Scotland's Census
Shaping our future
A' dealbhadh ar n-àm ri teachd

Scotland's Census 2022

Census Coverage Survey: Communal Establishment Sample Methodology

July 2020

National Records of Scotland

Contents

1. Plain English Summary 3
2. Executive Summary 3
3. Introduction and Background 4
4. Summary of 2011 Methodology 5
4.1 Small Communal Establishments 5
4.2 Large Communal Establishments 6
5. Proposed 2022 Methodology 6
5.1 Small Communal Establishments 6
5.2 Large Communal Establishments 6
5.3 Communal Establishment Boost Sample 7
6. Methods Considered 8
6.1 Boost Sample Allocation and Weighting 8
6.2 Boost Sample Stratification 10
6.2.1 Geographical Stratification 10
6.2.2 Estimation Area (EA) and Communal Establishment Type 13
6.2.3 Bed Space Grouping and CE Type 17
6.2.4 Proposed Approach 19
6.3 Boost Sample Clustering 20
6.3.1 Bias Analysis 20
6.3.2 Distance Analysis 22
6.3.3 Operational Considerations 23
6.3.4 Proposed Approach 25
7. Strengths and Limitations of Methodology 25
7.1 Allocation and Weighting 25
7.2 Stratification 25
7.3 Geographically Clustered Approach 26
8. Conclusion and Recommendations 26
9. References 28
Annex 1: Glossary of Acronyms 29
Annex 2: 2011 Census Processing Units 30

1. Plain English Summary

The Census Coverage Survey (CCS) is a survey that takes place after the census, to help create more accurate population estimates. Communal Establishments (CEs) are residential accommodations that are managed, for example hospitals or hotels. These are included in the CCS, however to improve on the method in 2011, we plan to include a Boost sample of CEs in 2022 to increase the total sample size.
We looked at different ways of grouping CEs across the country together, and suggest grouping CEs by location and Establishment Type (i.e. hospitals will be grouped separately from hotels). We plan to spread the sample across these groups based on the number of people that are within the groups, rather than the number of establishments themselves. We looked into clustering the sample so that only CEs that are close to CCS areas are included in the sample, but this does not seem to provide significant benefits, so we suggest using a non-clustered sample.

2. Executive Summary

As in 2011, the Census Coverage Survey (CCS) in 2022 will enumerate small Communal Establishments (CEs, under 100 bed spaces) opportunistically within its sample, and large CEs (100 bed spaces or more) will undergo a manual adjustment process. To improve on the statistical methodology from 2011, a Boost sample of small CEs is proposed in addition to CEs sampled within CCS areas. The overall CE sample is aimed to be 250, with the number of CEs included in the CCS sample informing the size of the CE Boost sample.

The proposal is to allocate the CE Boost sample to strata proportionally – based on the size of CEs, as opposed to the number of CEs themselves. For current purposes, the number of usual residents is the optimal measure of CE size, however usual resident information won't be available until January 2022, therefore bed space information was used as an alternative measure of CE size. Allocation proportions were created by calculating the proportion of bed spaces in each strata of the total bed spaces across the sample frame (a source list of CEs from which the sample will be drawn). These proportions were then weighted to account for the average number of bed spaces within strata, in order to reflect average CE sizes across strata.

Various stratification options were considered, with the proposed option involving stratification by Estimation Area (see Annex 2) and collapsed higher order CE type.

A geographically clustered approach was considered as a method of increasing the operational efficiency and reducing travel times associated with the enumeration of CEs within the Boost sample. This would be implemented by creating a sample frame that includes only the closest CEs to existing CCS areas. There was, however, concerns around the potential of intrinsic bias associated with this approach, given that some establishments by nature are more likely to be situated in closer proximity to heavily populated areas. Analysis was conducted, and inconsistent distributions across clustered and non-clustered sample frames supported the suggestion that the clustered approach may introduce bias. Additionally, operational timings of adopting a geographically clustered approach were considered, and distance analysis was conducted to assess the magnitude of impact on travel times that a clustered approach would yield. When evaluating the statistical and operational considerations together, the proposal is to implement a non-clustered approach.

