

ANZSCO imputation in the National Apprentice and Trainee Collection

BRIAN HARVEY
NATIONAL CENTRE FOR
VOCATIONAL EDUCATION RESEARCH

The views and opinions expressed in this document are those of the author and do not necessarily reflect the views of the Australian Government or state and territory governments.

© National Centre for Vocational Education Research, 2010

This work has been produced by the National Centre for Vocational Education Research (NCVER).

Apart from any use permitted under the *Copyright Act 1968*, no part of this publication may be reproduced by any process without written permission. Requests should be made to NCVER.

The views and opinions expressed in this document are those of the author and do not necessarily reflect the views of NCVER.

TD/TNC 100.14

Published by NCVER

ABN 87 007 967 311

Level 11, 33 King William Street, Adelaide, SA 5000

PO Box 8288 Station Arcade, Adelaide SA 5000, Australia

ph +61 8 8230 8400 fax +61 8 8212 3436

email ncver@ncver.edu.au

<<http://www.ncver.edu.au>>

<<http://www.ncver.edu.au/publications/2266.html>>

Contents

Introduction	4
Missing occupation codes	4
The need to impute	4
Caveat	4
Requirements	5
Consistency	5
Method (overview)	6
Implementation	8
Concluding remarks	9
Appendices	
1 Method (detailed)	10
2 System charts	13

Introduction

Missing occupation codes

Data relating to occupations have been collected in the National Apprentice and Trainee Collection since 1994. The coding used conforms to classifications endorsed by the Australian Bureau of Statistics (ABS). The latest version issued from the ABS is the Australian and New Zealand Standard Classification of Occupations (ANZSCO). The classification previous to ANZSCO is the second edition of the Australian Standard Classification of Occupations (ASCO). Currently, occupation data are collected using both ANZSCO and ASCO codes.

Although ANZSCO was only introduced to the Apprentice and Trainee Collection in 2007, states and territories were asked to back-code ANZSCO on all contracts that were active as at 1 July 2000. Therefore, while ASCO codes appear on records from the beginning of the collection, ANZSCO codes only appear on records from 1 July 2000. Thus, although the database has a field for ANZSCO, the value is missing for all records prior to this date.

The need to impute

An historical series of data classified by ANZSCO can only go as far back as 1 July 2000; however, there is a demand to produce series that go further back in time. To satisfy this demand, the missing ANZSCO codes need to be either assigned (a manual process), deduced or inferred.

Manually coding ANZSCO for the older records is not viable due to the extent of resources that would be required.

If ANZSCO codes had a one-to-one correspondence with ASCO codes, then it would be easy to calculate ANZSCO codes for the older records. If this were the case, then there would be no need to invoke a (complex) imputation procedure. However, there isn't a simple one-to-one mapping between the ASCO and ANZSCO categories.

The remaining option is then to implement an imputation procedure to assign ANZSCO codes to the older data.

Caveat

It is important to remember that imputation is associated with some degree of uncertainty. The assignment to individual records might not be correct even if broad-level aggregates of the data have the correct distributional properties. In general, the finer the level of aggregation, the more risk is involved in using imputed ANZSCO codes. Unfortunately, quantifying the level of risk is not a straightforward task.

Requirements

Consistency

The most important requirement was that the imputed ANZSCO coding had to be as consistent as possible with the way the existing ANZSCO had been coded. For example, if in a particular state, a given qualification has always been coded to a single ANZSCO category, then that category should be imputed whenever records for that state and qualification are missing a value for ANZSCO.

The basic idea underlying the process is to match a record that is missing an ANZSCO category with the most similar set of records among those with an ANZSCO category and assign the value held by those records.

When imputing an ANZSCO for a record, it sometimes occurs that the most similar set of records with ANZSCO codes do not have a unique value. In this case, it is necessary to select one of the values with a probability proportional to the frequency with which the values occur in the collection. This is a clear case of where the imputed value might be wrong for a given record; however, the distribution of imputed values should reflect the distribution of known values.

Method (overview)

This section gives a broad-level description of the imputation process. The steps are executed in the sequence described below. At any point in the process, records that have not yet had an ANZSCO code imputed are passed on to the next step.

More detailed descriptions can be found in the appendices.

1 Prepare data

Data from the apprentice and trainee database are read into SAS datasets, keeping only those variables required for the imputation process. Qualification, state, Australian Qualifications Framework (AQF) level and ASCO are retained as being particularly associated with ANZSCO.

For some contracts, every record will have an ANZSCO code. These records can be used to provide data for the imputation process.

Other contracts will contain both records with ANZSCO codes and some without codes. The records without an ANZSCO code are imputed to have the same ANZSCO code as the other records in the contract. These records can also be used to provide data for the imputation process.

