28 October 2019

United Nations Economic Commission for Europe

Conference of European Statisticians

Work Session on Migration Statistics
Geneva, Switzerland
29–31 October 2019
Item 4 of the provisional agenda
Use of longitudinal data for migration statistics

An error framework for longitudinal administrative sources; its use for understanding the statistical properties of data on international migration

Note by the Office for National Statistics, United Kingdom*

Summary

The Office for National Statistics (ONS) is committed to increasing its use of administrative data in the production of population statistics. The current focus is on improving international migration estimates, using administrative sources. ONS is engaged in a consultation exercise to understand better the migration statistics that our users need. We are also collaborating with the Home Office to understand how the government's plans to build a new end-to-end border and immigration system may provide new opportunities to use administrative data sources to measure international migration.

As part of this programme of research, methodologists working alongside migration experts within ONS have been investigating the statistical properties of administrative data.

We found that the Statistics New Zealand framework for reporting administrative data quality provides a valuable organising structure for our quality investigations. We have extended the framework to reflect administrative data that are longitudinal in nature and potentially have a multi-level structure. We are considering how statistical error propagates through the data over time, incorporating these insights into our conceptual thinking. Our knowledge of

*Prepared by Louisa Blackwell and Nicky Rogers, ONS

statistical error structures in the administrative sources is informing the design of quality indicators and further linkage, including longitudinal linkage. This paper describes our research and considers how we might optimise the design of integrated datasets, using these new insights and quality indicators.

Our error framework will be helpful for others seeking to design new longitudinal datasets using administrative data, and for those seeking to unravel the statistical properties of existing administrative sources that are longitudinal in nature.

I. Introduction

1. This paper describes an error framework that we have developed to help us to understand the quality of longitudinally-linked administrative data. Building on the framework developed by Statistics New Zealand for administrative sources, we present one framework for single longitudinal sources and one for multiple sources that are longitudinally linked. In the fullest version of our framework we seek to identify all the sources of statistical error that may appear in the data as they progress through administrative systems and statistical processing. We also show a simplified version of the framework, where we summarise the sources of error at each stage of the data journey. Some of the error we seek to understand is at the conceptual level, because it is not easily measurable, while some error can be quantified. In this paper we describe some quality indicators that we have developed to quantify error. This is work in progress and we are keen for your comments and views on the utility of this framework to you.

II. The context for this work; the ONS Population and Migration Statistics Transformation Programme

2. The Office for National Statistics (ONS) is the United Kingdom's largest independent producer of official statistics and is its recognised national statistics institute. ONS is responsible for collecting and publishing a broad range of statistics related to all aspects of social life; the economy, population and society at national, regional and local levels. Every ten years we also conduct the census of population for England and Wales.
3. Alongside many other National Statistical Institutes (NSIs), we are committed to increasing the use of administrative data in the production of our statistics, to reduce costs, reduce respondent burden and to improve the quality and granularity of our statistics. ONS is currently progressing a programme of research which aims to place administrative data at the core of our population statistics system by 2020. A critical component of that system is the estimation of international migration. We are working in partnership with colleagues across the Government Statistical Service to transform the statistics available, to improve the evidence base for migration, and to research and understand migrants' experiences in the UK.

A. Transforming international migration statistics

4. ONS currently produces statistics on short-term and long-term international migration and the migrant population, largely following United Nations definitions for these. Our estimates are primarily based on the International Passenger Survey (IPS), which has been collecting information from people entering and leaving the UK since 1961. ONS has accepted for some time that the IPS is being stretched beyond its originally-intended use. The survey-based

estimates do not provide local authority- level outputs, for example, at a time when our users are requesting more granularity in our data on international migration. The changing policy context means that decision-makers and the public are requesting more evidence on migration, including the impact of migration on the economy and society, at local levels. High demand includes calls from parliamentary committees, requesting better evidence on migration. For these reasons we are pursuing a programme of research and in January 2019 we published a research engagement report, updating our users on our population and migration statistics transformation journey.