3. Introduction and Background

The Census Coverage Survey (CCS)[1] is a voluntary, interviewer led, follow-up survey that is conducted six weeks after census day and samples approximately 1.5-2% of the household population of Scotland. The primary aim of the CCS is to collect data from a representative sample that can be matched to the 2022 Census data to inform the level of over or under coverage of the census data. This is achieved by matching the CCS data to the census data to determine persons captured by both, or those counted in one but not the other. The matched data is then put through Dual System Estimation (DSE)[2] which estimates the number of persons who may have been missed overall. The census data is then adjusted to account for these missing

[1] More information on the CCS sample methodology can be found in the CCS Sample Methodology and the CCS Sample Allocation and Reserve Sample papers
https://www.scotlandscensus.gov.uk/external-methodology-assurance-panels-emaps-0

[2] More information on DSE can be found in the Estimation and Adjustment methodology paper
https://www.scotlandscensus.gov.uk/documents/Scotlands%20Census%202021%20-%20SMDP%20-%20Estimation%20and%20Adjustment%20Methodology%20paper%20(pdf).pdf

persons, thereby providing a more complete estimate of the true population count.

In order for an accurate estimation process, the CCS sample needs to be an adequate and accurate reflection of the Scottish population. Communal Establishments (CEs), classified in Scotland's Census as managed residential accommodations, are an important component in the census enumeration process, and as such the CCS needs to account for the inclusion of CE data in some manner. There are currently separate enumeration options for small CEs in 2022 (<100 bed spaces) and large CEs (≥100 bed spaces). This paper will discuss the enumeration strategy for CEs in the 2011 CCS, and outline the chosen methodology for the enumeration of CEs in the 2022 CCS.

Note: On 17 July 2020 Scottish Government announced the decision to move Scotland's Census to 20 March 2022 following the impact of the COVID-19 pandemic.

4. Summary of 2011 Methodology

4.1 Small Communal Establishments

Small Communal Establishments (CEs, under 100 bed spaces) were enumerated in 2011 as an opportunistic sample, whereby if a small CE appeared within a CCS sample area then this was enumerated as part of the sample. Therefore, the sample methodology for CEs was consistent with the main CCS sample. Sample stratification consisted of geographical and demographic components, using Local Authority (LA) and the Hard to Count Index (HtC)[3] to create strata that ensure the sample is representative and evenly spread across the population.

Small CEs underwent Dual System Estimation (DSE), with estimates produced for ages 0-60 and 60+, and for all establishment natures. This was done all together across the whole of Scotland, rather than within Estimation Areas (see Annex 2) used for estimates of the population within households. Particular establishment

[3] More information on HtC can be found in the Developing a Hard-to-count Index paper https://www.scotlandscensus.gov.uk/external-methodology-assurance-panels-emaps-0

natures, such as prisons and defence establishments were excluded from the sample. The estimates only increased the number of residents within establishments existing in the Census data, and did not generate any new establishments.

4.2 Large Communal Establishments

Large Communal Establishments (CEs, with 100 or more bed spaces) were not included in the enumeration strategy for the CCS in 2011 given the time and resource that would be required for successful enumeration. Instead, a manual adjustment method was used. The number of returned questionnaires was compared to the number of questionnaires issued to the establishment as well as the number of bed spaces, to indicate if there was a considerable difference. In these cases where there was a considerable difference, the establishment was contacted to verify the correct number of usual residents of the establishment. The population count within the establishment was then increased proportionately to the identified number of usual residents, with new records generated in the Adjustment process to make up the shortfall.

5. Proposed 2022 Methodology

5.1 Small Communal Establishments

A similar approach is planned for small Communal Establishments (under 100 bed spaces) within the main CCS sample in 2022, whereby any CEs that fall within CCS areas will be opportunistically sampled. Similarly to 2011, Dual System Estimation (DSE) will be used for small Communal Establishments. These will be stratified by age group and establishment nature, with room to collapse this stratification if the number of responses is not sufficient. As in 2011, this will be done across the whole of Scotland, as opposed to within Estimation Areas. In addition, a Boost sample of small CEs is proposed, discussed further in Sections 5.3 and 6.

