MACHINE READABLE (CSV FORMAT) CORRESPONDENCE DOCUMENTATION

Introduction

Since the first release of the Australian Statistical Geography Standard (ASGS) in 2011, the ABS has created and provided a wide range of both area weighted and population weighted geographical correspondences. These files are produced primarily to assist users to make comparisons and maintain time series between the editions of the ASGS, as well as with the previously used statistical geography, the Australian Standard Geographical Classification (ASGC).

Area weighted correspondences are created through a simple geographic overlay process. They can be distinguished from population weighted files by the second character of their file names, which will be an “A” for Area in the case of an area based file (the naming convention is further described in the Correspondence File Name section).

Population weighted correspondences on the other hand have been created using specific population weighted grids which represent either Census Collection Districts (CD), or Mesh Blocks (MB). This creates a far more accurate correspondences than have been previously available. Population weighted correspondences will have a “G” for Grid as the second character in their file name. The methodology used for population weighted files is further described below.

Traditionally these correspondences have been provided in Microsoft Excel format with different components of the correspondence stored across various sheets within the Excel file. In some cases this has required users to manually combine all the correspondence records prior to converting data from one geographic boundary set to another, which can be quite a time consuming task. To circumvent this, the ABS is now providing correspondences in CSV format which will contain all the information in the one location. CSV format will still automatically open in Microsoft Excel, but the other purpose of using CSV format is that it is “machine readable”. This means that it can be loaded into other software and the data can then be converted using an automated process rather than having to be manually handled.

All the information currently provided in correspondences, such as the Individual Quality Indicator and Overall Quality Indicator, as well as records which have no corresponding "FROM" or "TO" region, known as "Null" records, will still be provided. This will also include records that have a minor interaction with another record, known as "Below Minimum Output Size" records.

This document details the methodology used to calculate the ratio that a particular "FROM" region will donate to a "TO" region, as well as how both the Individual and Overall Quality Indicators are derived. Information on the file naming convention used with the new format, and descriptions of the column headings and their meanings will also be provided. It should be noted that the methodology used to create the correspondences remains unchanged, and the aim of this document, while providing details on the methodology, is to describe the new output format.

Population Weighted Grid Correspondences

The population weighted grid method that the ABS has adopted is essentially a series of grid points that represent the underlying geographical distribution of the weighting unit, most often CD or Mesh Block population. Each grid point is then assigned a value based on this weighting. This is demonstrated in the example below.