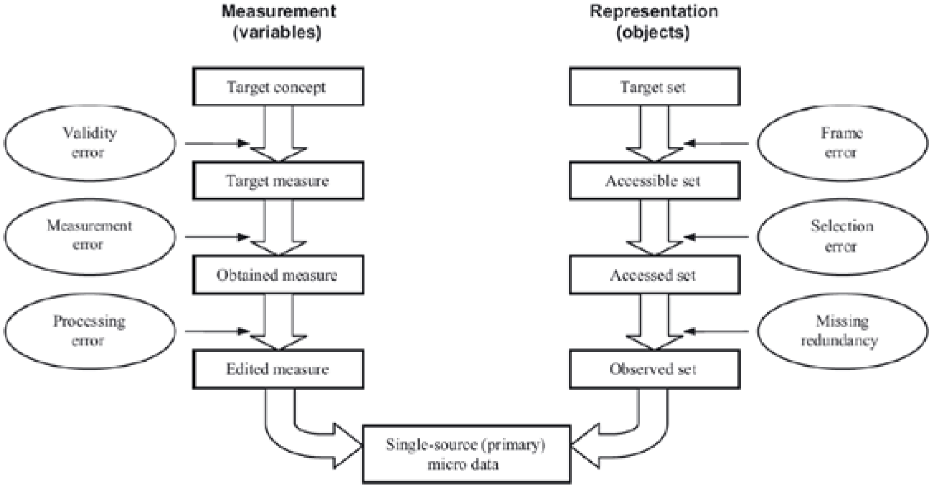
5. We are considering all possible sources of information on international migration, and on the experiences of international migrants. We are looking into using administrative data from the National Health Service, from the Higher Education and Statistics Agency, from Her Majesty's Revenue and Customs, from the Department for Work and Pensions and from the Home Office. We expect that the transformed international migration estimates will draw on several sources, enhanced through record linkage.
6. No single administrative source will meet all our data needs. We are taking care to examine the quality of all administrative data sources very closely.

III. The Statistics New Zealand error framework for administrative sources

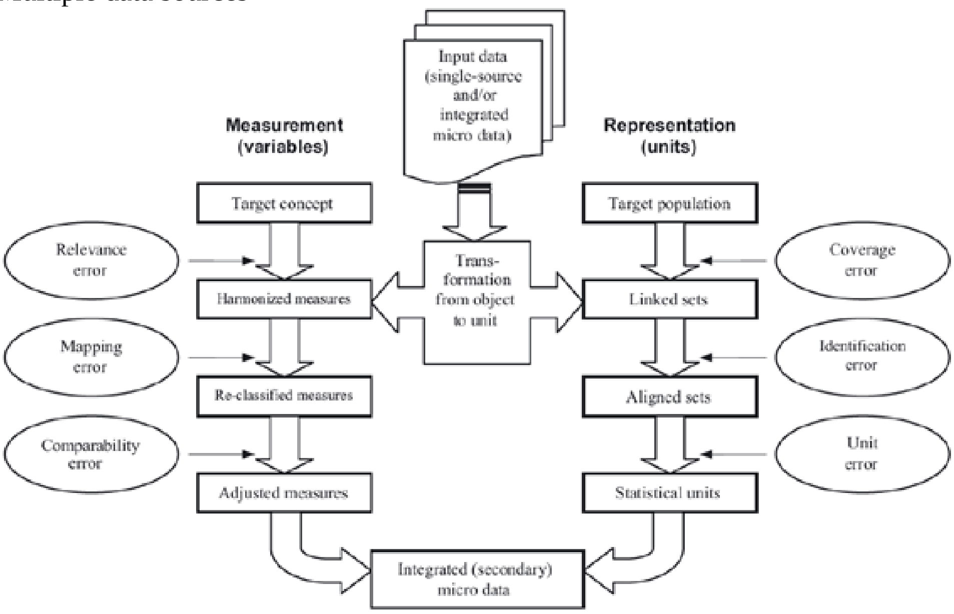
7. In this research, we have drawn heavily on the framework for measuring administrative data that has been developed by Statistics New Zealand (SNZ). The SNZ approach draws on the systematic identification of sample survey error described in Groves et al. (2004), adapted for administrative data sources by Zhang (2012). The framework is designed to help statisticians to understand the strengths and limitations of data, including administrative data, survey data and combinations of the two. The framework proposes a staged approach to understanding data quality, looking first at individual datasets (phase 1), before assessing the datasets created through the integration of different datasets (phase 2, see Figure 1).
8. We found this taxonomy of error to be highly valuable in helping us to analyse and understand potential sources of error in administrative data. We have also consulted the literature on longitudinal error, see for example Lynn (2001 and 2009), Lynn and Lugtog (2016) and Tourangeau (2018). The complexity of administrative data processing, and the fact that it is carried out beyond ONS, mean that we need quantitative indicators of data quality to support the use of this information for estimation purposes.