The remaining contracts have no ANZSCO codes for any records. These records will need to have an ANZSCO code imputed.

2 Create imputation tables

Group the records that already have ANZSCO values by qualification, state, AQF level and ASCO and ANZSCO categories. Tables that associate ANZSCO codes with all the combinations of the other variables are created from this process. As a result it is possible to identify which combinations of qualification, state, AQF level and ASCO are associated with which ANZSCO codes.

3 Match records with missing ANZSCO to the imputation tables

This is just a matching exercise. Records with missing ANZSCO codes are compared with the tables created in the previous step. If the combination of qualification, state, AQF level and ASCO match, then the ANZSCO code in the imputation table is used as the imputed value. If the match corresponds to multiple ANZSCO codes, then as previously mentioned, a random selection is made (with probability proportional to frequencies).

Unmatched records will move on to the next step.

4 Use ABS ASCO to ANZSCO concordance for imputation

The ABS provides information on the relationship between the older ASCO and the newer ANZSCO classifications. In particular, there is a description of which ASCO codes correspond to which ANZSCO codes.

From this concordance, use only the ASCO codes that map to a single ANZSCO. Whenever a record with missing ANZSCO categories has been coded with one of these ASCO codes, then use the mapped ANZSCO for imputation.

Unmatched records will move on to the next step.

5 Map ASCO to ANZSCO based on distribution in data

Where the ABS maps an ASCO code to more than one ANZSCO code, the apprenticeship collection can give information on how to impute a single ANZSCO code. It is possible that an ASCO code has only ever been associated with one ANZSCO code in the collection, even though the ABS concordance assigns that ASCO code to many ANZSCO codes. When this is the case, the imputed value is the associated code.

When the collection associates an ASCO code with many ANZSCO codes, then, as before, a random selection is made (with probability proportional to frequencies).

Unmatched records will move on to the next step.

6 Manually code remaining ANZSCOs

At this stage any records still missing an ANZSCO code have exhausted the process. Fortunately, the number of remaining records was small. These records were given an ANZSCO code by manually assigning a value based on the qualifications associated with those records.

7 Finalise imputation table

Append the data set of imputed ANZSCO codes from step 1 to the data set of matched records subsequently accumulated.

Implementation

The method described above was translated into a SAS programme. The apprentice and trainee database current at the end of December 2008 (collection 58) was used as the source data.

The small number of records that were assigned ANZSCO codes manually were associated with just two qualifications. The ANZSCO codes for these qualifications were hard-coded into the program as the last procedure of the imputation processes.

Concluding remarks

The fact that an ANZSCO code could be assigned to every record does not, by itself, prove the quality of the imputations nor validate the method used. As previously mentioned, the ANZSCO code assigned to any given record is not necessarily correct. The goal is to have aggregates of the data that exhibit the correct distribution of ANZSCO codes. As a general principle, the imputed values should 'perform' better with data that are aggregated at broader levels than finer levels

Furthermore, the records without ANZSCO codes span a time period from 1994 to 2000. The variables collected have changed over time (early records do not even have qualifications included) and include the introduction and uptake of training packages phased in over time. It can reasonably be expected that the quality of the imputation is better for the records closer in time to 2000, progressively deteriorating for records from earlier years.

Appendix 1: Method (detailed)

1 Prepare data

- 1.1 Create an extract from the apprentice and trainee database.
- 1.2 Sort the records so that contracts are in time order.
- 1.3 Identify contracts with at least one record missing ANZSCO.
- 1.4 Split records into three data sets:
 - 1.4.1 **HAVES**. Records with ANZSCO present.
 - 1.4.2 **IMPUTED**. Records with ANZSCO missing but present in other records belonging to the same contract: impute the ANZSCO from those records.
 - 1.4.3 **HAVE_NOTS**. Records with ANZSCO missing for all records belonging to the same contract.
- 1.5 Append a copy of **IMPUTED** to **HAVES**.

2 Create imputation tables

- 2.1 Sort **HAVES** by qualification, state, AQF, ASCO and ANZSCO.
- 2.2 Compress **HAVES** to one record per 'by group', keeping count of how many records contribute to each (keep as variable 'numerator'). Save as **TEMP**.
- 2.3 Split **TEMP** into two data sets based on the number of ANZSCO categories per qualification, state, AQF and ASCO groups. Create data set **ASCO_UNIQUE** to store records where there is only one ANZSCO per group. Create data set **NON_UNIQUE** to store the remainder.
- 2.4 Sort **ASCO_UNIQUE** by qualification, state, AQF and ANZSCO and store as **AQF_UNIQUE**.
 - 2.4.1 Compress **AQF_UNIQUE** to one record per 'by group'.
 - 2.4.2 Retain only those records that are one record per qualification, state and AQF group.
- 2.5 Sort **AQF_UNIQUE** by qualification, state and ANZSCO and store as **STATE_UNIQUE**.
 - 2.5.1 Compress **STATE_UNIQUE** to one record per 'by group'.
 - 2.5.2 Retain only those records that are one record per qualification and state group.
- 2.6 Sort **STATE_UNIQUE** by qualification and ANZSCO and store as **QUAL_UNIQUE**.
 - 2.6.1 Compress **QUAL_UNIQUE** to one record per 'by group'.
 - 2.6.2 Retain only those records that are one record per qualification group.