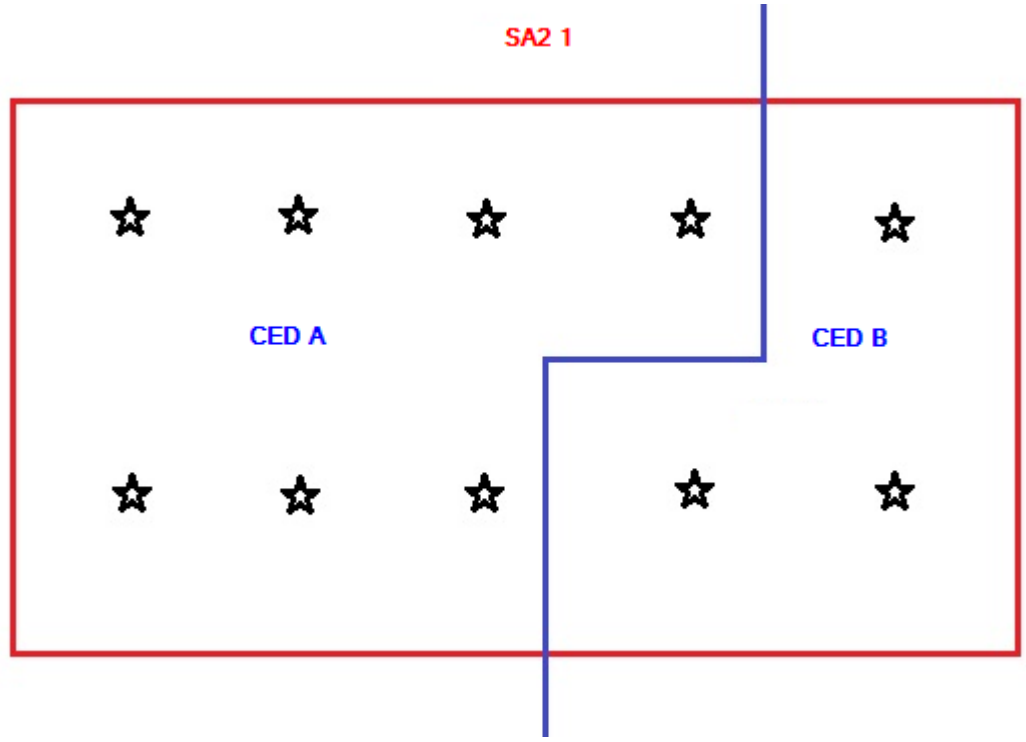

Figure 1: A correspondence example from a Statistical Area Level 2 (SA2) with population grid points and Commonwealth Electoral Division (CED) regions.

The correspondence in Figure 1 is from a SA2 (the "FROM" region) to a CED (the "TO" region), 2016 Mesh Block population is the weighting. This hypothetical SA2 contains 200 persons, represented by ten evenly distributed grid points, each grid point representing 20

persons.

The next step in the correspondence generation process is to determine the proportion that the SA2, as the “FROM” region, is donating to the respective CED “TO” regions. In Figure 1 there are seven grid points in CED A, and three in CED B. Given that each grid point represents 20 persons, 140 persons are located in CED A and 60 in CED B. The ratio is then calculated by dividing the population found in each of the “TO” regions by the total population of the “FROM” region. Therefore the ratios are as follows:

- CED A: 140 / 200 which gives a ratio of 0.7.
- CED B: 60 / 200 which gives a ratio of 0.3.

The result is that the SA2 in question is donating 70 per cent of its data to CED A, and 30 per cent of its data to CED B.

The benefit of using this method is that any two sets of geographic regions can have a correspondence generated for them, and that any attribute value can be distributed across the grid to be used as the weighting unit.

Quality Indicator

With the increased demand for correspondences, it was determined that a measure of quality should be derived to provide users with an indication of how well data will be converted to each respective “TO” region. This will inform users of where the converted data values are likely to be accurate, and where caution will need to be used when assessing the results.

The method that has been developed to generate the quality indicator involves a number of steps. Firstly it looks at the value that a “FROM” region donates to a “TO” region as a ratio of the whole “FROM” region. The next step is to examine the value that the “FROM” region donates to the “TO” region as a ratio of the whole “TO” region. These two values are then multiplied together to provide the component for that “FROM” region. This process is then repeated for each donating “FROM” region, with the component values then added to provide the overall score for the “TO” region. Based on the score returned, a textual description is then applied as to how well the ABS expects data to be converted to the “TO” region. This is highlighted in the example below.