Figure 1
The error framework used by Statistics New Zealand from Zhang (2012)

Phase 1
Single data sources



Phase 2
Multiple data sources



IV. Our longitudinal admin data error framework

9. Much of the population statistics transformation research in ONS involves longitudinal linkage by diverse teams and we have tried to create a simplified version of the framework which is also labelled as intuitively as we could manage. We welcome comments on whether this is desirable or has been achieved. We have attempted to give a fuller account of the potential sources of error at each processing stage, drawing on what we have observed in our data, and in an attempt to give the framework broader utility within ONS. Aside from re-orienting the infographic, perhaps the biggest difference between our framework and the SNZ one is that while we recognise the conceptual difference between objects and their attributes, we also acknowledge that the data come to us as a coherent dataset already formed, we have little

influence over the design. We still consider these separate dimensions of the data. But, rather than representing them as two parallel conceptual flows, we have at the heart of our framework the dataset which we have not designed, *it is what it is*, but which we then transform through our data integration processes.

10. The framework suggests a staged approach to understanding data quality, looking first at single datasets before assessing error in the production of datasets that are the result of integrating two or more single sources. Figures 2 and 3 describe the generalised error frameworks for single and multiple sources. Tables in Appendices 1, 2 and 3 define the errors at each stage.
11. The single source framework identifies four different stages of the data journey – target data, accessible data, accessed data and processed data. Errors are represented as a conceptual difference between data at each of these stages. Target data is conceptual – the ideal data to be collected and so errors between target data and accessible data are conceptual. Errors are split between objects and attributes: errors occurring to objects relate to the entity the data are for, be that people, events, businesses etc. Errors relating to attributes relate to what you are measuring for the objects. Errors for both objects and attributes can affect each other (represented by a double arrow between the two). Not all types of errors will be applicable to each source of data and there may be other sources of error identified, particularly for processed data.
12. Our initial thinking was that error propagates through the framework, building at each stage. We believed it to be cumulative. But following further research we believe it is more nuanced than that. The processors of administrative data may proactively manage the quality of data as they pass through their systems and therefore mitigate accumulation of error. While error can accumulate over time, there is also the possibility that compensating errors may complicate matters. Compensating errors, such as undercoverage which is masked by duplicates in the data, may be invisible in cross-sectional comparisons. They will, however, compound in longitudinal analysis.
13. The framework also incorporates longitudinal error created by the collection of data over time – seasonality/periodicity, attrition and censoring (see Appendix 2).