- 2.7 With the data set **NON_UNIQUE** (created in 2.3), do the following:
 - 2.7.1 Accumulate ‘numerator’ (see 2.2) for each qualification, state, AQF, ASCO group (keep as variable ‘denominator’). Store as data set **DIVISORS**.
 - 2.7.2 Merge **NON_UNIQUE** and **DIVISORS**. For each ANZSCO category associated with a qualification, state, AQF, ASCO group, calculate variable ‘imputation weight’ from ‘numerator’ and ‘denominator’. Store in data set **NON_UNIQUE**.
 - 2.7.3 Calculate ‘cumulative_weight’ by accumulating ‘imputation_weight’ within qualification, state, AQF, ASCO groups.
 - 2.7.4 Restructure **NON_UNIQUE** so that each qualification, state, AQF, ASCO group has one record which contains details of all associated ANZSCO categories.

3 Match records with missing ANZSCO to the imputation tables

- 3.1. Sort **HAVE_NOTS** by qualification, state, AQF and ASCO.
- 3.2. Match **HAVE_NOTS** against the imputation table **ASCO_UNIQUE**. Store matched records in **MATCHED** and let the imputed ANZSCO be the corresponding value from the imputation table. Store unmatched records as **NOT_MATCHED**.
- 3.3. Match **NOT_MATCHED** against the imputation table **AQF_UNIQUE**. Store matched records in temporary data set **MATCHED2** and let the imputed ANZSCO be the corresponding value from the imputation table.
 - 3.3.1. Unmatched records remain in **NOT_MATCHED**. Append **MATCHED2** to **MATCHED**.
- 3.4. Match **NOT_MATCHED** against the imputation table **STATE_UNIQUE**. Store matched records in temporary data set **MATCHED2** and let the imputed ANZSCO be the corresponding value from the imputation table. Unmatched records remain in **NOT_MATCHED**.
 - 3.4.1. Append **MATCHED2** to **MATCHED**.
- 3.5. Match **NOT_MATCHED** against the imputation table **QUAL_UNIQUE**. Store matched records in temporary data set **MATCHED2** and let the imputed ANZSCO be the corresponding value from the imputation table. Unmatched records remain in **NOT_MATCHED**.
 - 3.5.1. Append **MATCHED2** to **MATCHED**.
- 3.6. Match **NOT_MATCHED** against the imputation table **NON_UNIQUE**. Store matched records in temporary data set **MATCHED2** and let the imputed ANZSCO be the value selected (randomly) from the imputation table. Unmatched records remain in **NOT_MATCHED**.
 - 3.6.1. Append **MATCHED2** to **MATCHED**.

4 Use ABS to ANZSCO concordance for imputation

- 4.1. Import concordance into SAS from Excel.

- 4.2. Sort concordance by ASCO.
- 4.3. Where an ASCO code maps to a single **ANZSCO** code, store that record in the data set **UNIQUE**. Otherwise store in data set **MULTIPLE** (currently not used in subsequent processing).
- 4.4. Sort **NOT_MATCHED** (from 3.6) by ASCO.
- 4.5. Match **NOT_MATCHED** against the imputation table **UNIQUE**. Store matched records in temporary data set **MATCHED2** and let the imputed ANZSCO be the corresponding value from the imputation table. Unmatched records remain in **NOT_MATCHED**.
 - 4.5.1. Append **MATCHED2** to **MATCHED**.