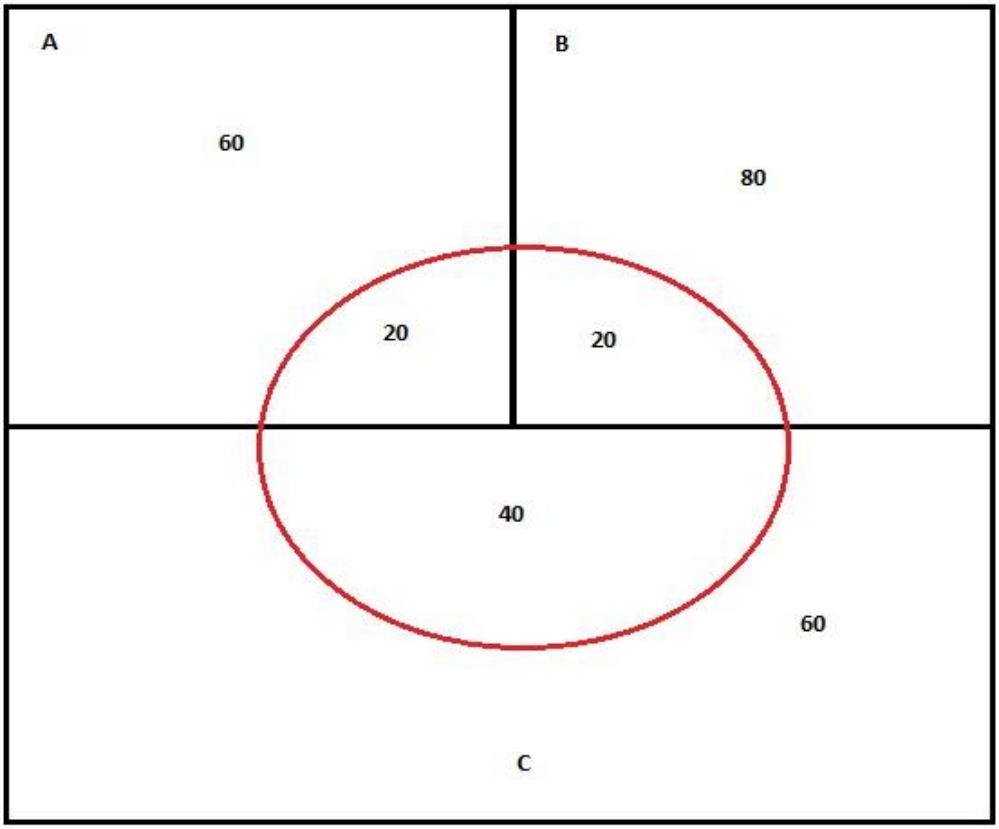

Figure 2: Illustration of 3 “FROM” regions to 1 “TO” region.

In Figure 2 there are three “FROM” regions represented by the black boundaries. The “TO” region is represented by the red ellipse.

Region A donates 20 persons to the “TO” region, while there are a further 60 that are not donated to the “TO” region. Therefore the ratio of “FROM” Region A is 20 / 80, or 0.25. The next step is to look at the value that is being donated from Region A compared to the total value of the “TO” region. Region A donates 20 persons, and the total population is 80. So in this case the ratio is 20 / 80, or 0.25. Region A's component score is then calculated by multiplying 0.25 x 0.25 giving Region A a component score of 0.0625.

The same process is then applied to “FROM” Regions B and C. Region B donates 20 persons with a further 80 persons in the remainder of the “FROM” region. Therefore its ratio is 20 / 100 or 0.2. Region B donates 20 persons and the total population of the “TO” region is 80 so the ratio is 20 / 80 or 0.25. Region B's component score is therefore 0.2 x 0.25 or 0.05.

Similarly Region C donates 40 persons with another 60 in the remainder of “FROM” region. The ratio is 40 / 100 or 0.4. The 40 persons donated are then compared against the total population of the “TO” region of 80, so the ratio is 40 / 80 or 0.5. This results in the component score for “FROM” Region C being 0.4 x 0.5 or 0.2.

The final step is to add the three component scores. In this case:

- Region A = 0.0625

- Region B = 0.05
- Region C = 0.2

The final result is that the “TO” region in this example would have a quality indicator score of 0.3125, a score that the ABS would regard as being poor, meaning that caution would have to be used when using the results of data converted to the “TO” region.

Overall Quality Indicator

An overall quality indicator is given to each correspondence. The aim of this is to provide users with a reasonable idea of how well the correspondence will convert data across the whole of the correspondence.

The overall quality indicator is derived from multiplying the population of each “TO” region with its quality indicator score, based on the methodology described above. The values produced by this multiplication for each “TO” region are then added together. This aggregated value is then divided by the total population of the “TO” regions. This will return a result similar to the individual quality indicator scores. Similar textual descriptions are then applied.

Correspondence File Name

The file naming convention remains predominantly the same. The first character of the file name will always be “C” for Correspondence. The second character can be one of two letters and indicates how the file is weighted:

A = weighted by Area
G = weighted by a standard population Grid

The example below relates to a correspondence where 2016 SA2s are being corresponded to 2017 CEDs using a standard population grid.

File name:
Statistical Area Level 2 2016 TO Commonwealth Electoral Division 2017

and

CG_SA2_2016_CED_2017 - All.csv

Table 1: Character and meaning of the file name.

Character	**Meaning**
C	Correspondence
G	In this case ‘G’ indicates population Grid based correspondence

SA2	Represents the name of the “FROM” region, in this case SA2s
2016	The year that this version of the “FROM” region was released
CED	Represents the name of the “TO” region, in this case CEDs
2017	The year that this version of the “TO” region was released
.csv	The format that the file is being supplied in, Comma Separated Value

Appearance and Definitions

As the file is in CSV format, double clicking on it will automatically open the file in Microsoft Excel and it will be displayed similarly to the example below.