Figure 2
Single source error framework

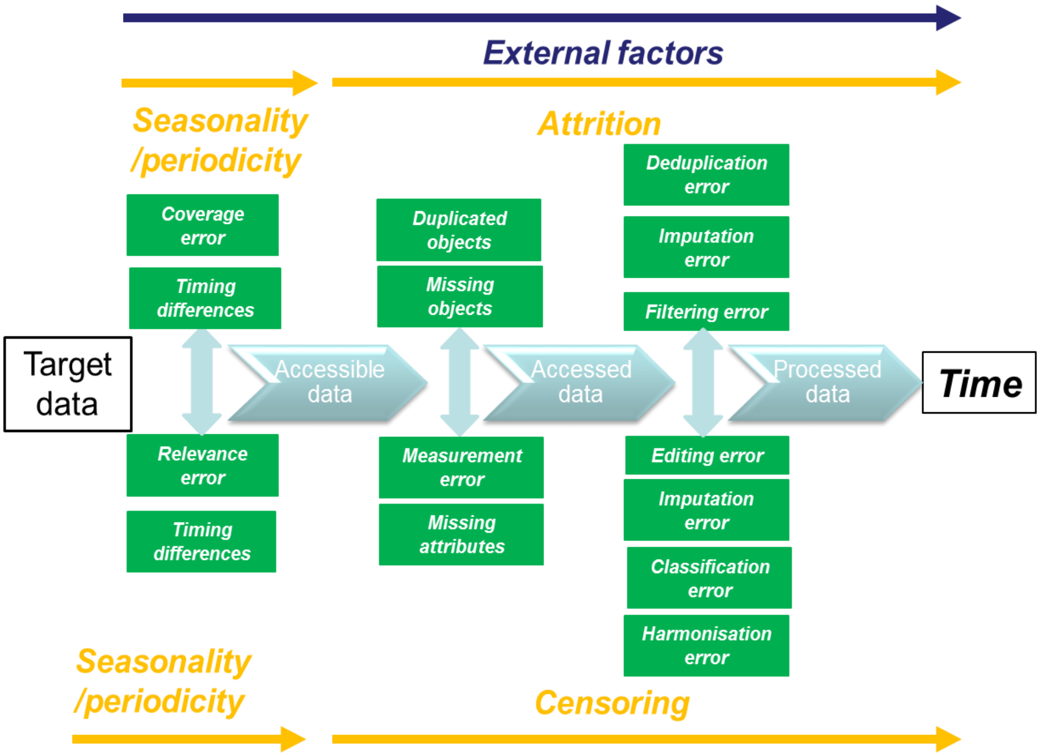
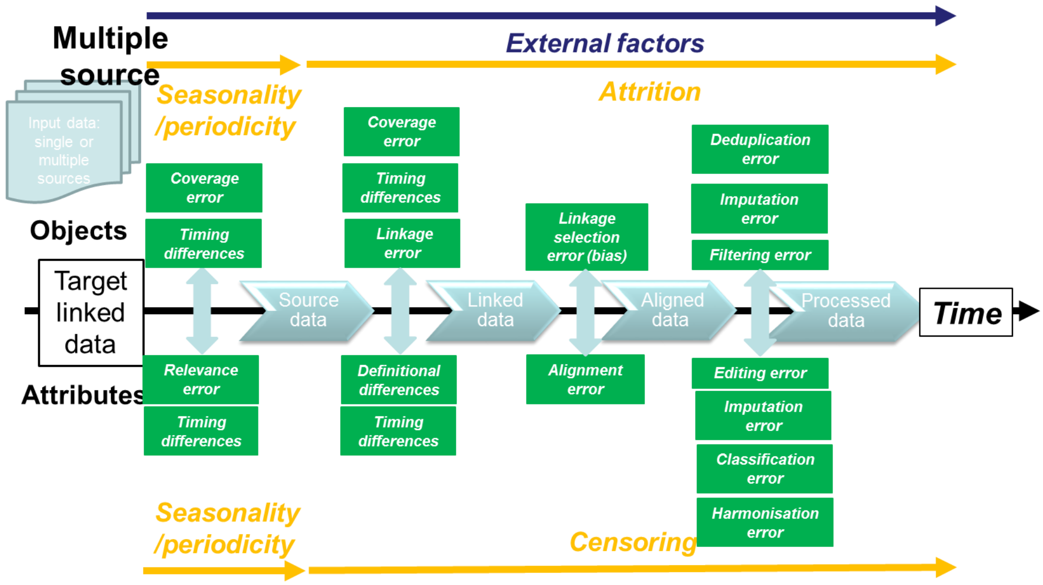


Figure 3
Multiple source error framework



14. The multiple source framework (Appendix 3) is for the linkage of multiple data sources. It consists of five stages – target linked data, source data, linked data, aligned data and processed data. Errors are represented as a conceptual difference between data at each of these stages. Many of the errors in the multiple source framework are conceptually similar to those in the single source framework. The main difference is the fact that we are measuring errors between

the source datasets and the ideal target linked data (rather than target data) as well as errors between the source datasets to be linked. These conceptually similar errors have the same name between the single source and the multiple source frameworks. For example, timing differences, coverage error, relevance error, imputation error, selection error and processing error. Target linked data are different to the target data for each individual source. Your target linked data are likely to be specific to the group of objects that you hope to measure through linkage of multiple sources. For the final stage (processed data), processing may have occurred in the single source or in the multiple source framework. For example, imputation may have already occurred at single source, or it may be done after linkage.

