ABME Uncertainty Estimation Options

Key Messages of Paper 2
Purpose 2
Recommendation 2
Key Asks of MARP 2
Executive Summary 3
Introduction 5
Background 5
Uncertainty in ABMEs 6
How do we calculate uncertainty in other population estimates at the ONS? 7
Proposed methods for calculating uncertainty 8
Methods options 14
Option 1: Calculate uncertainty in adjustment and projection 14
Option 2: Calculate RAPID coverage by comparing sub-totals with ISA totals. 14
Option 3: Linkage of RAPID and ISA datasets 14
Option 4: In addition to Option 3, measure uncertainty in the data itself, by looking at error rates within the administrative data. 15
Additional recommendations from Methodology and Quality for implementation by MSD: Improve or update the adjustments to limit temporal bias 15
Future Work 15
Bibliography 16
Appendix 1 – ABME Calculation Process 17
Non-European Union (Non-EU) nationals 17
European Union (EU) nationals 17
British nationals 19

Key Messages of Paper

Purpose

In this paper we draw together proposals that have been considered or suggested for uncertainty estimation for admin-based migration estimates (ABMEs).

Recommendation

We describe some combinations of data and methods that could be used. We highlight four main options for uncertainty calculations. Option 1 and Option 2 can be tested without requiring new data; all the others will require acquisition of data and therefore may need collaboration from parties within and outside ONS.

There are some outstanding questions surrounding prioritisation and timeliness; we recommend Options 1 and 2 for focus in phase one of the work, in order to allow production of an estimate of quantitative uncertainty in time for the upcoming ABME publication in May 2023, with additional work into Options 3 and 4 in phase two, between May and November 2023. We recommend focusing primarily on the non-EU component of migration, extending to EU and British nationals respectively, since the non-EU component of migration has the largest impact on total figures. The non-EU component of the estimation process has the most stable methodology and is least likely to change in the near to midterm future. Furthermore, methods comparable to those currently used for non-EU migration are in development for application to EU migration; as such, uncertainty methods designed for use on the non-EU component are likely to be easily replicable for use on EU migration methodology in the future.

Key Asks of MARP

We ask for general feedback on the methods proposed for measuring uncertainty in ABMEs, including the following:

- Are there any key methods you think we should consider which are not detailed here?
- Do the approaches outlined provide an acceptable approximation of uncertainty in international migration estimates?
- Is the uncertainty analysis proposed proportionate to the problem?

Executive Summary

- A new method, based on administrative data, has been developed to estimate long-term international migration. It splits the methodology into three distinct population groups: British nationals, EU nationals, and non-EU nationals. The method captures non-British migrants based on interactions with either border control or tax and benefits administration, and adjusts for timeliness and coverage issues through different administrative or survey data sources. It gives separate estimates for immigration and emigration. Estimation of migration for British nationals currently still relies on the International Passenger Survey (IPS) and is not yet based on administrative data.
- The aim of this paper is to outline the admin-based migration estimates (ABMEs) production process, highlighting key areas that contribute to uncertainty in the estimates, and propose methods to quantify some measure of error for all three population groups detailed above. Measuring uncertainty will be crucial in attaining National Statistics accreditation for international migration estimates. Further benefit from uncertainty measures will be seen in the future, when migration estimates become part of the input to the dynamic population model (DPM), which will use uncertainty estimates to appropriately weight input data when producing estimates of population size and characteristics.
- Counts derived from administrative data have errors for a myriad of reasons including missingness, definitional uncertainty, and measurement error.
- For ABMEs, the two main administrative datasets used are the Registration and Population Interaction Database (RAPID) and Initial Status Analysis (ISA) datasets, developed respectively by DWP (Department for Work and Pensions) on tax and benefits, and by the Home Office on border systems data. The datasets were not initially designed to estimate international migration and so some population groups are structurally missing. As such, data adjustments had to be made to arrive at estimates of long-term international migration, which have a number of underpinning assumptions. Also, the classification process for deciding whether an individual is (or may become) a long-term migrant, according to international definitions, is not always straightforward from the current datasets, and as such is another source of potential uncertainty. These all contribute to the overall uncertainty of the produced estimates. Full details on the process for estimating long-term international migration, including adjustments, can be found in Office for National Statistics (2022a).
- For each component of the ABME process, we outline potential options for uncertainty calculations. These options can be grouped together as such:
    - Methods making use of currently available data sources and methods:
        - Option 1: Calculate uncertainty in adjustment and projection. This process follows the current method and datasets step-by-step, and would be easiest to implement.
        - Option 2: Estimate coverage of RAPID by comparing the raw sub-aggregate totals by nationality and age-sex to ISA data. This uses current datasets and circumvents the adjustments by assuming theoretical perfect coverage of non-EU immigration and emigration in ISA data.
    - Methods requiring new data or adjustments to existing ABME estimation methods:
        - Option 3: Linkage of RAPID and ISA datasets to estimate coverage rates for both datasets. The current methods use different processes to estimate migration of EU and non-EU nationals. A linkage of the two main datasets could help us understand the accuracy of all the steps in the ABME process, including adjustments; it may even be possible to recommend replacement of some of the current adjustments in the international migration calculation process. It is, however, a more resource-intensive and ambitious option.
        - Option 4: In addition to Option 3, estimate uncertainty within the RAPID and ISA datasets themselves, by comparing them to benchmark data.

