# Disclosure control proposal for Future Population and Migration Statistics

## Contents

## Introduction

All outputs from the Office for National Statistics (ONS) are subject to the Data Protection Act 2018 and the Statistics and Registration Services Act 2007, and so must not contain information that identifies an individual, household or business. Our responsibility to protect confidentiality is also made clear in the UK Statistics Authority's Code of Practice for Statistics.

A number of statistical disclosure control methods and processes are used to protect ONS outputs, including those in tabular form and record-level data. These methods introduce uncertainty in the values or combine categories, and while this provides disclosure protection, it also reduces the utility of the outputs (see Hundepool, A. and others (2012), 'Statistical Disclosure Control'). There is a balance to strike between the two, usually based on a qualitative judgment of risk. For example, Census 2021 outputs had several disclosure control methods applied to reduce the risk of identification and disclosure, namely targeted record swapping, cell key perturbation, and disclosure checks.

As we consider the future of population and migration statistics and the wide range of outputs from such a system, including alternatives to traditional censuses, we also need to consider how best to protect these new data assets against disclosure. This methodology outlines some of the early thinking around this decision, including the research we are proposing, the methods we are considering to protect against disclosure in future population and migration outputs, and some of the factors and practicalities in making this decision. Notably, the data structure and anticipated frequency of some of the low-level outputs may make methods such as targeted record swapping impractical, in which case cell key perturbation would be used as the main form of protection.

## Future Population and Migration Statistics data and outputs

The Future of Population and Migration Statistics (FPMS) covers a large programme consisting of a variety of statistical products, as outlined in the consultation. It seeks to make greater use of administrative data to produce improved population and migration figures. Some of the outputs will be highly aggregated estimates, such as the population of local authorities. The Office for National Statistics (ONS) also plans to publish multivariate statistics at low levels of geography.

Because of the potential high coverage of administrative data, the range of topics, and the low level of geography, the disclosure risk of these multivariate statistics would be similar to a census. However, there are several significant differences between traditional census data and data use in estimates built around administrative data. This may mean that current disclosure methods applied to census are less effective or cannot be used.

The most obvious difference is the source of data and how they are made. Administrative data are drawn from operational uses of other government departments.

The outputs which would present the biggest concern for disclosure are those produced directly from a linked set of administrative data. Such a dataset could contain one row per record, with each record representing a real individual. This scenario makes the link between a person and the data more tangible, so that a person may feel they could identify themselves in the published statistics.

A linked dataset may also contain data on a wide variety of topics since it may originate from many sources. Characteristics may also be obtained by linking information from survey data, although these would pose a lower disclosure risk since the data would come from a relatively small sample. This means that many of the values would need to be estimated or modelled, with uncertainty in the true values.

Our main form of protection for census data was targeted record swapping, which involved swapping “risky” households with each other, or, in practice, changing their geographical information. This may not be practical for administrative data if there are several sources of geographical information, one from each data source that was linked, all of which would list the “correct” or pre-swapping geography. Changing these values may not be possible if they are held in multiple versions of the data, particularly if there are several or unknown stores of the pre-swapping geography, or if the data including the geography are updated regularly.

Another difference is the outputs we expect to produce. In particular, we expect that our future use of administrative data will enable more timely and frequent outputs. This introduces a new risk arising from longitudinal data: small differences in counts over time.

For example, in a given release, there are four persons with particular characteristics in an area. Then, in the next release, there are six, which reveals that two new people have moved into the area. Or, if an intruder knows there are two new arrivals, this reveals their characteristics.

Differencing in this way produces a “net” difference. It could be that in the period between outputs a large number of residents migrated out of the area, and a large number also migrated inwards, resulting in a net difference of two. In this case, the differences in outputs represents a much larger group of persons rather than just two individuals. The extent of this risk will depend on the geographical level and frequency of outputs, with smaller areas and more frequent outputs more likely to contain small longitudinal differences, perhaps of one migrant person or household.

Our current proposal is to also use cell key perturbation, as we did for Census 2021. Perturbation adds “noise” to the outputs, which makes small changes to the values; for example, adding noise of plus one so a count of four appears as a five. This is particularly effective at protecting against “differencing”, where multiple datasets are combined or “differenced” so that users cannot be sure whether small differences represent a real person or are caused by the perturbation.

Unlike Census 2021, it is possible we may not be able to add protection by swapping to the administrative data. In this case, more protection will need to be applied through cell key perturbation in the form of a higher rate of perturbation. The more frequent outputs we expect, the higher the risk of longitudinal differencing will be, and the higher the rate of perturbation will need to be.

Using the same “record keys” (that is, pseudo random numbers which determine which cells receive noise) for each release would provide very similar perturbation choices for each set of outputs. This is not desirable in a longitudinal set of outputs as it highlights changes over time. However, we do want consistent perturbation within the same release, so for example producing the same table twice for a given time period will provide the same results.