A. Use of the error framework in design

15. The purpose of applying the administrative data error framework is to identify and examine sources of error to inform statistical design decisions in the production of further linkage. We note the interaction between different sources of error, and the need to seek data designs that are optimal for the intended purpose of the linked dataset. Sometimes end-user requirements drive decisions about how the data are processed, for example there could be an over-riding concern to avoid false positive matches. Something to bear in mind is the trade-off between linkage, coverage and imputation error. In our experience, records that have poor quality data, possibly through measurement error, are also harder to link. Perhaps this is due to the quality of the identifiers used for linkage. One option is to develop sophisticated record linkage methods to minimise false negative matches (therefore accepting more false positives) and in this way you could aim to maximise the coverage of objects in the linked dataset. However, there is a possibility that the attributes that relate to these objects are also of poor quality, and will generate either missingness in the attribute fields, or will require imputation. Imputation is often undesirable in longitudinal data, since it can introduce spurious outcomes. The avoidance of missing data and imputation may be the over-riding concern, rather than maximising coverage.
16. We also note the interaction between errors in objects and in attributes, and between the single and multiple source datasets.
17. The administrative data error framework is helpful in evaluating the quality of linked datasets. Figures 4 and 5 provide a simplified version of our framework, with errors at each stage grouped within an over-arching error term. On these figures we show some of the quality indicators that we have been developing to measure dataset quality.

B. Use of the error framework for data quality assessment

Figure 4
Single source error framework summary

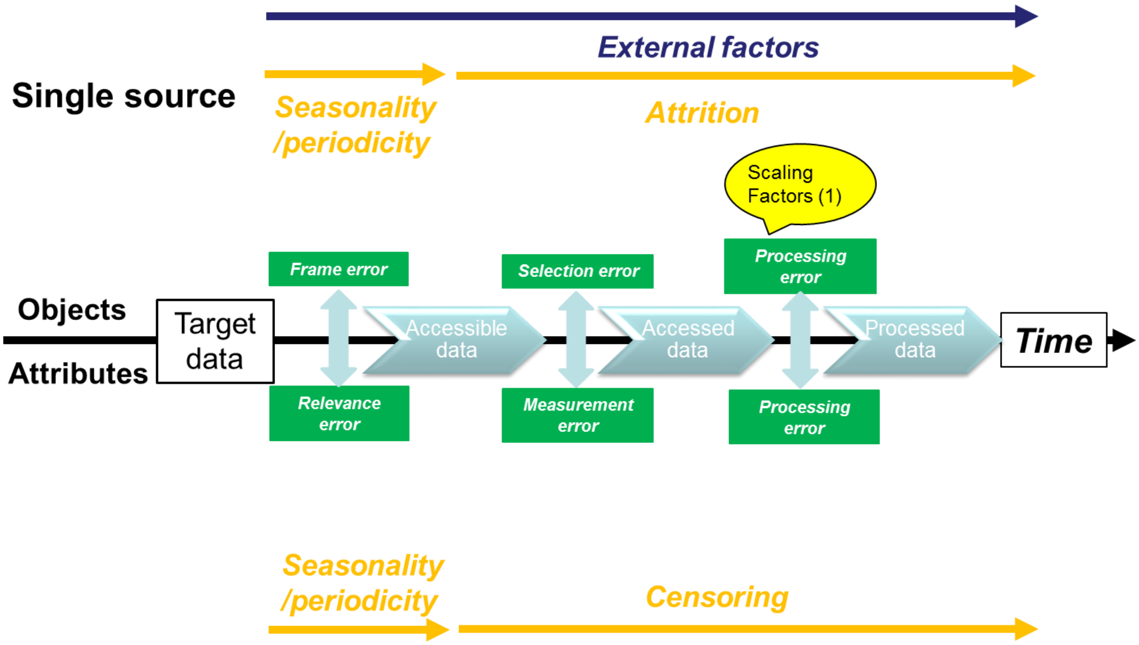
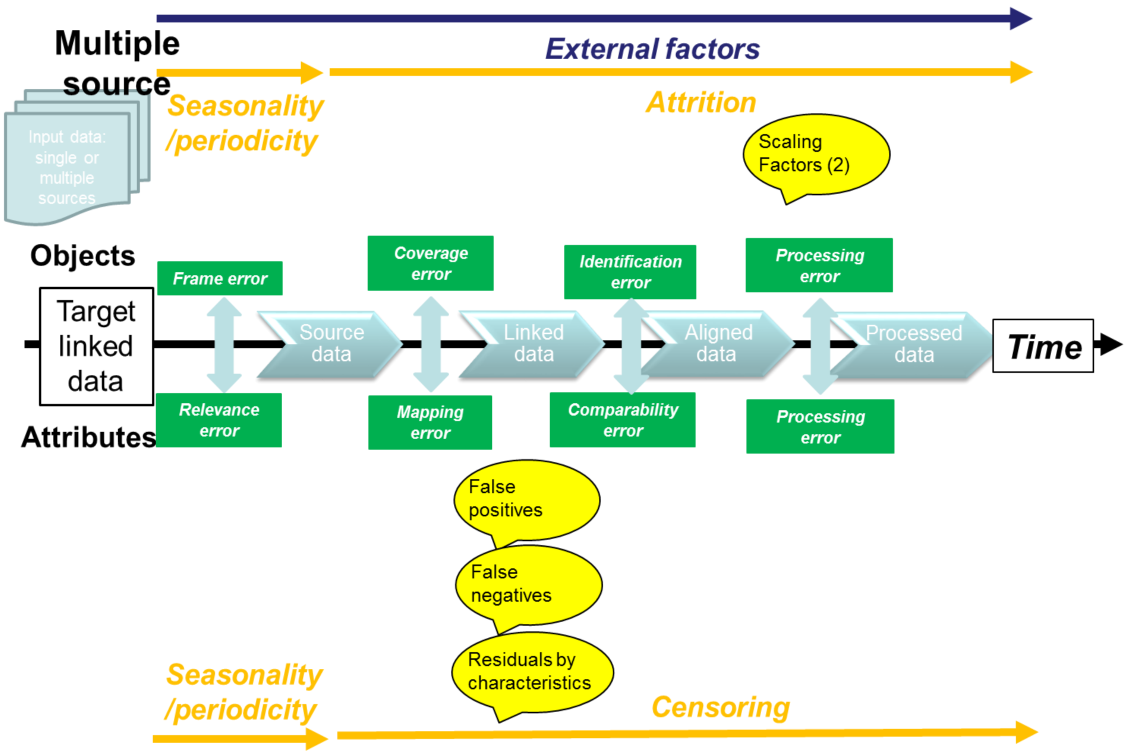


