NISRA
Northern Ireland
Statistics and Research Agency
Gníomhaireacht Thuaisceart Éireann
um Staitisticí agus Taighde

Aggregating Data Zones to produce statistics for higher-level geographies

February 2023

www.nisra.gov.uk/geography

Contents

1. Introduction

The phased release of standard Census 2021 outputs by the Northern Ireland Statistics and Research Agency (NISRA) up to summer 2023 is focussed on reporting information for the following geographies below Northern Ireland (NI) level: Local Government District; District Electoral Area; Super Data Zone; and Data Zone. These four geographies are hierarchical, with the 3,780 Data Zones nesting within the 850 Super Data Zones, which in turn nest within the 80 District Electoral Areas and 11 Local Government Districts.

The Data Zones and Super Data Zones have been newly created by NISRA to support Census 2021 outputs, while Local Government District and District Electoral Area are established administrative geographies, introduced in 2015. The paper 'New statistical output geographies for Northern Ireland derived from Census 2021' (PDF, 7 MB) provides information on the development of the new census geographies.

NISRA recognises that there will be user interest in Census 2021 outputs for other geographies outside of this hierarchy. This is based on the responses to the outputs consultation held in 2021, with some users indicating a requirement for Census 2021 outputs by geographies other than Local Government District. The range of geographies that Census 2011 outputs are available for via the commissioned table service and the NI Neighbourhood Information Service is further evidence of this user interest. The former Local Government Districts, Electoral Wards, Assembly Areas/Parliamentary Constituencies, Health and Social Care Trusts and Super Output Areas are among the most popular alternative geographies (note that Super Output Areas are a statistical rather than administrative geography but have been grouped with the others for the purpose of this paper).

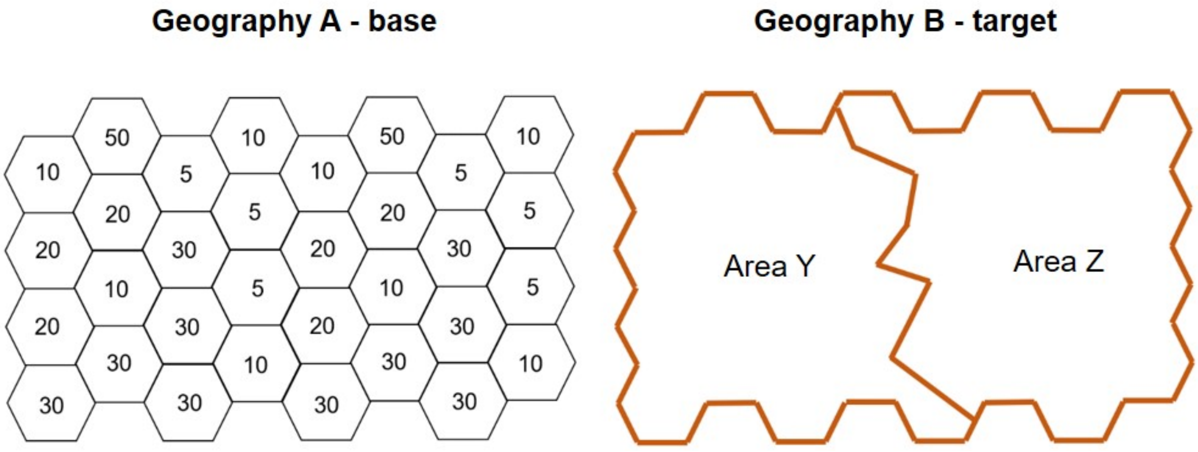
NISRA is considering how best to meet this user need; however, this must be balanced with the requirement to protect the census data and prevent individuals and/or households from being identified in published information, the risk of which is increased when data are released for multiple geographies.

This paper has three objectives:

- describes the use of aggregated Data Zones to produce approximated Census 2021 statistics for certain administrative geographies outside of the aforementioned four-level hierarchy
- outlines the NISRA policy for determining whether a given administrative geography is suitable for the Data Zone aggregation method
- presents an alternative aggregation method to produce approximated Census 2021 statistics for administrative geographies deemed unsuitable for Data Zone aggregation

2. Data Zone aggregation method

The basic principle of the aggregation method is one-to-one relationships between the base and target geographies using a best-fit approach (illustrated in Figure 1). Figure 1a shows two simplified geographies, A and B, that cover the same location; A comprises 32 small areas and B is made up of two large areas (Y and Z). The population count for each of the areas in A is included.