One proposal for applying different perturbation for each release but consistent perturbation within the same release is to add an index for the release number or release date to the record keys. For example, each record will receive a “base” record key, and for each release, the record key for that time period would be the base record key plus a number generated from the month or year. A simpler alternative would be producing a new set of record keys for each release, but this would use data storage unnecessarily and may be confusing for analysts working with the data.

The last disclosure control method applied to census outputs is the automated disclosure checks. Users can select which variables they would like to be cross-tabulated using our Create a custom dataset tool. The disclosure checks determine which geographical areas the data can be made available for based on measures of disclosure risk for that area, such as sparsity and low counts in marginal totals.

The disclosure checks enable a large degree of flexibility in outputs and cross-tabulation of data from different topics while avoiding releasing data for areas that have a high disclosure risk. This approach has been successful for census, although the availability of data will have to increase before this approach is possible for future population and migration statistics.

## Methodological considerations

### Targeted Record Swapping

Targeted record swapping was the main form of protection for both 2011 Census and Census 2021. Households considered most at risk of identification or disclosure are swapped with “matched” households in another geographical area by swapping their geographical information. The matched households are chosen to be similar on basic characteristics, such as the number of people in a household, and most swaps are performed within a local authority.

In this way, considerable doubt is introduced to low counts at low geographies, while higher level information at the local authority level is mostly unaffected. Another benefit of this approach is that swapping can be targeted to protect specific risks, for example targeting low marginal counts on certain variables which pose a greater risk in a flexible outputs environment.

As mentioned in the section ‘Future Population and Migration Statistics (FPMS) data and outputs’, a set of linked administrative datasets would be processed and stored by different means to census data. A linked set of multiple administrative sources would be expected to contain multiple sources of geography. It may also contain disagreements between sources of geography as a result of migration which has not been updated in some datasets, or because of versioning or uncertainty in linkage. To effectively swap households, outputs would need to be consistently produced from a single source of geography which has swapping applied.

Sets of outputs will be produced annually or biennially, although it is possible that future outputs will become more frequent where the data allow this and there is a user need. This raises two potential issues. Firstly, if outputs are more frequent, it may be more difficult to apply the swapping process in a timely manner. A simplified swapping process may be needed to achieve this.

Secondly, after several years, several sets of data and characteristics may be held in some form in the data model. How these data are stored and version controlled will affect how the swapping needs to be applied. It’s important that data deliveries and updates do not overwrite the changes made through swapping to the geography variable, which would remove the protection applied.

## Perturbation rate and parameters

Arguably, the biggest methodological question that needs to be answered is what level of perturbation should be applied to the data. This requires an assessment of the disclosure risk of the data and assurance that perturbation adequately protects against these risks.

The level of noise added firstly needs to be sufficiently high to prevent “unpicking” of the changes made. If very few changes are made to the cell counts, comparing the totals across different tables can highlight which cells have been perturbed and by how much. In this case a motivated intruder could effectively remove any protection added.

The main factor in the parameterisation is the anticipated level of disclosure risk, and different risk scenarios that we expect from the outputs. This includes the risk of directly identifying individuals in small counts, differencing multiple outputs to produce more granular breakdowns of data, and, as mentioned, in the case of frequent releases, differencing outputs from different time periods highlighting longitudinal differences. These risks will depend on the granularity, breadth, and frequency of outputs, all of which are expected to increase over time.

Lastly, we need to consider the utility of the data and ensure that the perturbed outputs still meet user needs. Disclosure control always needs to balance the risk and utility of data, aiming to provide outputs with sufficiently low disclosure risk but with maximum utility (or in some cases sufficiently high utility and minimum disclosure risk).

Once a perturbation rate has been chosen, the perturbation parameters will be stored in a “ptable” file. The ptable will need to be available for outputs production, most likely involving an ingest and management process in a secure environment.

For Census 2021, perturbation of zeros was applied. This provided the possibility of cells that were zero being perturbed upwards to non-zero counts. This introduces a good source of uncertainty as small counts in the data may not represent real individuals but instead be the product of perturbation. This aspect of the perturbation can be complex to apply but we intend to apply this to FPMS data.

**Record keys**

We intend to apply Laplace-shaped noise though the cell key perturbation, in line with international best practice. For further information, see Abowd (2023) ‘Confidentiality Protection in the 2020 US Census of Population and Housing, Annual Review of Statistics and Its Application’.

To enable this change, we intend to increase the range of record keys to allow more precision in the noise distribution. Previously, record keys and cell keys were integer values uniformly distributed between 0 to 255. Each cell key contains a perturbation value, which could either leave the cell unchanged, reduce the cell value by one, increase the cell value by one, and so on. By changing one key, we can change the rate of perturbation only in plus or minus 0.4% increments.

