived, not if they stayed and worked, or left the country again. Hence this is an active area of potential research – are these differences real or are they the result of matching effects?

We can get a sense of how representative the WED MWS data is by comparing to other sources. By merging MWS data into ASHE records, we can infer the proportion of employees who are migrant workers, as shown by the solid black line in Figure 1. Over the period 2015-2021, this proportion remained fairly stable at around 12-13%. This pattern aligns closely with the linked HMRC PAYE-MWS data, represented by the black dash line. In comparison, figures from the Annual Population Survey (APS) (using the variable “date last arrived in the UK” as an indicator of migration) indicate a slightly higher proportion, averaging around 14-15% between 2016-2019, before experiencing a 2% decline in 2020. Until 2020, the change over time in all data sources is very similar.

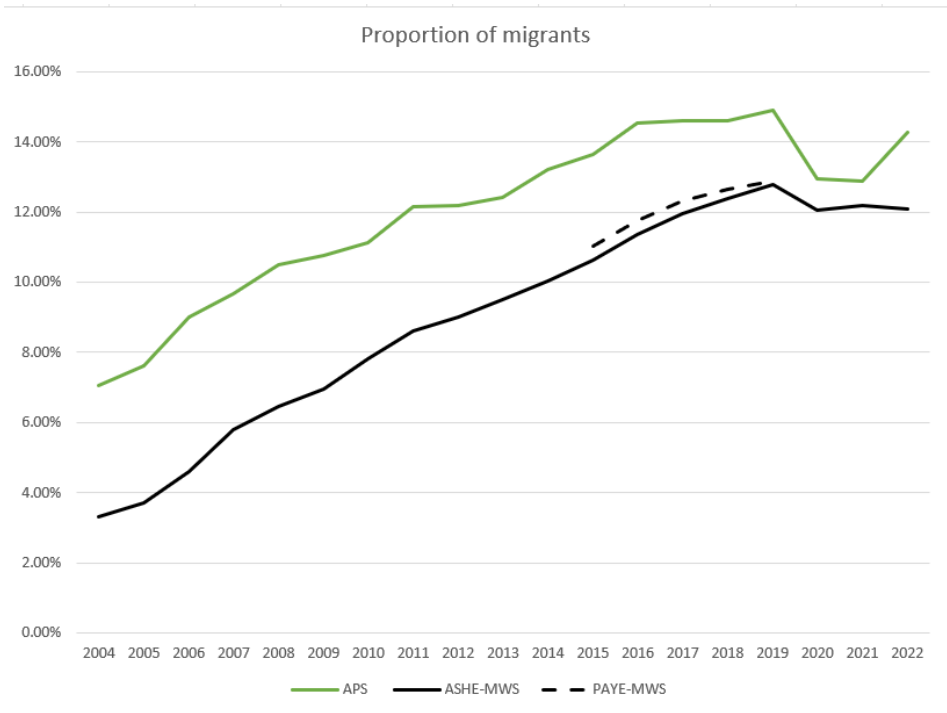


Figure 2 Proportion of employees who are migrants, across different datasets (%); authors' calculations

In contrast, the self-employment data may suggest a different observed pattern. Figure 2 shows the proportion of migrant workers who are reported as self-employed in APS and the proportion of migrants in MWS who appear in the HMRC-SA data.⁴ Over time, the differences between APS and HMRC-SA have narrowed, with the figures converging closely in 2018.