    - Additional recommendations from Methodology and Quality for implementation by MSD (Migration Statistics Division):
        - Improve or update the current adjustment methods through new and more recent datasets. This requires new datasets from outside ONS; we could combine this option to Option 1 depending on what data is available.
- By May 2023, we plan to implement Option 1, with work progressing on Option 2, in time for a working paper release alongside the next MSD publication on international migration in May 2023. Our preferred methods to be investigated for the second phase of work (May to November 2023) would be ones that will give the most information about the datasets used for migration estimates and therefore the most robust estimates of uncertainty, namely Option 3 and potential extension to Option 4, assuming a successful linkage exercise between data sources. Additionally, we recommend to MSD that an update of the calculation of adjustments should be undertaken, with the potential for new and updated datasets to be added to the current ABME process.

Introduction

Since 2020, new methods have been used by the Office for National Statistics to estimate long-term international migration (Office for National Statistics, 2021a; 2022a). They rely on people's interaction with administrative datasets to estimate entrance to, and exit from, the United Kingdom for 12 months or more for migration purposes. They split the population between EU, non-EU, and British nationals, as until January 2021 these were subject to differing laws relating to travel and hence appear in differing administrative records. (Note that these populations are not mutually exclusive; migrants with dual passports appear in the population associated with the passport with which they apply for visas or services.) The target populations, after capture by the admin datasets, are tested to see if they fit the long-term migrant criteria. The migration totals are then compiled from these sources and several adjustments are applied in order to arrive at the aggregate totals.

The Office for Statistics Regulation (OSR) distinguishes between direct and indirect uncertainty (Office for Statistics Regulation, 2022). Indirect uncertainty is the expression of the knowledge of relative insufficiency in quality about a method or claim. It often takes the form of a qualitative list of sources of uncertainty. Direct uncertainty, in comparison, is the expression of uncertainty about the estimate or fact and would most frequently take the form of an uncertainty range around numerical estimates. A thorough analysis of uncertainty in a process hence starts by listing the biases and limits of the data and proposed methods – indirect uncertainty – and then will aim to quantify as many of these as possible – direct uncertainty. The Aqua Book, a guidance on producing quality analysis for government, further emphasizes the importance not only of understanding causes of uncertainty, but then quantifying and clearly communicating them and their implications (HM Treasury, 2015).

Background

International migration estimates are based on the concept of long-term international migration. The definition of a long-term international migrant is, according to the United Nations, “a person who moves to a country other than that of his or her usual residence for a period of at least a year (12 months), so that the country of destination effectively becomes his or her new country of usual residence” (United Nations, 1998). Using this definition, we wish to make inferences about the total flow of the long-term migrant population in and out of the UK per annum. The problem is hence twofold: capturing the people who are entering or leaving the UK, and correctly identifying them as long-term immigrants or emigrants.

Until 2020, estimates of international migration were calculated using the International Passenger Survey (IPS), but following the pandemic, a new method was devised. It relies on administrative data to measure international migration for non-British nationals, a substantial change in methodology. As a high-level overview, migration into and out of the UK is estimated by observing length of activities within the following administrative datasets:

- Registration and Population Interaction Database (RAPID): developed by the Department for Work and Pensions (DWP), it includes benefits, employment, self-employment, pensions, and in-work benefit, as well as demographic information from the Migrant Worker Scan (MWS), which shows all non-UK nationals who have registered for a National Insurance Number (NINo). These persons are (by definition) 16 or older. It is updated yearly in March and covers the years from 2011 to present.
- Initial Status Analysis (ISA) border data: developed by the Home Office, it captures entry and exit from the UK for all individuals holding a visa to enter the country. It is delivered to the ONS quarterly, with some additional information updated yearly in July, and covers the years from 2016 to present.

- Higher Education Statistics Agency (HESA) data: covers all student enrolments from academic year 2011/12 to 2021/22. It includes information on a student's course and institute of study, as well as demographic characteristics of the student.
- Pay As You Earn Real Time Information (PAYE RTI) – HESA linked data: international student inflows for the academic year ending 2016 to the academic year ending 2018 are linked to their corresponding records in PAYE RTI from April 2014 to April 2019 via the Demographic Index.
- Graduate outcomes (LEO) data: developed by the Department for Education, it captures employment and earning outcomes of higher education graduates in England. It is a one-off dataset from the 2017-2018 academic year.
- Migrant Journey data: developed by the Home Office, it gives the proportion of non-EU nationals that gain British citizenship within a set time period. It is a one-off dataset from 2017.