5 Map ASCO to ANZSCO based on distribution in data

- 5.1. Append **MATCHED** to **HAVES** and store as data set **ASCO_MAP**.
 - 5.1.1. Sort **ASCO_MAP** by ASCO and ANZSCO.
 - 5.1.2. Compress **ASCO_MAP** to one record per ‘by group’, keeping count of how many records contribute to each (keep as variable ‘weight’).
 - 5.1.3. For ASCOs that associate with only one ANZSCO, store that information in data set **UNIQUE**. Otherwise, store in **NON_UNIQUE**.
- 5.2. Match **NOT_MATCHED** to **UNIQUE** and store as data set **ASCO_MAP**. Store matched records in temporary data set **MATCHED2** and let the imputed ANZSCO be the corresponding value from **UNIQUE**. Unmatched records remain in **NOT_MATCHED**.
 - 5.2.1. Append **MATCHED2** to **MATCHED**.
- 5.3. Accumulate ‘weight’ on data set **NON_UNIQUE** by ASCO group (keep as variable ‘denominator’). Store in **ASCO_TOTALS**.
- 5.4. Merge **NON_UNIQUE** and **ASCO_TOTALS** and recalculate weight as a proportion.
 - 5.4.1. Delete any records with missing ASCO values.
 - 5.4.2. Restructure **NON_UNIQUE** so that there is one record for every ASCO and calculate cumulative weights.
- 5.5. Match **NON_UNIQUE** and **NOT_MATCHED**. Store matched records in temporary data set **MATCHED2** and let the imputed ANZSCO be the value selected from **NON_UNIQUE**. Unmatched records remain in **NOT_MATCHED**.
 - 5.5.1. Append **MATCHED2** to **MATCHED**.

6 Manually code remaining ANZSCOs

- 6.1. Hard-code the values for the remaining qualifications on **NOT_MATCHED**.
- 6.2. Append **NOT_MATCHED** to **MATCHED**.

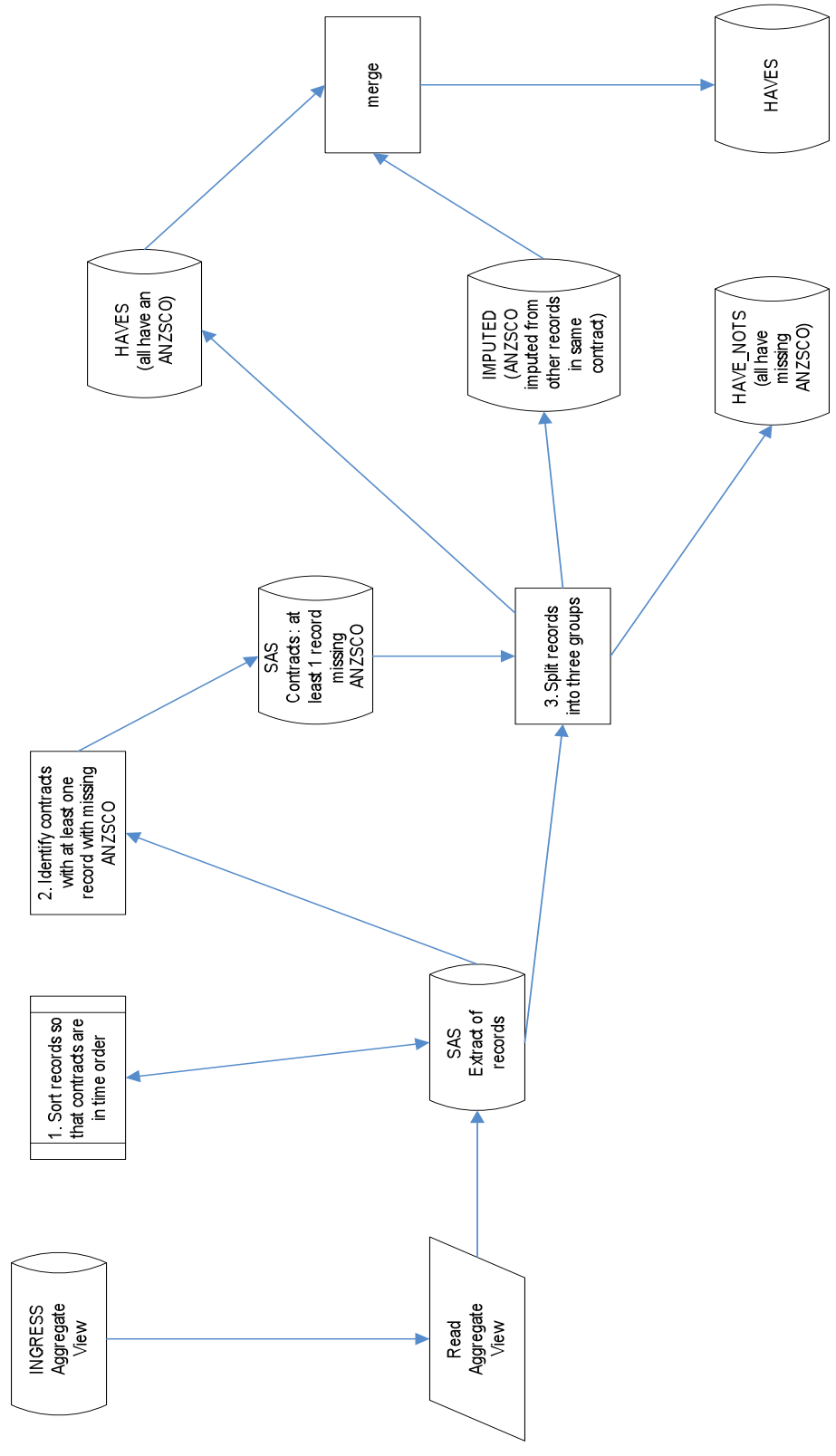
7 Finalise imputation table

- 7.1. Append **IMPUTED** (from 1.4.2) to **MATCHED**.