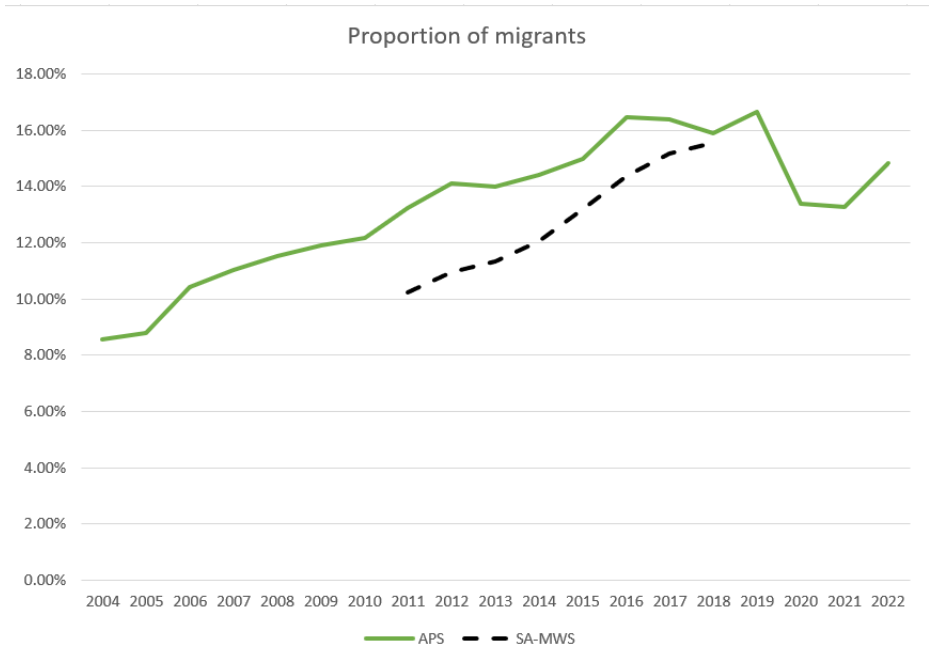


Figure 3 Proportion of migrants who are self-employed (APS) and who have self-assessment returns (SA-WED MWS) (%); authors' calculations

We can also consider the distribution of migrants in the working population. Figure 4 presents a stacked bar chart depicting the changing composition of migrant employees in the UK labour market from 2004 to 2022 as recorded in the APS.