Uncertainty in ABMEs

The ABME process splits migration estimation in three parts: it uses a different method depending on the nationality of the migrants. Full details can be found in Appendix 1, but a summary of key adjustments and steps where uncertainty may be introduced can be seen in Figure 1.

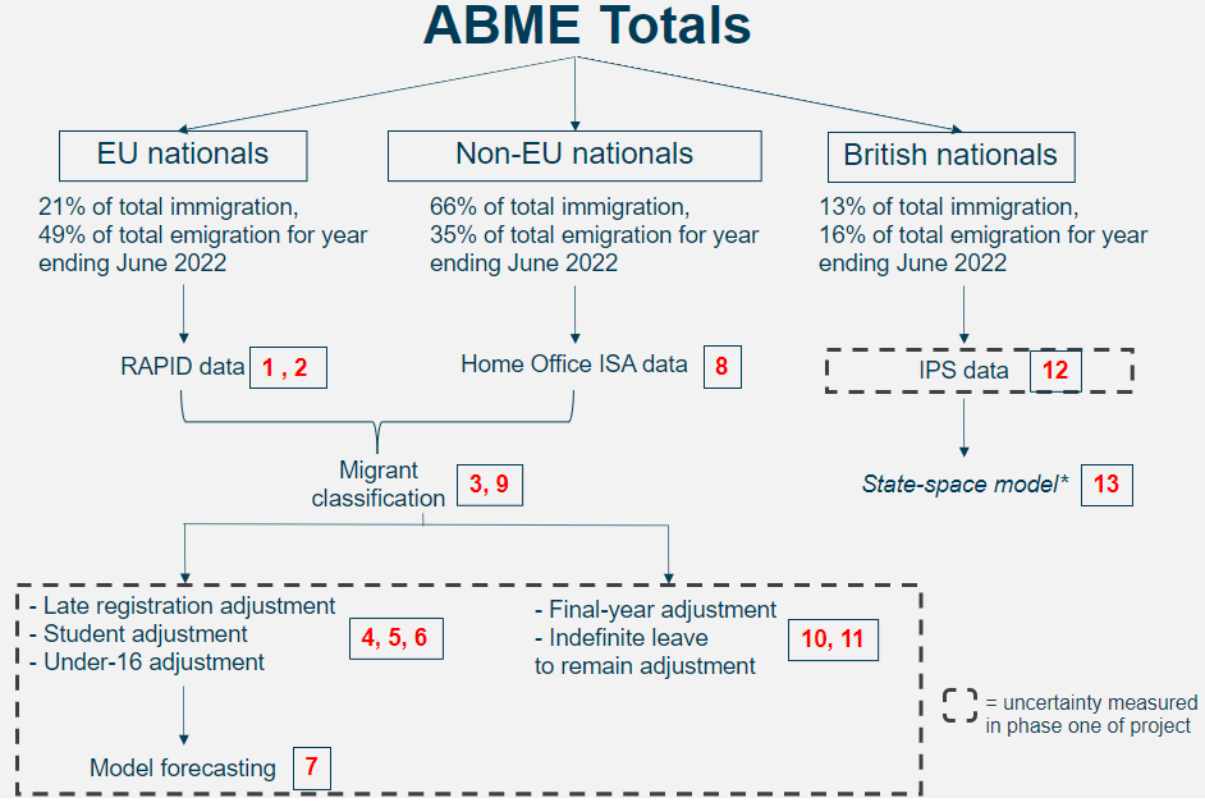

Figure 1: Outline of ABME process. The numbered labels in red indicate some of the points where uncertainty is introduced to the process. They are detailed below in Table 1. Dashed boxes highlight the areas which are being examined as part of phase one of the project. The state-space model is shown in the diagram for completeness; however, the state-space model was only used to infill gaps in the IPS when it was suspended between March 2020 and January 2021, and therefore is no longer used directly in estimating international migration, although the outputs from the state-space model are used in derivation of some other steps of the process.

The ABME method relies on some assumptions at each step:

- 1, 2: Coverage of RAPID data: the largest coverage errors in RAPID data are corrected through the adjustments for under-16s and non-working undergraduate students. There is no method to capture missingness among population groups expected to be in RAPID.
- 3, 9: Migrant classification: self-reported date of arrival is assumed correct and is not further analysed and/or corrected. In general, the administrative variables are assumed correct.
- 4: Late-registration adjustment: Looking at difference between year of arrival and year of registration in previous years of RAPID data, this adjusts current RAPID counts by reallocating a proportion of the counts to the past two years of immigration. This assumes that past patterns hold true for current year.
- 5: Student adjustment: Proportion of working students in previous years is used to correct for current year students. This assumes that the proportion is constant over time.
- 6: Under-16 adjustment: Proportion of EU IPS under 16 to over 17 counts is assumed to be representative of EU migration.
- 8: Coverage of ISA data: no coverage corrections, so coverage of ISA data is assumed to be precise.
- 10: Final-year adjustment: Proportion of immigrants and emigrants in end-of-year months in previous years is assumed to hold for current year.
- 11: Indefinite leave to remain adjustment: A proportion of people that have indefinite leave to remain do not stay in the country. This assumes that past patterns of emigration hold true for current year.