SA2_MAINCODE_2016	SA2_NAME_2016	CED_CODE_2017	CED_NAME_2017	RATIO_FROM_TO	INDIV_TO_REGION_QLTY_INDICATOR	OVERALL_QUALITY_INDICATOR	BMOS_NULL_FLAG
101021007	Braidwood	113	Eden-Monaro	0.9317249	Good	Good	0
101021007	Braidwood	120	Hume	0.0682751	Acceptable	Good	0
101021008	Karabar	113	Eden-Monaro	1	Good	Good	0
101021009	Queanbeyan	113	Eden-Monaro	1	Good	Good	0
101021010	Queanbeyan - East	113	Eden-Monaro	1	Good	Good	0
101021011	Queanbeyan Region	113	Eden-Monaro	0.9744189	Good	Good	0
101021011	Queanbeyan Region	120	Hume	0.0255811	Acceptable	Good	0
101021012	Queanbeyan West - Jerrabomberra	113	Eden-Monaro	1	Good	Good	0
101031013	Bombala	113	Eden-Monaro	1	Good	Good	0
101031014	Cooma	113	Eden-Monaro	1	Good	Good	0
101031015	Cooma Region	113	Eden-Monaro	1	Good	Good	0
101031016	Jindabyne - Berridale	113	Eden-Monaro	1	Good	Good	0
101041017	Batemans Bay	116	Gilmore	1	Good	Good	0
101041018	Batemans Bay - South	116	Gilmore	1	Good	Good	0
101041019	Bega - Tathra	113	Eden-Monaro	1	Good	Good	0

Below is a description of each column heading.

SA2_MAINCODE_2016
This is a unique code associated with each “FROM” region, in this case the unique 2016 SA2 code.

SA2_NAME_2016
This is a textual label associated with the unique code of the “FROM” region, in this case the textual label for each 2016 SA2.

CED_CODE_2017
This is the unique numerical code representing the “TO” region, in this case the unique 2017 CED code.

CED_NAME_2017
This is a textual label associated with the unique code of the “TO” region, in this case the textual label for each 2017 CED.

RATIO_FROM_TO
This field describes the Ratio of the “FROM” region that is being donated to the “TO” region. The Ratio is a figure between 0 and 1.

INDIV_TO_REGION_QLTY_INDICATOR
This field describes how well data is likely to be converted to the "TO" region. There are three values associated with the Individual Quality Indicator:

> Good – The ABS expects that for this "TO" region the correspondence will convert data to a high degree of accuracy and users can expect the converted data will reflect the actual characteristics of the geographic regions involved.
>
> Acceptable – The ABS expects that for this "TO" region the correspondence will convert data to a reasonable degree of accuracy, though caution needs to be applied as the quality of the converted data will vary and may differ from the actual characteristics of the geographic regions involved.
>
> Poor – The ABS expects that for this "TO" region there is a high likelihood the correspondence will not convert data accurately and that the converted data should be used with caution as it may not reflect the actual characteristics of many of the geographic regions involved.

OVERALL_QUALITY_INDICATOR
An overall quality indicator is applied to each correspondence. The aim of this is to provide users with a reasonable idea of how well the correspondence will convert data across the whole of the correspondence. There are three values associated with the Overall Quality Indicator and they are identical to those listed above for the individual quality indicator.

BMOS_NULL_FLAG
This is the acronym for the Below Minimum Output Size/Null values flag. The CSV file may contain records that are below a pre-set minimum output size (typically below a ratio of 0.01). These are records where the proportion of the "FROM" region that are being donated is very small and is deemed as being statistically insignificant. Records that fall in to this category will be flagged by a specific code as described below.

Null values are records where part or all of a "FROM" region does not have a corresponding "TO" region, or vice versa. An example of when this may occur is when one geography dataset contains islands which are not included in the other dataset. If the correspondence contains Null records, they will be flagged by a specific code as described below.

BMOS_NULL_FLAG Codes

> 0 – Record is not below the specified minimum output size and has both a "FROM" and "TO" region present.
> 1 – Record is below the specified minimum output size, but both a "FROM" and "TO" record are present.
> 2 – Record has a null "TO" unit.
> 3 – Record has a null "FROM" unit.
> 4 – Record is below the specified minimum output size and the "TO" unit is null.
> 5 – Record is below the specified minimum output size and the "FROM" unit is null.

Further Information

More information on the ASGS and ABS Statistical Geography can be found by visiting the ABS website: http://www.abs.gov.au/geography