Geography A - base

Geography B - target

Figure 1a. Base and target geographies for application of aggregation method.

The aim is to select areas in the base geography A to produce an approximated representation for the larger target geography B. When geography B is put on top of geography A, the base areas wholly within one of the two target areas are assigned to those (Figure 1b).

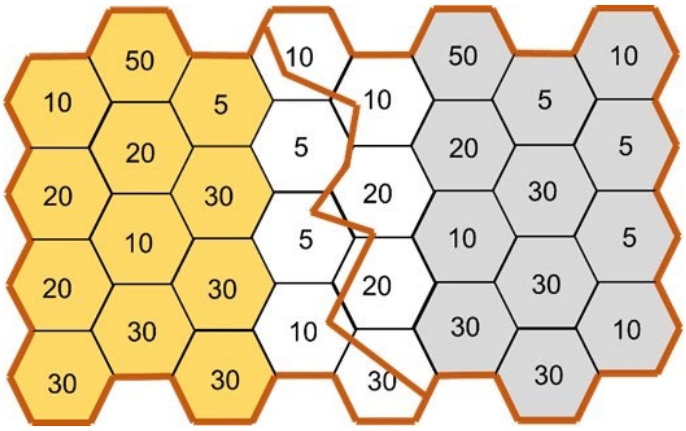


Figure 1b. Base geography areas assigned to target geography areas on the basis of being wholly within the target geography area (gold- and grey-coloured areas assigned to area Y and Z of target geography, respectively).

For areas in geography A that are intersected by the boundary running through the centre of geography B, they are assigned to the target area containing the majority of its population. Figure 1c shows the distribution of household XY coordinates in the topmost geography A area intersected by the central boundary of geography B. Among the six households in this area, the largest proportion of usual residents are in area Z of geography B; therefore this base geography area is assigned to area Z.

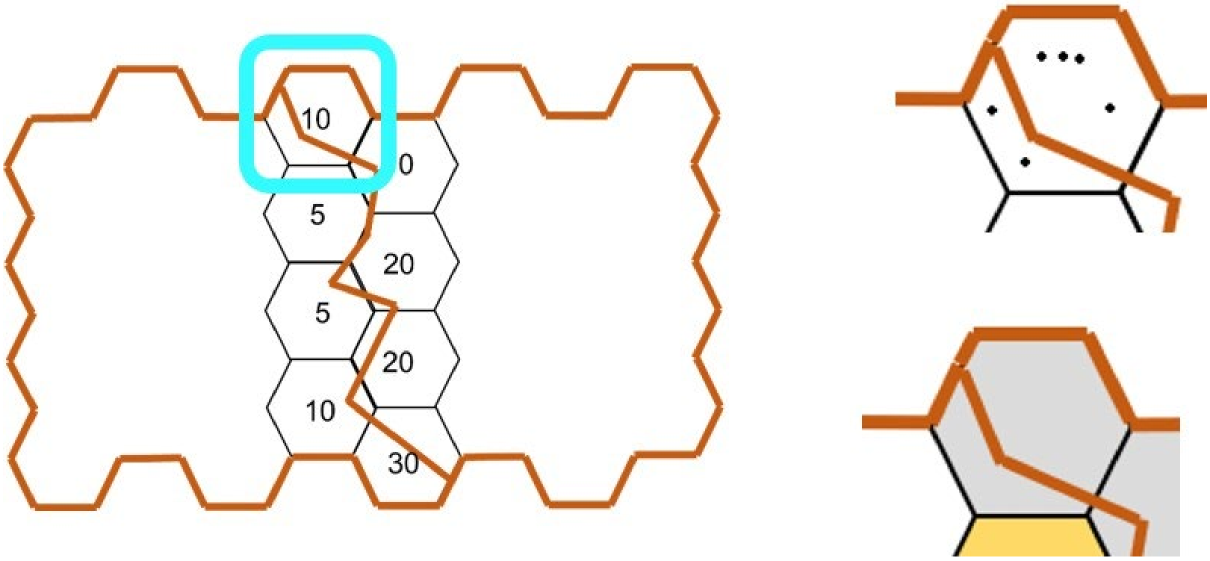


Figure 1c. Allocation of topmost base geography area intersected by central boundary of target geography; largest proportion of population in the four households to the right of the boundary resulting in this base geography area being assigned to area Z (grey-coloured).