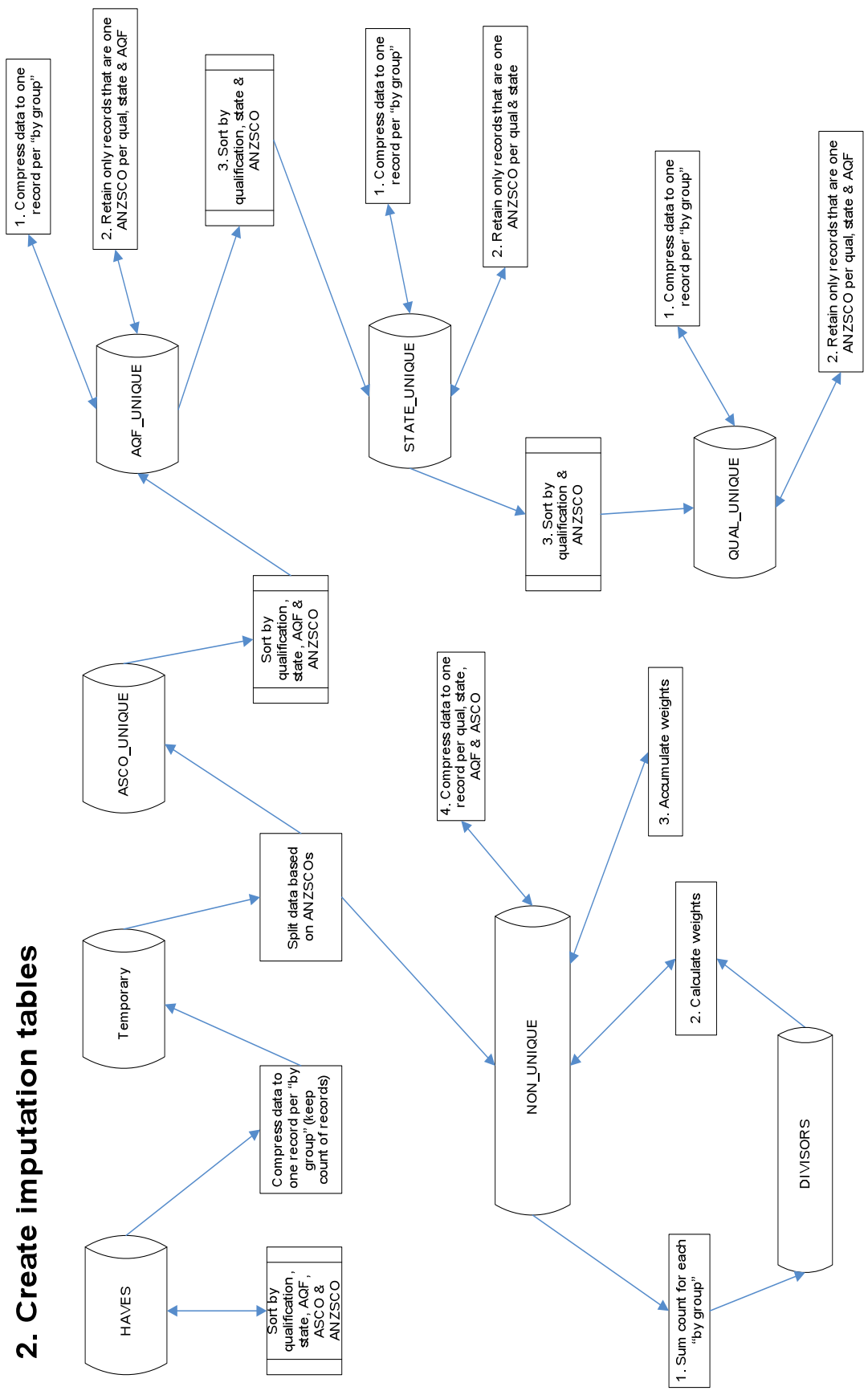
Appendix 2: System charts

This appendix gives a pictorial representation of the imputation process described in appendix 1. Because of the detailed nature of the process the diagram has been split into several parts.