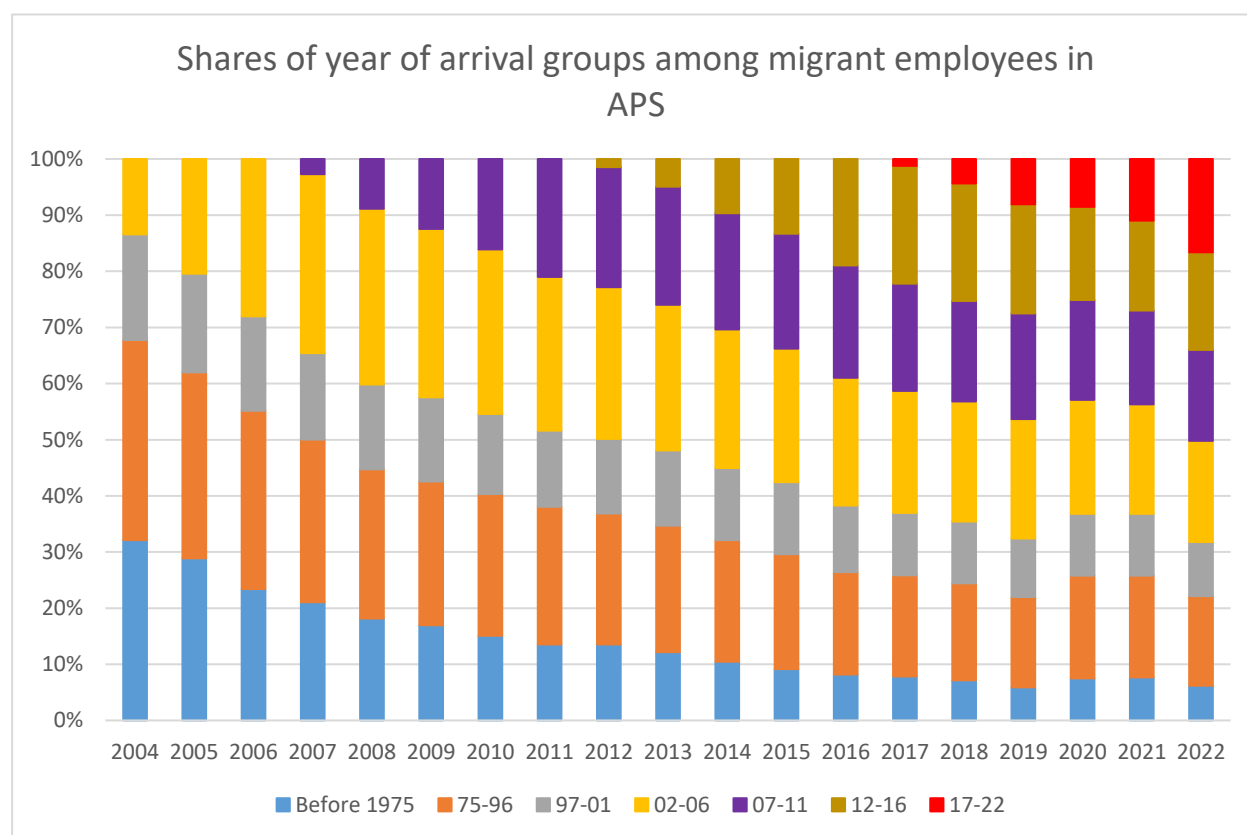


Figure 4 Distribution of arrival year for migrant employees in the UK by APS year; authors' calculations

The proportion of immigrants who arrived before 1975 shows a steady decline during this period because this group is aging and retiring. The 2002-2006 immigrant cohort demonstrates a significant trend that aligns with the EU enlargement by 2004 when ten new countries, particularly from Eastern Europe, joined in the union. Therefore, this group accounted for approximately 29% of all migrant workers in 2007. The proportion of this cohort still consistently persisted through the late 2000s but slowly dropped to about 15% by 2022. The UK's immigrant workforce continues to evolve through its newer arrival groups. The migrants who arrived between 2012-2016 maintained a substantial presence by representing 15-21% of all migrant workers throughout the second half of the 2010s. The latest cohort (2017-2022) has steadily integrated into the workforce, growing to represent 16.5% of all migrant workers by 2022.

We compare this with ASHE-WED MWS in Figure 5. Unlike APS, the immigrant cohort who arrived before 1975 is not captured in ASHE-WED MWS. The 1975-1996 migrant cohort experienced a substantial decrease to approximately 7% by 2022. The 2002-2006 cohort demonstrates a substantial increase, nearly twice the APS proportion following the EU enlargement after 2004. Both datasets exhibit the same general patterns such as the progressive decrease in earlier arrival

cohorts and substantial effects of Brexit and the COVID-19 pandemic around 2019-2020 which caused major drops across most cohorts.