How do we calculate uncertainty in other population estimates at the ONS?

There are two main population estimates that follow similar methodologies, and that we could use as guidelines to calculate uncertainty for ABMEs.

The Mid-Year Estimate (MYE) is an estimate of population totals at local authority level. It relies on the decennial census totals and a cohort component method to adjust for the years since the census. The previous year's population is aged-on by one year and then adjusted for births, deaths, net international migration, net internal migration and special populations (such as members of the armed forces and prisoners). The data for these adjustments come from several sources. Births and deaths come from the General Register Office administrative registers. International migration estimates come from the IPS, supplemented in the case of in-migration by a range of administrative sources (as described above). Data on asylum seekers and their dependants come from the Immigration and Nationality Directorate of the Home Office. Internal migration data are primarily based on the NHS Patient Register. Uncertainty is then calculated for each of the three components that make up the MYE (census-base, internal migration and international migration) and aggregated together. To do so, it uses a mixture of non-parametric and parametric bootstrapping: either a distribution is assumed for the population we are working on, e.g. allocating migrants to local authorities assumes that the previous distribution holds and hence uses its population distribution – or not, and then simulates from the distribution or directly from the sample. The MYE estimation process is followed in a step-by-step fashion, and uncertainty is measured at each point in the process where it could be introduced (Office for National Statistics, 2016). This framework has heavily informed our options for calculation of ABME uncertainty and is reflected through the following sections.

Admin-based population estimates (ABPEs) are produced through linkage of administrative data and the application of a set of rules to estimate the usual resident population. To estimate uncertainty, local authorities are grouped together by similar proportions of population by sex and single year of age. A Generalised Additive Model (GAM) is fitted to each cluster to a variable comparing population

estimates from census to ABPE totals, to produce model residuals. The residuals are then resampled, and confidence intervals are built by taking the x^{th} smallest and largest residual measures; in the case of ABPEs, these are the 2.5^{th} and 97.5^{th} percentiles, or the 26^{th} and 975^{th} of the 1000 simulated values (Office for National Statistics, 2020). This method is more difficult to apply to ABME uncertainty due to the lack of an obvious benchmark in the case of international migration.

Proposed methods for calculating uncertainty

All the effects described in the preceding sections contribute to uncertainty in estimates. We will identify the different components of the ABME process that could lead to biased or erroneous estimates. This modular approach mirrors the approach taken in calculation of MYEs, as described in the preceding section.

Uncertainty measures of non-EU national migration is the first priority in this line of work; indeed, the method is least likely to change in the coming years and is the largest component of overall migration. With changing document requirements for EU nationals recently, it is expected that high quality Exit Checks data will be available for EU nationals as well in the future, and hence that the current method will be amended. As mentioned previously, a revision for British nationals will happen in the future due to changes to the IPS and its future uses.

Table 1 below breaks down the components and processes which are used at present to estimate international migration, following the steps and labels set out in Figure 1. For each step in the process, we list the key areas of under and over coverage, options for uncertainty calculation in that step, advantages and disadvantages of the option, and a minimum viable timeline with reference to the May and November 2023 publication dates for international migration figures by Migration Statistics Division (MSD).

Table 1: An overview of the strengths and limitations of potential methods for estimation of uncertainty associated with each step of the current ABME calculation process, broken down by nationality group. Each step relates directly to the numbered stages detailed in Figure 1.

ABME step and description		**Proposed improvement**			
		Scope	**Strengths**	**Limitations**	**Timeline**
EU nationals					
1. Coverage rate of RAPID	Under-coverage: some population groups are structurally excluded (including under 16s, non-working students, immigrants not working or claiming any UK benefits and asylum seekers), while others will be missed due to differing administrative patterns. Over-coverage: possibility of over-coverage if linkage between RAPID sources is not done perfectly, or from delay between actual exit from the UK and removal from administrative systems.	1a: Linking RAPID and Exit Checks data. Since January 2021, all EU nationals also require visas (or membership of the EU settlement scheme) to enter and stay in the United Kingdom. Requesting the ISA data for EU nationals, and then linking it with RAPID, would let us estimate the coverage rate of RAPID for EU nationals directly.	Aggregate datasets already used by ONS. Linkage would be extremely useful for all of the components of the ABME estimation process.	Requires additional work between government departments. This coverage assessment would ideally be done yearly. ISA data may not be reliable enough for EU nationals.	Not possible for the next May publication: requires access to new data and linkage work.
		1b: Comparing non-EU RAPID and Exit Checks sub-aggregate totals To estimate RAPID coverage rate, we can compare non-EU RAPID and Exit Checks totals aggregated to similar stratifications – for instance, age-sex group by year. This assumes that Exit Checks has negligible under coverage, or that its missing coverage (i.e., known biases which are currently adjusted for) can be excluded from the RAPID totals. This also assumes, like Option 1a, that we can estimate EU RAPID totals by applying trends from non-EU totals.	Datasets already used by ONS. No additional data acquisition needed.	Assumes coverage distribution holds across population groups.	Can be tested immediately for next publication.