Using an increased range of 0 to 4,095 allows a choice of more precise rates of perturbation, in plus or minus 0.02% increments. It also allows large perturbations to be applied with low probabilities, introducing the possibility of large changes in cell values, although keeping these changes infrequent to preserve utility. Any range of record keys can be used for perturbation, though larger ranges increase the file size of the ptable (proportionally).

Although record keys are generated as random uniform numbers, record keys should stay consistent once generated. Providing the same record keys and ptable to the perturbation method ensures the results are consistent and repeatable, so that the same piece of analysis carried out at multiple times, or by different users, will produce the same result.

The simplest option for producing record keys is to add record keys to datasets when they are ingested into the secure environments in the Office for National Statistics (ONS). But how the data are linked, stored, and managed has the potential to affect the cell key implementation. For example, if new versions of data deliveries are supplied separately, rather than as updates to the same file, we may need to produce record keys for each iteration and develop a method to produce them consistently. One option to do this is to store the record keys on a separate dataset and link the record keys to the new versions of data for each iteration.

### Minimal viable product versus future planning

The current improvements we are making to population and migration statistics is part of a long-term programme which will evolve over time, with new outputs and variables being included. This presents a choice: to prepare a system of protection for the outputs we expect in the future or to protect each set of outputs on its own merits, changing the extent of the protection each time. We propose to take the former approach and provide a system of disclosure methodology based on the expected long-term scope, coverage, granularity, and frequency of outputs.

This is analogous to a “worst case” scenario which requires the most protection. This approach may overprotect simpler early FPMS outputs. However, we expect this will be preferable for users, as the alternative approach of applying the minimum required protection at each stage of the programme would involve frequent changes to the statistical disclosure control for each set of outputs, with each increase in outputs accompanied by an increase in the required disclosure control, along with the associated communications of what has changed and why.

It would also involve an assessment of how much the level of risk has changed and therefore how much additional protection is needed. Depending on the size of the changes, the changes in risk level could be gradual and small and the data-utility benefits of a reduced level of protection (for example, a lower rate of perturbation applied) are likely to be minimal.

### Geographic detail in the Future of Population and Migration Statistics

One major factor in the level of disclosure risk is the level of detail available in outputs, including the level of geography, with “lower” or smaller areas of geography presenting a higher disclosure risk. The proposed approach for many topics is to produce outputs at the Lower layer Super Output Area (LSOA) level of geography, which contain between 1,000 to 3,000 usual residents. As we develop our approach for producing estimates for smaller areas, we will consider the effect these outputs will have on disclosure risk.

Whether outputs will also be made available for alternate geographies, such as wards, postcodes and parishes is unclear, and a decision is not expected for some time. Having the same set of data available for several sets of overlapping areas presents a greater risk of disclosure by differencing and may affect the level of perturbation required.

### Modelling estimates

Some outputs which form part of the FPMS programme, such as those from the Dynamic Population Model, will be modelled estimates. These figures are generated from Bayesian models using several data sources as inputs. The disclosure risk of such outputs are subject to debate.

There is a clear separation of the estimates from the individuals contained in the datasets, and often a known, measurable level of uncertainty in the model, which can make it hard to argue that a real individual could be identified within modelled outputs. We will need to consider the disclosure risk of such modelled outputs and what level of protection is necessary. It is likely that for basic demographics of age by sex at local authority level, the uncertainty from the modelling will be sufficient to protect against the risk of disclosure, and no additional protection will be needed.

## Future developments

This methodology has outlined the main methodological and practical considerations for protecting Future Population and Migration Statistics (FPMS). It has outlined the main disclosure methods used for Census 2021, record swapping and cell key perturbation, which would be suited to protecting statistics as the population and migration outputs evolve, including multivariate statistics at low levels of geography if this becomes possible.

Many of the challenges relate to aspects of the new system that are unknown, or are likely to change over time. Similarly, this methodology makes some assumptions about outputs that will become available in future. In response to this, the Office for National Statistics (ONS) will need to prepare for the highest level of disclosure risk, occurring with the most detailed, frequent outputs, and keep up to date with the FPMS plans and systems.

Specifically, our next steps are to:

- carry out a literature review of methods used by other national statistical institutions for administrative data outputs
- investigate how the data model works with linked administrative and survey data, and how updates or data deliveries could affect variables held (such as geography, record keys, and so on)
- attempt a pilot application of record swapping to ensure the method is feasible, with the data model in mind

- discuss the anticipated content and frequency of outputs, including the content of the consultation, which will determine what disclosure control is required
- develop a system for producing record keys for FPMS data (composed of multiple linked datasets)
- decide a rate of perturbation based on the expected scope and frequency of outputs, and considering whether the data will have protection from swapping
- create the parameter files that apply this protection and make these available next to the FPMS data ready for output production