In terms of the aggregation method, the approximated population count for areas Y and Z of geography B is the sum of the counts for the gold- and grey-coloured base areas in geography A (305 and 325 for areas Y and Z, respectively).

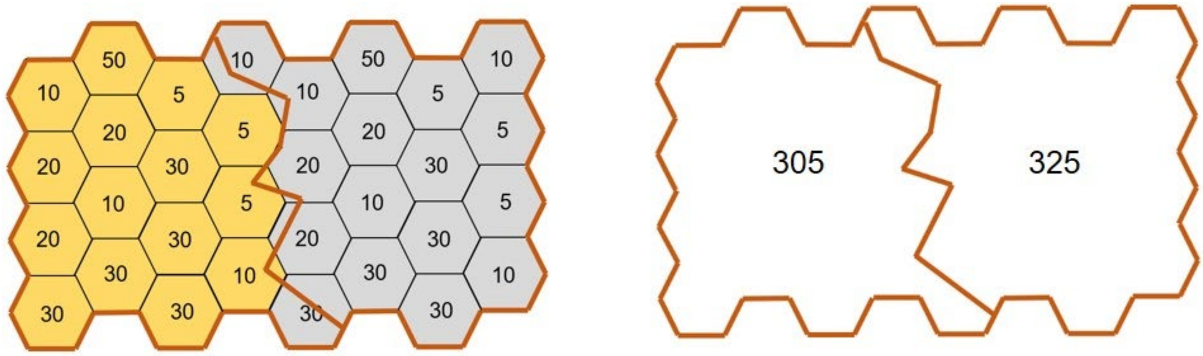


Figure 1d. Overall allocation of base geography areas (colour-coded) and resulting approximated population count for target geography areas Y and Z.

The method dictates that each area in geography A is assigned to a single target area in geography B; each geography is approximated in this manner. This influences the accuracy of the resulting counts since the approximated count for the target areas will differ to the exact count. This is on the basis of intersected areas being wholly allocated to target areas according to their population proportion. As a general rule, the aggregation method is more accurate the smaller/more numerous the base geography areas are in comparison to the target geography areas. For example, the application of the aggregation method using the 3,780 Data Zones produces more accurate approximated counts for the 18 Parliamentary Constituencies than it does for the 462 Electoral Wards; this is due to a smaller proportion of the base geography areas being intersected by a target geography boundary.

3. Data Zone aggregation policy

The identification of administrative geographies that can be approximated sufficiently accurately by the Data Zone aggregation method is determined by an accuracy threshold of five percent. The approximated counts for a target geography using the Data Zone aggregation method are compared with the exact counts (generated using a Geographic Information System program); the counts refer to usual resident persons and households from Census 2021. The policy being implemented by NISRA is as follows:

1) where the absolute difference across all target geography areas is less than five per cent, Data Zone aggregation is deemed a sufficiently accurate method of producing Census 2021 statistics for the target geography
2) where the absolute difference across any of the target geography areas is more than five per cent, Data Zone aggregation is not considered to be a sufficiently accurate method of producing Census 2021 statistics for the target geography
3) if an administrative geography is unsuitable for the Data Zone aggregation method according to criterion (2), NISRA will assess the use of Grid Squares as a suitable base geography for aggregation

4. Geographies assessed for Data Zone aggregation

NISRA has assessed a range of administrative/statistical geographies in terms of their suitability for the Data Zone aggregation method. This involved the creation of a spatial lookup between Data Zone and each geography based on the method outlined in section 2.

4.1. Suitable for Data Zone aggregation

Table 1 lists the geographies that NISRA considers suitable for Data Zone aggregation based on being within the five per cent accuracy threshold for both the usual resident person and household count from Census 2021.

Table 1. Administrative geographies considered suitable for the Data Zone aggregation method of generating Census 2021 statistics, based on meeting the specified accuracy requirement.