2. Data quality of RAPID	The current estimation method assumes that self-reported date of arrival in the RAPID dataset is correct (for identification of target population), as well as that the rest of the data is precise. In particular, self-employment data is incomplete for some tax years, and newly arriving migrants from the EU do not have access to income related benefits at first.	2: Set up a structural equation model to determine measurement error, based on self-employment data. This would be similar to previous work in MQD on other data sources (Office for National Statistics, 2022b). Using RAPID variables that measure similar data critical for Long-Term International Migrant (LTIM) identification – notably self-reported date of arrival – we could estimate uncertainty in the RAPID variables. This would be easier if there is linkage, even only partial, with ISA data.	Methods are based on activity observations. Would be good to estimate errors in these observations.	Unlikely that there are candidate variables that are significant for RAPID coverage and usable here. Necessitates work outside ONS.	Very unlikely to be ready for next publication; although work on data quality may have already started.
3. Identification of EU target population	The current estimation method uses the number of weeks of activity as a proxy for presence in the UK. It hence assumes that this metric is sufficient, and that the activity captured is continuous. Under-identification: does the method in place fail to capture some migrants? Over-identification: should be unlikely, except in the case of delay between actual exit from the UK and removal from administrative systems	3: Linkage of RAPID and Exit Checks data and comparison of methods. If we link the RAPID and Exit Checks data, we can compare the two methods and understand the probability of inclusion of an individual in both methods. A partial linkage should be sufficient for an analysis of the validity and accuracy of the method, but further investigation is required.	Linkage would be useful for each ABME process. Could help us improve current method. Wouldn't necessarily have to be done every year.	Same as 1a/1b.	See 1a. Not possible for the next May publication: requires access to new data and linkage work.
4. Late-registration adjustment	The method calculates the average proportion of people that appear in RAPID one and two years after their arrival date, based on yearly data since 2010.	4: Fit a distribution for the proportion of late registrations. Sample from that distribution and iterate the adjustment. Create confidence interval from the bootstrapping process.	Easy to implement. In line with uncertainty methods of MYE/ABPEs.	Would need distribution to have increasing variance to account for increasing temporal bias.	Ready for next publication.

5. Student adjustment	The method extrapolates from the average proportion of working students in 2016 to 2018, and hence assumes it stays constant. This could lead to either an over or under-correction, for a variety of reasons.	5: Fit a distribution for the proportion of students and hence the size of the adjustment. Sample from that distribution and iterate the adjustment. Create confidence interval from the bootstrapping process.	Easy to implement. In line with uncertainty methods of MYE/ABPEs.	Could need distribution to have increasing variance to account for increasing temporal bias.	Ready for next publication.
6. Under-16 adjustment	The adjustment uses an adult-to-child ratio derived from immigrants (of all nationalities) captured by the IPS. Where IPS data are not available (namely in 2020), a five-year average ratio (2016 to 2019, 2021) is applied. This ratio is then applied to RAPID to estimate the number of under-16s to add to the RAPID estimate.	6b: Estimate sampling error directly from IPS, where estimates and confidence intervals are calculated from a normal distribution.	Straightforward, well established theoretical approach	May be skewed by complex IPS weighting strategy	Ready for next publication.
7. Model forecasting	Necessary to fill a time gap, as the migration estimates are published yearly in May/June but RAPID data is annual for year ending March. The data is hence temporally disaggregated using the Denton-Cholette method (Dagum & Cholette, 2006) and monthly IPS data for EU nationals, with the state-space model for the gap where the IPS was suspended.	7: Calculate uncertainty in IPS indicator series using sampling theory, and combine with disaggregation model variance to produce an approximate uncertainty for the projected months from March-May.			Ready for next publication.