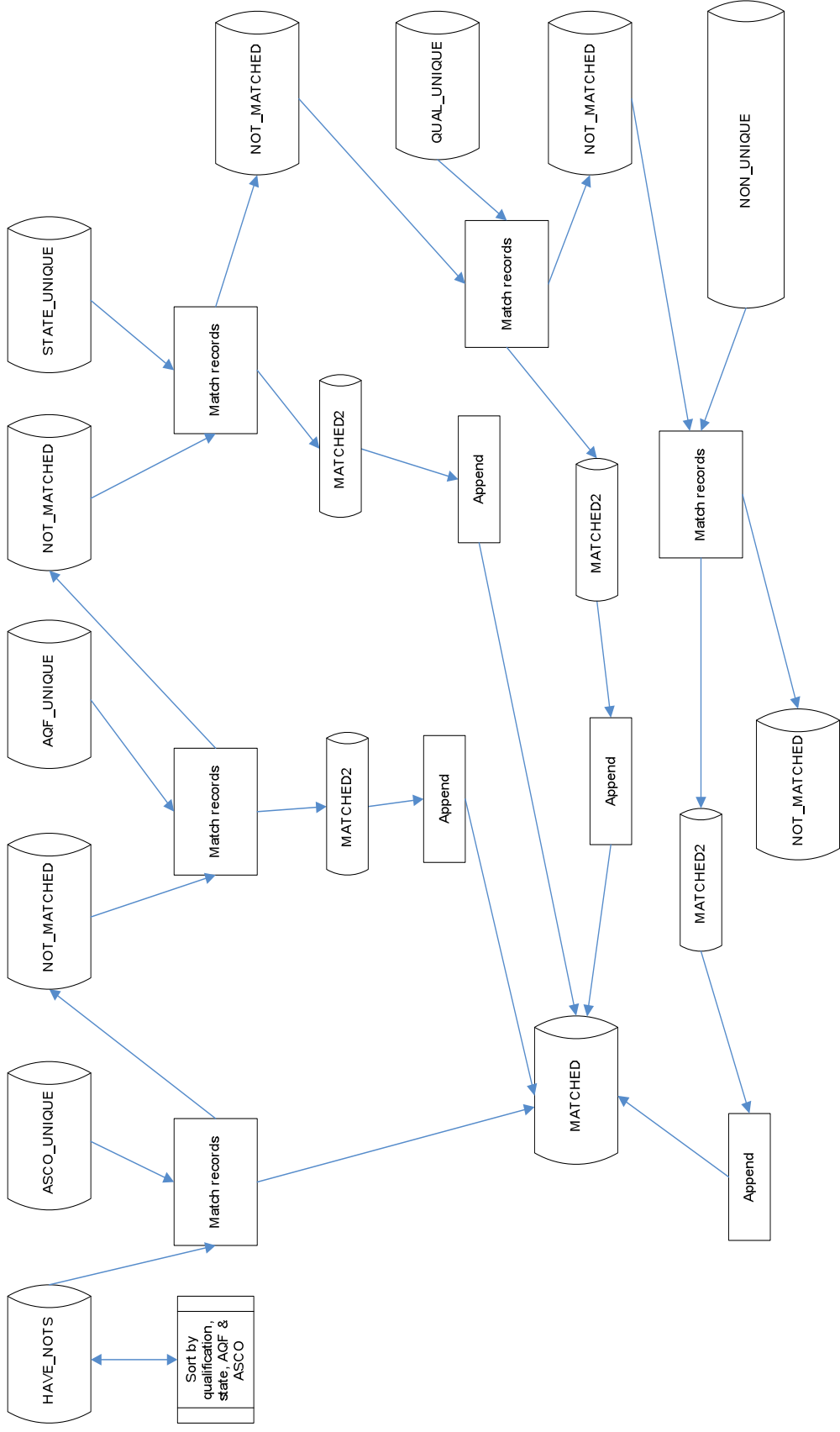
1. Prepare data



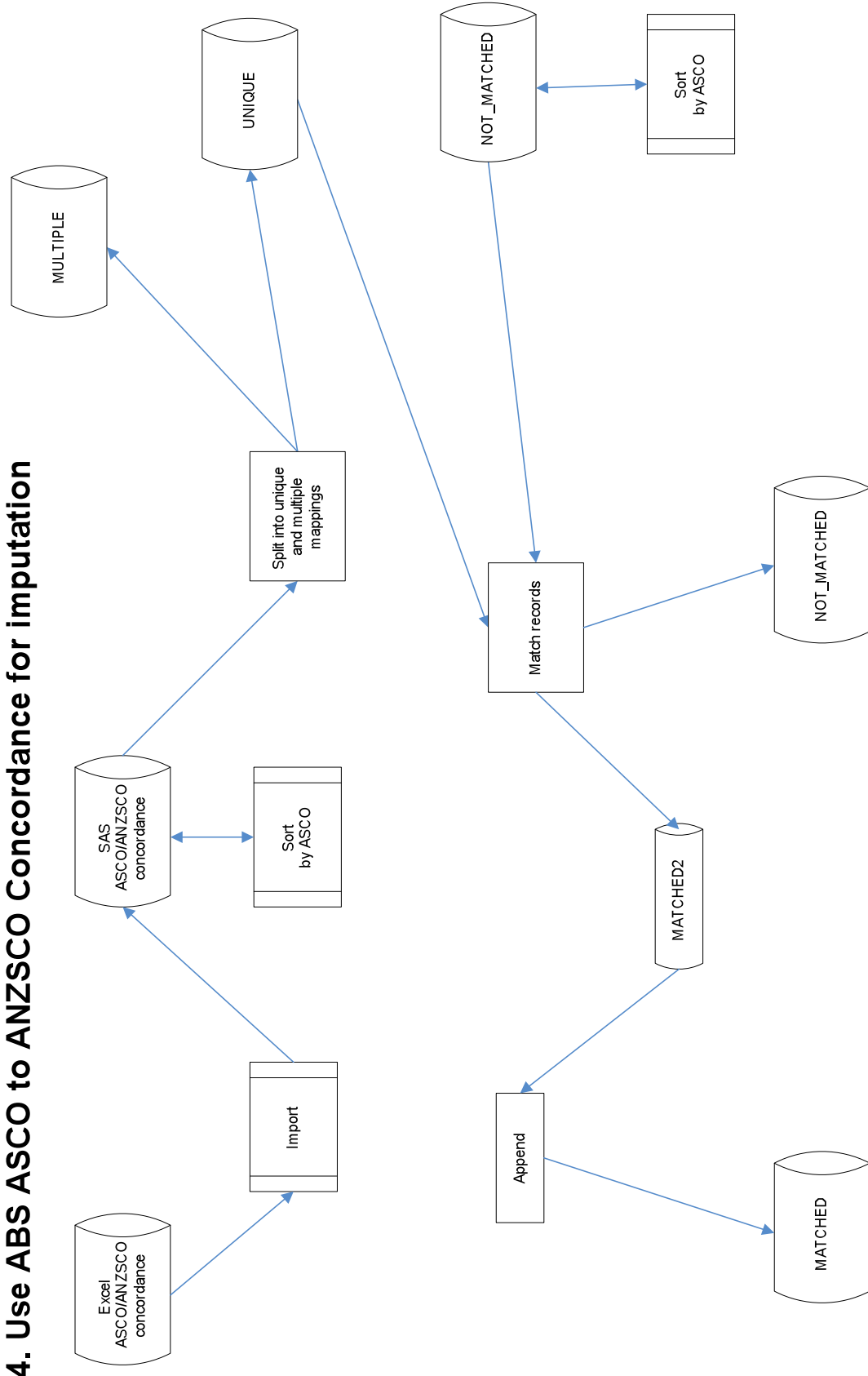
2. Create imputation tables



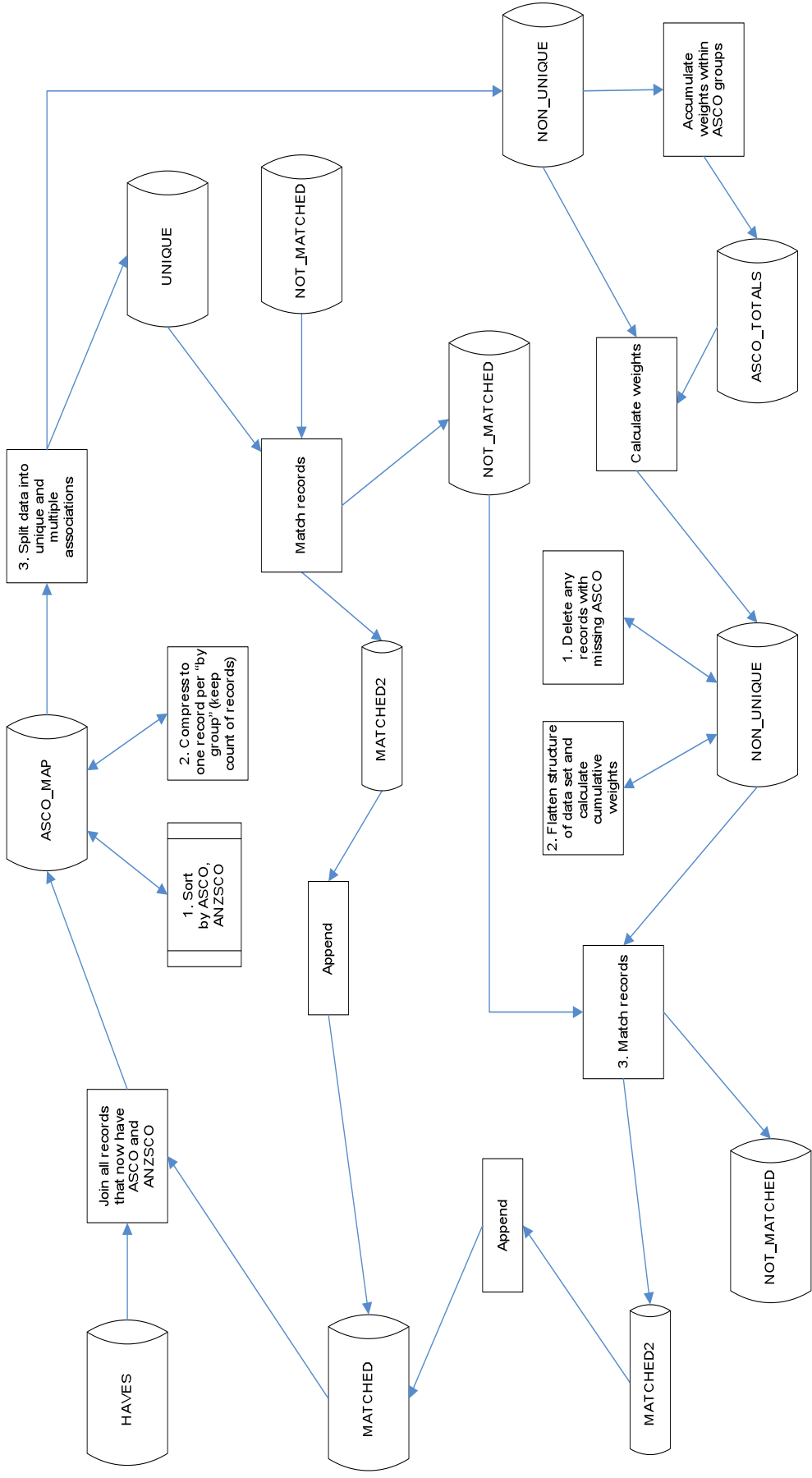