Administrative geography	Number of areas in geography	Number of areas (proportion in parentheses) within the five per cent accuracy threshold	
		Census 2021 usual resident persons	Census 2021 usual resident households
Local Government District [1]	11	11 (100%)	11 (100%)
District Electoral Area [1]	80	80 (100%)	80 (100%)
County	6	6 (100%)	6 (100%)
Parliamentary Constituency/ Assembly Area	18	18 (100%)	18 (100%)
Health and Social Care Trust	5	5 (100%)	5 (100%)
Former Local Government District	26	26 (100%)	26 (100%)
Settlement [2]			
Band A – Belfast City	1	1 (100%)	1 (100%)
Band B – Derry City	1	1 (100%)	1 (100%)
Band C – Large town	14	14 (100%)	14 (100%)

Administrative geography	Number of areas in geography	Number of areas (proportion in parentheses) within the five per cent accuracy threshold	
		Census 2021 usual resident persons	Census 2021 usual resident households
Band D – Medium town	10	10 (100%)	10 (100%)
Band E – Small town	17	14 (82%)	17 (100%)
Band F – Intermediate settlement	24	20 (83%)	21 (88%)

[1] Data Zones nest within District Electoral Area and Local Government District; these two geographies are included in the table for completeness.

[2] Settlement as defined in the Review of the Statistical Classification and Delineation of Settlements (2015) (PDF, 1.2 MB)

The large size of the areas in the County, Parliamentary Constituency/Assembly Area, Health and Social Care Trust and former Local Government District geographies lend themselves to suitability for application of the Data Zone Aggregation method. To add, regarding Parliamentary Constituency/Assembly Area, analysis has shown that the Data Zone aggregation method is also suitable when used with the latest proposed boundaries (as of February 2023) as part of the ongoing 2023 Review of Parliamentary Constituencies in NI.

As described in the 'New statistical output geographies for Northern Ireland derived from Census 2021' (PDF, 7 MB) information paper, one of the design aims in building the Data Zone geography was to broadly align them with the larger settlements; this is reflected in the Data Zone aggregation method meeting the accuracy requirement for most settlements in Band A to F. Although not all of the 17 Band E settlements are within the five per cent accuracy threshold in terms of Census 2021 usual resident persons, NISRA approves the use of the Data Zone aggregation method for the 14 settlements in this Band that are within the accuracy threshold. The three above-threshold settlements are intersected by a District Electoral Area boundary in a manner that impacts the broad alignment of Data Zones with the extent of these settlements. When the census usual resident household count is used, these three settlements are within the five per cent accuracy threshold. This is due to the different spatial distributions of these two variables; the household count is solely based on the Irish Grid X and Y coordinates of each usual resident household, whereas the person count is based on the number of usual residents in each household along with those residing in communal establishments such as student halls of residence and care homes.

The same scenario is evident with three of the twenty-four Band F settlements, while the location of a fourth settlement in this Band in relation to its District Electoral Area boundary was unsuited to alignment with the Data Zones. However, the other 20 Band F settlements are within the five percent accuracy threshold for both census usual resident persons and households, so NISRA endorses the use of the Data Zone aggregation method for these. Table A1 in the Annex lists the settlements deemed suitable for Data Zone aggregation.

4.2. Unsuitable for Data Zone aggregation – Alternative method

As outlined in section 4.1, while the Data Zone aggregation method can be applied to the majority of the Band E and F settlements, a small number fail to meet the required level of accuracy, namely (in alphabetical order within each Band):

- Coalisland (Band E)
- Greenisland (Band E)
- Magherafelt (Band E)
- Castlewellan (Band F)
- Culmore (Band F)
- Hillsborough and Culcavy (Band F)
- Keady (Band F)

In addition, the following administrative/statistical geographies do not meet the five per cent accuracy requirement and are therefore considered by NISRA to be unsuitable for application of the Data Zone aggregation method:

- Electoral Ward (462 areas)
- Former Electoral Ward (582 areas)
- Super Output Area (890 areas)
- Neighbourhood Renewal Area (36 areas)