Non-EU nationals					
8. Coverage rate of ISA	Under-coverage: some population groups are structurally excluded if they do not require a visa to enter the country Over-coverage: should be very limited, except due to delay between actual exit from the UK and removal from administrative systems (i.e., leaving before visa expiry)	8: Through a person-level linkage with RAPID data, set up a Dual System Estimator. This necessitates almost no over-coverage for the ISA data, and would give us an estimate of ISA coverage rate. It also assumes that the probability of capture of individuals on either dataset is independent.	Same as 1a/1b.	Same as 1a/1b.	Same as 1a/1b.
9. Identification of non-EU target population	Under-identification: does the method in place fail to capture some migrants? Over-identification: unlikely to occur, except due to delay between actual exit from the UK and removal from administrative systems (i.e., leaving before visa expiry)	Similar issue to 3. The method in place to identify non-EU nationals as migrants through length of aggregated visas could lead to some classification errors. We need linked data with either Census as a one-off analysis of migrants in 2021, or with RAPID, to understand the levels of error in migrant classification.	Could solve both EU and non-EU migration classification at once if RAPID and Exit Checks are linked.	Linkage of Home Office data with other sources may not be possible. The classification method is constantly refined and improved and hence this adjustment will need re-actualisation.	Same as 3.
10. Final-year adjustment	The method provides counts of early leavers and returners (comparing to their visa information) by visa type for three years of data. It then projects final-year immigration and emigration using the previous years' average of this ratio for current year.	10a: Through a person-level linkage with RAPID data, we would be able to better understand the relationship between visa ending and leaving the UK, and we could refine our emigration adjustment. Once we have this, we can assume a distribution and bootstrap from it to create confidence intervals.	Same as 1a/1b. Bootstrapping mirrors MYE/ABPE methods.	Same as 1a/1b. Assumes that RAPID emigration methods are correct.	Same as 1a/1b.
		10b: Fit a distribution to the adjustment ratios by visa type directly, bootstrap from it	No additional datasets or work.	Distribution to sample from not obvious.	Ready for next publication.

		10c: Use the IPS as a benchmark to compare with implemented method, and measure uncertainty by spread/ratio with IPS.	As with 6.	As with 6.	Ready for next publication.
11. Indefinite leave-to-remain adjustment	This method looks at prior counts of emigration from people who have indefinite leave to remain, and had been assumed of staying in the UK. It applies the average proportion of emigration among this group to the current year.	11: Fit a distribution for the proportions of leavers. Sample from that distribution and iterate the adjustment. Create confidence interval from the bootstrapping process.	As with 5.	As with 5.	Ready for next publication.

British nationals

12/13. IPS (and state-space modelling of IPS data)	There are two different sources of uncertainty: the survey uncertainty due to the IPS itself, and the correction uncertainty due to data gaps in the IPS when it was suspended. Both are already measured from sampling theory and as part of the modelling process. The state-space model was only used to infill gaps in the IPS when it was suspended between March 2020 and January 2021, and therefore is no longer used directly in estimating international migration, but outputs from the state-space model are still currently used to form a back series for some other parts of the estimation process e.g. the Denton-Cholette model forecasting in step 7.	Ready for next publication.

Methods options

The ABME process breakdown detailed above in

Table 1 lends itself to different options to measure its uncertainty. The modular nature of the process means that we can estimate uncertainty for different parts relatively independently of each other, and that we can hence improve certain adjustments or coverage estimations without overhauling the entire methodology. With that in mind, we hence present different options, that can be read in increasing level of potential quality, data needs, and resources.

Option 1: Calculate uncertainty in adjustment and projection

For this option, we focus on uncertainty from the size of the adjustments to the baseline (RAPID and ISA) data, and from the projections for data gaps. There are hence no new datasets needed nor any linkage required for this approach, and we are able to implement this approach in time to publish results in a working paper alongside MSD's international migration publication in May.
Similarly to the MYE uncertainty calculation, we would derive probability distributions for each of the adjustments, and bootstrap a range of simulations for each adjustment contribution to overall uncertainty. No quantification is given in this option for the uncertainty due to the coverage rate and quality of RAPID or ISA data.

The options from the table above that follow Option 1 are: 4, 5, 6, 7, 10b, 11 & 12

Option 2: Calculate RAPID coverage by comparing sub-totals with ISA totals.

To estimate RAPID coverage rates at an aggregate level, we compare the totals we receive through RAPID data with those that we receive through ISA data. As all nationalities are captured by RAPID data, we would be able to compare non-EU ISA data, that is currently used for ABMEs, to non-EU RAPID data. We could test different levels of data stratification – by age-group, sex, nationality – depending on the data that we receive. We could then either extrapolate EU totals from non-EU totals, or calculate them separately, depending on the reliability of ISA EU data.
This option doesn't require new datasets and has the added potential of simplifying the ABME estimation process, dependent on outcomes.

The options from the table above that follow Option 2 are as for Option 1, with the addition of 1b.

Option 3: Linkage of RAPID and ISA datasets

A linkage of RAPID and ISA datasets would enable us to estimate coverage rates of both datasets. Both datasets are subject to structural missingness that is relatively well understood. Also, RAPID should be subject to non-systematic missingness, due to people who do not interact with the RAPID component administrations for a variety of reasons. Meanwhile, ISA data can be assumed to have negligible over coverage, and any under coverage can be corrected for using adjustments. We could hence set-up a Dual System Estimator.