Figure 5
Multiple source error framework summary



C. Scaling factors

18. Scaling factors benchmark the administrative data against other sources. Often it is not possible to equivalise the administrative and comparator data, perhaps because of definitional differences or because of known coverage differences. Even if both sources have systematic differences or errors, the relationship between them over time is illuminating, shining a light on both variance and potential bias. In other work, logged scaling factors have been modelled, as a means to attempt to measure statistical uncertainty (see ONS 2017).

D. False positives/negatives

19. A clerical examination of linked records can be used to check whether there are any matches that have been made that appear to be erroneous. This provides an indication of false positive linkage error. A similar exercise can check whether there are falsely unmatched records in the residuals, to estimate false negative matching.

E. Analysis of residuals

20. When two sources are linked, it's important to understand the representativeness of the resultant linked dataset. One way to begin to approach this is to consider the characteristics of the matched dataset, and whether these are the same or different to the unmatched residual records.

V. Summary

21. ONS is transforming its population statistics system and a key strategic aim is to make better use of administrative data, reducing our reliance on survey data. It has long been recognised that we need better data on international migration. The International Passenger Survey is being stretched beyond its original purpose and cannot support the more granular statistics that our users need. We are urgently seeking to enhance our estimation processes with administrative data. To achieve this, we need to understand the statistical properties of administrative data.
22. The framework for single- and multi-source integrated data in use by Statistics New Zealand (SNZ) has provided a valuable taxonomy of potential data error, and an organising framework for understanding administrative data quality. The framework we present here extends the SNZ approach, to focus on linked administrative data that is longitudinal in nature. Our framework is valuable for making design decisions where trade-offs between different types of error could be finely balanced. Shrewd judgement requires an understanding of the error dynamics at play in the data.
23. Some of the errors to be considered cannot be easily quantified. In these cases it is important that they are conceptualised so that the scale and implications of the errors can be borne in mind and managed. Quality indicators seek to quantify data errors: scaling factors that benchmark the data against other sources. Even if both sources have errors, the relationship between them over time is illuminating, shining a light on both variance and potential bias. In other work, logged scaling factors have been modelled, as a means to attempt to measure statistical uncertainty (see ONS 2017). Qualitative investigation can produce estimates of false positive linkage. Likewise, false negatives can usefully be quantified. The age and sex distributions of linkage residuals provide some insight on representativeness of our linked datasets.
24. *We welcome* comments *on our journey so far.*

References

2. Groves, R. M., F. J. Fowler Jr., M. Couper, J. M. Lepkowski, E. Singer and R. Tourrangeau (2004), Survey methodology, Wiley, New York.