3. Match records with missing ANZSCOs to the imputation tables



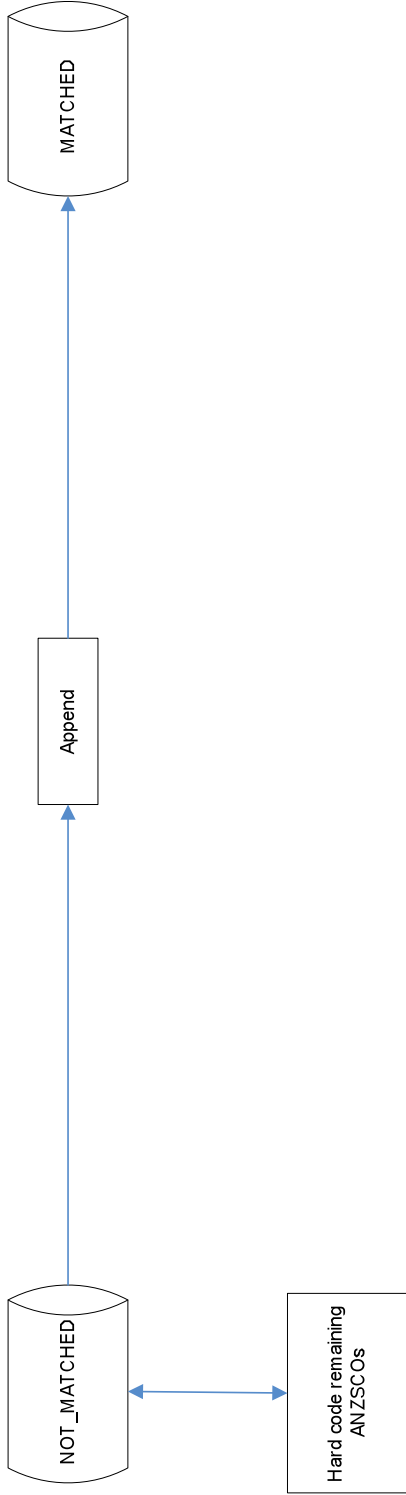
4. Use ABS ASCO to ANZSCO Concordance for imputation



5. Map ASCO to ANZSCO based on distribution in data so far



6. Manually code remaining ANZSCOs



7. Finalise imputation table