A second source of uncertainty is in the long-term migration classification step. Linking RAPID and ISA data would let us check whether people who would be labelled as emigrants from the RAPID emigration methodology have indeed left the UK or not, and hence adjust the totals accordingly.

This option is more costly. It is unclear whether a linkage of RAPID and ISA data would be possible, both for privacy and practical reasons, but communication with MSD and with the data providers is ongoing. The datasets come from two different government sources; prior efforts to link them were unsuccessful due to difficulties in acquiring the data from differing sources at record level.

The options from the table above that follow Option 3, in addition to the core steps from Option 1, are: 1a, 2, 3, 9, 10a

Option 4: In addition to Option 3, measure uncertainty in the data itself, by looking at error rates within the administrative data.

Additional recommendations from Methodology and Quality for implementation by MSD: Improve or update the adjustments to limit temporal bias

In the ABME process, some of the adjustments apart from the Emigration Adjustment are based on analysis of one-off data.
Each of these analyses could either be re-actualised, or done systematically every year. We assume that this would reduce overall uncertainty, as it would reduce the temporal bias associated with each adjustment. However, it would require work from within and outside ONS to make these datasets available for use.

Future Work
Ongoing work on ABME uncertainty is planned to take place in two key phases, approximately aligned to the regular publication schedule for MSD's international migration estimates. Phase one, until May 2023, focuses on producing an initial estimate of uncertainty centred around the effect of adjustments and projections (Option 1). In phase two, between May and November 2023, we are planning further analysis which may include:

- Linkage approaches for assessment of data coverage, for increased quality of uncertainty estimates
- Additional testing of the validity of MSD's assumptions in the estimation creation process

For phase one, we are currently working on preliminary testing of the options discussed in this paper, the outcomes of which have recently been submitted to MaRAG for technical assurance. Following feedback from MaRAG, we plan on releasing a working paper alongside the publication of MSD's international migration estimates in May 2023. The paper submitted to MaRAG will focus on the methodology used, which will primarily be demonstrated as a feasibility study based on data from the previous publication on international migration (from November 2022). The subsequent working paper, however, will be applied to the estimates released in MSD's contemporary publication for May 2023.

Transformation of migration statistics is ongoing, and the methodology used for ABMEs is continually evolving. As such, ongoing work will be necessary to update uncertainty measures in line with updates to the estimation methods, especially to ensure appropriate uncertainty analysis is being used for new data and methods not yet seen in the international migration estimation process.

Bibliography

Dagum, E., & Cholette, P. A. (2006). Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series. In *Lecture Notes in Statistics.* (pp. 80, 82). New York: Springer-Verlag.

HM Treasury. (2015). *The Aqua Book: guidance on producing quality analysis for government.* Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/416478/aqua_book_final_web.pdf

Office for National Statistics. (2016). *Methodology for measuring uncertainty in ONS local authority mid-year population estimates: 2012 to 2016.* Retrieved from https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/methodologyformeasuringuncertaintyinonslocalauthoritymidyearpopulationestimates2012to2015

Office for National Statistics. (2020). *Admin-based population estimates and statistical uncertainty: July 2020.* Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/adminbasedpopulationestimatesandstatisticaluncertainty/july2020

Office for National Statistics. (2021a). *Methods for measuring international migration using RAPID administrative data.* Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/methodologies/methodsformeasuringinternationalmigrationusingrapidadministrativedata

Office for National Statistics. (2021b). *Using statistical modelling to estimate UK international migration.* Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/datasets/usingstatisticalmodellingtoestimateukinternationalmigration

Office for National Statistics. (2022a). *Methods to produce provisional long-term international migration estimates.* Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/methodologies/methodstoproduceprovisionallongterminternationalmigrationestimates

Office for National Statistics. (2022b). *Using structural equation modelling to determine measurement error in different administrative data sources*. Retrieved from https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/usingstructuralequationmodellingtodeterminemeasurementerrorindifferentadministrativedatasources

Office for Statistics Regulation. (2022). *Approaches to communicating uncertainty in the statistical system.* Retrieved from https://osr.statisticsauthority.gov.uk/wp-content/uploads/2022/09/Approaches_to_presenting_uncertainty_in_the_statistical_system.pdf

Appendix 1 – ABME Calculation Process

Non-European Union (Non-EU) nationals

The ISA data combines visa and travel information to link an individual's travel movements into and out of the country. Visas are required for nearly all non-EU nationals. The method for estimation of non-EU migration using ISA data is implemented as follows:

- Identify travellers meeting the definition of long-term migration, first by filtering to exclude those on long-term visit visas (i.e., those on long trips but are not changing their country of residence).
- Use arrival and last departure dates within visa period as approximation for length of stay in UK
- If either first arrival or last departure information is missing, or if the end date on the visa is later than the date on which data is extracted, visa start date or end dates are used as a proxy.
- Visa periods constructed by linking together consecutive or concurrent visas held.
- Previous visa period is looked at to determine if this is a new long-term immigrant, or one who has previously been in the country.
- If no presence is identified in the country during the 12 months preceding first arrival, or the previous visa period had a length of stay of less than 12 months, then this pattern of travel is identified as a new long-term immigrant.
- For the final year of data, we cannot deduce whether someone will stay if they have just migrated, or if someone else will emigrate. We hence adjust the totals through final-year adjustments, using the average ratio of people who have emigrated in previous years in the months that follow the publication.