The base geography used to aggregate Census 2021 statistics for these geographies needs to be smaller than Data Zone. NISRA will use census Grid Squares and is currently assessing the accuracy of this approach. Note that a Grid Square aggregation method would rely on the publication of Census 2021 data for the Grid Square geography, which is currently scheduled for no earlier than summer 2023 in the release plans for Census 2021 statistics. Furthermore, as with Census 2011, the published Census 2021 Grid Square product will provide a more restricted range of outputs compared to the standard outputs; this is to ensure the confidentiality of individual census returns. Given the wide range of census outputs that will ultimately be released down to Data Zone level, the publication of further outputs for an additional geographic base (Grid Squares) increases the risk of statistical disclosure via differencing. Consequently, Grid Square outputs are mainly univariate.

5. Considerations

5.1. Appropriate data

The Data Zone aggregation method relates to count data only; it does not apply to percentages, rates or other derived statistics. Users wishing to create such statistics will need to apply this guidance to the baseline count data before creating the derived statistics.

5.2. Other variables

The described accuracy measures are based on Census 2021 usual resident person and household counts. For other Census 2021 statistics relating to the likes of population sub-groups or locally clustered groups (for example, the elderly or ethnic minorities), there is less certainty that the Data Zone aggregation method will provide this level of accuracy.

5.3. Impact of statistical disclosure control

In general, the use of aggregation to generate statistics for higher level geographies should be done with caution, taking into account any statistical disclosure control (SDC) methods applied to the data. The ‘noise’ introduced to the data via the application of SDC methods is propagated by aggregation.

As an example, all Census 2021 information released by NISRA is subjected to an SDC method called ‘cell key perturbation’. This approach adds ‘noise’ to the data to protect against the disclosure of information on individuals, households or groups. The method involves making small changes to cells in output tables, and has more impact on sparse data (for example, small geographic areas). This SDC method is described in more detail in our guidance note ‘Statistical disclosure control methodology for 2021 Census’ (PDF, 206 KB).

When NISRA produces census information for the administrative geographies in Table 1, this will be based on Data Zone aggregation; however, the perturbation will be applied to the final table as opposed to the base Data Zones, thereby minimising the ‘noise’ propagation.

If users wish to aggregate census data themselves, the guidance is to use larger base geographies if data are available for them and they are located within the target geography. For example, if a user wanted to create an area that was equivalent to the Ards and North Down Local Government District plus a small bit of East Belfast, it would be more accurate to add the figures for Ards and North Down to the relevant figures for the Data Zones that approximate the part of East Belfast, rather than aggregate the figures for all Data Zones in the defined area.

Annex

Table A1. Settlements considered by NISRA to be suitable for application of the Data Zone aggregation method to produce Census 2021 statistics.

Band as defined in the 2015 Review of the Statistical Classification and Delineation of Settlements (PDF, 1.2 MB)	Settlement
A	Belfast City
B	Derry City
C (Large town)	Metropolitan Newtownabbey
	Craigavon Urban Area including Aghacommon
	Bangor
	Metropolitan Castlereagh
	Lisburn City
	Metropolitan Lisburn
	Ballymena
	Newtownards
	Carrickfergus
	Newry
	Coleraine
	Antrim
	Omagh town
	Larne
D (Medium town)	Banbridge
	Armagh
	Dungannon
	Enniskillen
	Strabane
	Limavady
	Cookstown
	Holywood
	Downpatrick
	Ballymoney
E (Small town)	Ballyclare
	Comber
	Warrenpoint/Burren
	Portstewart

Band as defined in the 2015 Review of the Statistical Classification and Delineation of Settlements (PDF, 1.2 MB)	Settlement
	Newcastle
	Carryduff
	Donaghadee
	Kilkeel
	Portrush
	Dromore (Armagh City, Banbridge and Craigavon)
	Ballynahinch
	Ballycastle
	Crumlin
	Randalstown
F (Intermediate settlement)	Moira
	Maghera
	Whitehead
	Eglinton
	Waringstown
	Tandragee
	Saintfield
	Ahoghill
	Dungiven
	Castlederg
	Lisnaskea
	Ballygowan
	Killyleagh
	Broughshane
	Richhill
	Rostrevor
	Bessbrook
	Newbuildings
	Cullybackey
	Portaferry