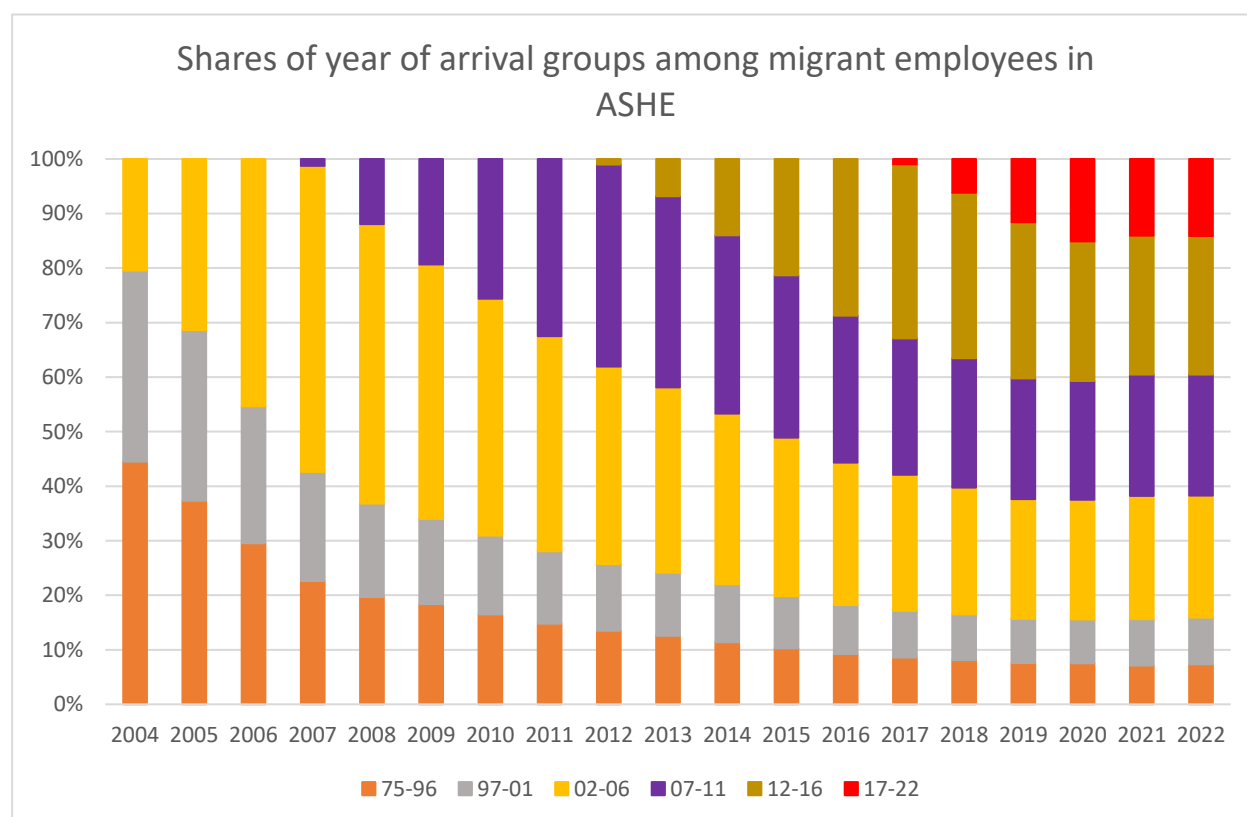


Figure 5 Distribution of arrival year for migrant workers in the UK by ASHE year; authors' calculations

Suggested improvements

The main problem is the data collection phase, with postcode and country of previous residency being the most likely variables to be omitted.

Suggested future data linkages

There are not many obvious linkages. One potential is the DWP benefits data⁵, but migrants might not be eligible for many of the benefits covered. Similarly, education and training data in LEO would be useful for those who came as children, but generally working migrants complete their compulsory education before arriving in the UK.

A significant linkage would be a dataset that gives an indication of whether migrants have left the country; this would allow much better analysis of working lives and earnings by avoiding the survivor bias when returnees are unaccounted for. It's not clear if this data exists, except possibly in Home Office files. It may be possible to make some inferences from identifying whether continual non-return of End of Year PAYE or SA returns; this becomes feasible if the HMRC data is extended beyond the current 2011-2018.

Conclusion

The WED MWS dataset is relatively simple to deal with. It is of limited interest when used in isolation, but when used to enhance other datasets with details of migration the potential seems significant. Its characteristics and use in analysis are yet to be fully exploited, but when explored this initial analysis (and the high rate of linkage with other HMRC datasets) suggests it will be a valuable addition to the portfolio of wage datasets.

References

Office for National Statistics, released 08 February 2024, ONS SRS Metadata Catalogue, dataset, Annual Survey of Hours and Earnings - GB, <https://doi.org/10.57906/x25d-4j96>

Disclaimer

This work was produced using administrative data accessed through the ONS Secure Research Service. The use of the data in this work does not imply the endorsement of the ONS or data owners in relation to the interpretation or analysis.

This work uses research datasets which may not exactly reproduce National Statistics aggregates. National Statistics follow consistent statistical conventions over time.

Acknowledgements

This work was supported by ADR UK (Administrative Data Research UK). ADR UK is a partnership transforming the way researchers access the UK's wealth of public sector data, to enable better informed policy decisions that improve people's lives. ADR UK is an Economic and Social Research Council (ESRC) investment (part of UK Research and Innovation). [Grant number: ES/W005298/1]

Contact

Name: Felix Ritchie, Van Phan

Email: felix.ritchie@uwe.ac.uk; van4.phan@uwe.ac.uk