European Union (EU) nationals

It is not possible to estimate migration of EU nationals using the ISA data because of free movement between the EU and UK up until January 2021, and continued free movement of EU nationals that were granted residency through the EU Settlement Scheme. A different methodology is therefore needed to estimate migration of EU nationals. The current method uses the RAPID dataset created by DWP.

- RAPID provides a single coherent view of citizens' interaction across DWP, HMRC, and local authorities through housing benefits. Anyone arriving in the UK who needs to apply for a National Insurance Number (NINo) to work, claim benefits, or apply for a student loan is captured.
- The DWP and ONS rely on information from the Migrant Worker Scan to identify all non-UK nationals registering for a NINo from 1975 onwards, and hence to define rules of residency in the UK.
- Records are categorised as either long- or short-term migrants by looking for patterns of interaction with the tax and benefits, as well as self-reported date of arrival information.
- Long-term emigrants out of the UK are then estimated; individuals who have no interaction with the RAPID sources of administration after a full tax year are assumed to no longer be resident in the UK.

There are a few caveats to the use of RAPID data to estimate migration. Since it is based on the acquisition of a NINo, some groups of population will be missed entirely. This is detailed in Table 2 below.

Table 2: Data limitations of RAPID data, and how they are accounted for in the EU nationals migration estimate totals

Limitations of RAPID data	How these are addressed
Children under 16 years of age cannot apply for a NINo. Whilst child benefit data are contained within RAPID, it does not provide any evidence of the nationality of the child and is hence not suitable for the analysis of migration into or out of the UK.	In April 2021, we acknowledged that there is a coverage gap in RAPID for those under 16 years old. Between April 2021 and November 2022, migration estimates for EU nationals were only published for those over the age of 16 years. In November 2022, a new adjustment to RAPID was introduced to help fill this coverage gap. The adjustment uses an adult-to-child ratio derived from immigration of all migrants on the IPS. Where IPS data are not available (namely in 2020), a five-year average ratio (2016 to 2019, 2021) is applied. This ratio is then applied to RAPID to estimate the number of under-16s to add to the RAPID estimate. This ratio is calculated separately for both inflow and outflow. This is the first step into looking at this cohort and the migration statistics division are continuing to investigate and develop their understanding of migration patterns of EU nationals under 16 years of age.
People may not undertake any activity that verifies residency, and hence could be considered as having left the UK and thus emigrants.	To identify EU students immigrating into the UK long-term, we use Higher Education Statistics Authority (HESA) data as the best available data source. Our latest method links this to newly acquired HMRC Pay as You Earn Real Time Information (PAYE RTI) data to better understand how many international students are in employment alongside their studies.
RAPID classifies everyone under their nationality at registration even if they have subsequently gained British citizenship. Consequently, migrants who gain British citizenship and subsequently emigrated could potentially be counted in emigration figures for both British and non-British nationals. Furthermore, it could be problematic for counting immigrants, as it excludes people who have never been in the UK previously but who have dual nationality and present themselves at immigration with British documentation.	The UK naturalisation adjustment uses the Migrant Journey data to estimate the proportion of non-EU nationals who have gained UK citizenship within 10 years of their visa being issued.

The RAPID dataset is available yearly in March, while the migration estimates are published yearly in June. To account for the 3-month difference, the missing quarter is predicted using the Denton-Cholette method (Dagum & Cholette, 2006); the predicted IPS series was applied to the financial

year RAPID estimates to both disaggregate it to a monthly series and then to predict this RAPID-based measure for April to June 2022.

British nationals

It is very complex to measure British nationals' migration through administrative data. Indeed, most of them would not appear in administrative data sources from very early on, and hence migration events could be missed. As such, the International Passenger Survey is still the main source of information.

After being stopped for several months due to COVID, the IPS was reinstated in January 2021, and we use these data as our estimates for January 2021 to June 2022. To cover the period when the IPS was suspended (March to December 2020), we use the state space modelling (SSM) time series analysis (Office for National Statistics, 2021b). This takes the available IPS and administrative data and uses the relationship between them to estimate the missing IPS data. We assume that the pattern of British and UK nationals' immigration to the UK is equivalent to non-EU nationals' emigration from the UK, as measured by ISA, and the other way around. The validity of this assumption is unclear and may require further clarification. However, applying confidence estimates to historical data is not within the current scope of this work, and since this modelling period was temporary, it's unlikely to be relevant to future outputs.