3. Home Office (2019), Fourth report on statistics being collected under the exit checks programme, available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/826381/fourth-report-on-statistics-being-collected-under-the-exit-checks.pdf

4. Lynn, P. (2009), Methods for Longitudinal Surveys, in Lynn, P. (ed) Methodology of Longitudinal Surveys, Wiley, Chichester.

5. Lynn, P. (2001), A Quality Framework for Longitudinal Studies, available at: https://www.iser.essex.ac.uk/files/ulsc/standards/framework/2001-09-26.pdf

6. Lynn, P. and Lugtog, P. (2016) Total Survey Error for Longitudinal Surveys, Understanding Society Working Paper Series 2016-07, September 2016, available at: https://www.understandingsociety.ac.uk/sites/default/files/downloads/working-papers/2016-07.pdf

7. Office for National Statistics (2019), Developing our approach for producing admin-based population estimates, England and Wales: 2011 and 2016, available at:

8. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/developingourapproachforproducingadminbasedpopulationestimatesenglandandwales2011and2016/2019-06-21

9. Office for National Statistics (2017), Methods for measuring statistical uncertainty in local authority mid-year estimates, Newport, Wales, available at: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/methodologyformeasuringuncertaintyinonslocalauthoritymidyearpopulationestimates2012to2015

10. Statistics New Zealand, Guide to reporting on administrative data errors, see: http://archive.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality/sources-of-error.aspx#

11. Tourangeau, R. (2018) How Errors Cumulate: Two Examples, Hansen Lecture, Washington DC, October 11, 2018, found at: http://washstat.org/hansen/2018Tourangeau.pdf

12. Zhang, L. (2012), Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica, 66: 41-63. doi:10.1111/j.1467-9574.2011.00508.x

Appendix

Appendix 1. Errors in the single-source framework

Single source errors	
Objects[1]	**Attributes**[2]
Frame error: *Coverage error* Assessing objects that are not in the target data, or not being able to access objects that are in the target data. *Timing differences* Objects in the ideal target data that are not accessible because of a discrepancy in the time window for obtaining observations.	*Relevance error:* *Validity error* The difference between ideal measurement of attributes sought about an object and the operational measure used to collect it. *Timing differences* A conceptual discrepancy in the timing of the measurement of attributes between the ideal target data and accessible data.
Selection error: *Duplicated objects* Objects that are represented more than once in the accessed data. *Missing objects* Objects that *in theory* are accessible but are not in the accessed data.	*Measurement error:* *Measurement error* Errors arising from attributes that are not recorded accurately. *Missing attributes* Attributes that are missing from the accessed data (could be for specific objects or all of the objects).
Processing error: *Deduplication error* Errors arising from deduplication of objects in the accessed dataset. This could include both deduplicating objects that are actually different (false positive error) or failing to deduplicate objects that are the same (false negative error). *Imputation error* Errors arising from the imputation of missing objects. *Filtering error* Errors arising from the selection or de-selection of accessed objects to an ideal target set.	*Processing error:* *Editing error* Errors arising from editing the value of an attribute. This could include editing as a result of validation or QA checks. *Imputation error* Errors arising from the imputation of missing attribute values. *Classification error:* Errors arising from classification of values into groups or derivation of new attributes. *Harmonisation error:* Errors arising from the harmonisation of values of attributes to an ideal or target concept

Attrition The loss of research objects or units over time. Occurs naturally, through death (or an unobserved migration). Also occurs through failure of follow-up, a refusal to take part, in the case of survey data, or through missing information or linkage failure, in administrative sources.	*Censoring* Where the value of a measurement or observation is only partially known. Right censoring is when the research object drops out of the data before the end of the observation window or does not experience the event of interest during the observation window. Left censoring is when the event of interest has already occurred, before the observation window begins.
Periodicity/seasonality error Objects are not observed because the data capture is not frequent enough (periodicity) nor adequate to capture seasonality in the data (seasonality).	*Periodicity/seasonality error* Measurement of attributes over time are not frequent enough (periodicity) nor adequate to capture seasonality in the data (seasonality).

[1] *This refers to data units and could be events, transactions, persons, households, firms or other entries in an admin dataset.*
[2] *This refers to the measures or variables that have been collected that relate to the data objects/ units*.

Appendix 2. Longitudinal error

Longitudinal error – applies to both single source events and multisource longitudinal datasets	
Objects[1]	**Attributes**[2]
Attrition The loss of research objects or units over time. Occurs naturally, through death (or an unobserved migration). Also occurs through failure of follow-up, a refusal to take part, in the case of survey data, or through missing information or linkage failure, in administrative sources.	*Censoring* Where the value of a measurement or observation is only partially known. Right censoring is when the research object drops out of the data before the end of the observation window or does not experience the event of interest during the observation window. Left censoring is when the event of interest has already occurred, before the observation window begins.
Periodicity/seasonality error Objects are not observed because the data capture is not frequent enough (periodicity) nor adequate to capture seasonality in the data (seasonality).	*Periodicity/seasonality error* Measurement of attributes over time are not frequent enough (periodicity) nor adequate to capture seasonality in the data (seasonality).

[1]*This refers to data units and could be events, transactions, persons, households, firms or other entries in an admin dataset.*
[2]*This refers to the measures or variables that have been collected that relate to the data objects/ units*.

Appendix 3. Errors in the multiple source framework

Multiple source errors	
Objects[1]	**Attributes**[2]
Frame error: *Coverage error* Observing objects that are not in the target linked data, or not being able to access objects that are in the target linked data. *Timing differences* Objects are not observed due to conceptual discrepancies in the timing of the capture between the target linked data and source data.	*Relevance error:* *Relevance error* The differences between ideal measurement of attributes sought about an object and the operational measures used to collect it in each source dataset. *Timing differences* A conceptual discrepancy in the timing of the measurement of attributes between the target linked data and the source data.
Coverage error: *Coverage error* Objects are not linked due to discrepancies in the coverage of objects between data sources. *Timing differences* The difference between observed objects in source datasets due to the data being captured at different times. *Linkage error* Errors arising from linking objects together incorrectly (false positive error) and failing to link objects together that should have been linked (false negative error).	***Mapping error:*** *Definitional differences* The differences between how attributes are operationally measured in each of the source datasets. *Timing differences* The differences between the values of attributes for a linked object between source datasets caused by the data being captured at different times.
Identification error: *Linkage selection error (bias)* Errors arising from the selection of linked objects (or de-selection of unlinked objects) due to biases in the linkage, or through error in the resolution of conflicting links.	*Comparability error:* *Alignment error* Errors arising from the alignment of the conflicting values of attributes across sources.

Processing error: *Imputation error* Errors arising from the imputation of missing objects. *Filtering error* Errors arising from the selection or de-selection of accessed objects to an ideal target set.	*Processing error:* *Editing error* Errors arising from editing the value of an attribute. This could include editing as a result of validation or QA checks. *Imputation error* Errors arising from the imputation of missing attribute values. *Classification error:* Errors arising from classification of values into groups or derivation of new attributes. *Harmonisation error:* Errors arising from the harmonisation of values of attributes to an ideal or target concept.
Attrition The loss of research objects or units over time. Occurs naturally, through death (or an unobserved migration). Also occurs through failure of follow-up, a refusal to take part, in the case of survey data, or through missing information or linkage failure, in administrative sources.	*Censoring* Where the value of a measurement or observation is only partially known. Right censoring is when the research object drops out of the data before the end of the observation window or does not experience the event of interest during the observation window. Left censoring is when the event of interest has already occurred, before the observation window begins.
Periodicity/seasonality error Objects are not observed because the data capture is not frequent enough (periodicity) nor adequate to capture seasonality in the data (seasonality).	*Periodicity/seasonality error* Measurement of attributes over time are not frequent enough (periodicity) nor adequate to capture seasonality in the data (seasonality).

[1]*This refers to data units and could be events, transactions, persons, households, firms or other entries in an admin dataset.*
[2]*This refers to the measures or variables that have been collected that relate to the data objects/ units.*