5.2 Large Communal Establishments

Similarly to 2011, large Communal Establishments will not be included within the CCS sample, instead being manually adjusted for. In 2022, an additional question on the CE Managers' Census Questionnaire will allow the collection of information around how many male and female usual residents each establishment has, broken down into different age groups. This will allow us to make a similar correction to that of 2011, without needing to contact the establishment unless further information is required. As this usual resident information is broken down into age groups, we can also be more accurate in the records we generate in an establishment to reflect the different response rates within age groups.

5.3 Communal Establishment Boost Sample

In 2022, a Boost sample of Communal Establishments is proposed to increase the overall sample size of CEs within the CCS. In the 2011 CCS, the small CE sample size resulted in estimate stratification by age alone, reducing the granularity of the estimates. By incorporating a boost sample to increase the sample size of CEs enumerated in the 2022 CCS, this should increase estimate granularity through more comprehensive stratification.

The Boost sample approach involves sampling additional CEs out-with CCS areas, with the number of opportunistically sampled CEs within CCS areas informing the Boost sample size – to provide a total CE sample of 250. This sample size, alongside the reduced individual question set being implemented in 2022, was chosen based on previous research by NRS that showed the improved simulated Relative Standard Error (RSE=0.465% for boosting to a total of 200 CEs) compared to the 2011 approach (RSE=0.7%), as well as practical and financial considerations.

The proposed source for this sample is an extract of the Communal Establishment Register (CER) that is collated by the Enumeration team from Census records, which contains information from every CE in Scotland recorded in the Census. The latest version of this will be finalised after the Enumeration Address Check is undertaken, expected in January 2022. At this point, a sample frame can be created from this list, reflecting only the CEs that are within scope for the CCS, i.e. CEs under 100 bed

spaces. Armed Forces and Detention establishments are excluded from the CCS given the typically high number of bed spaces and difficulty enumerating as a result of access issues, for example.

6. Methods Considered

6.1 Boost Sample Allocation and Weighting

The preferred allocation option for the CE Boost sample is proportional allocation. Although this does not take into account variation in response rates across strata, given the comparably small sample size required for the additional boost sample of CEs, this may not pose a significant risk. When considering proportional allocation, there are two main questions to be considered. The first relates to the definition of stratum size, upon which the sample would be allocated proportionally; and the second is whether this is achievable based on the strata selected.

When defining stratum size, the number of residents is of more interest than number of CEs themselves within a stratum, as in general only usual residents are enumerated in CEs within the CCS. For example, a stratum could contain a large number of CEs, but if these CEs have low numbers of usual residents in them, the volumes of potential data to be collected would be low. Conversely, a small number of CEs may have high numbers of usual residents, increasing the level of data to be gathered. Therefore, defining stratum size by potential volumes of data that could be collected i.e. number of usual residents, appears more informative and meaningful than the number of CEs located within a stratum.

Given that usual resident information will not be available until the latest version of the CER extract is available (as mentioned, this is expected in January 2022), bed space information was used as an alternative measure of establishment size.

Allocation proportions were calculated by dividing the total number of bed spaces in each stratum by the total number of bed spaces in the sample frame as a whole. These were then weighted to account for varying CE size across strata, to reflect the average number of bed spaces in the strata. For each stratification option, a

weighting was created by dividing the mean number of bed spaces across the sample frame by the mean number of bed spaces across each stratum. These weightings were then applied to the corresponding allocation proportions, in order to more accurately reflect variations in average bed space numbers across the strata. This should avoid bias towards larger CEs, given their proportionately larger number of bed spaces. An example of weightings being applied is illustrated in Table 1.

Stratum	*Allocation Proportion*	*Total Mean Beds*	*Mean Beds per Stratum*	*Weighting*	*Weighted Allocation Proportion*
1	0.053215	33.9	28.8	1.177773	0.062676
2	0.009968	33.9	21.3	1.590838	0.015858
3	0.032956	33.9	54.8	0.618655	0.020388

Table 1: Weightings were then calculated to reflect the average number of bed spaces in each stratum. These were then applied to the original proportions to create weighted allocation proportions, to reflect variance in stratum size.

When considering stratification by Estimation Area and higher order CE type – discussed in more detail in Section 6.2.2 – the bed space distribution of the sample frame (solid line) was compared to that of an unweighted sample drawn (dashed line), as demonstrated in Figure 1.

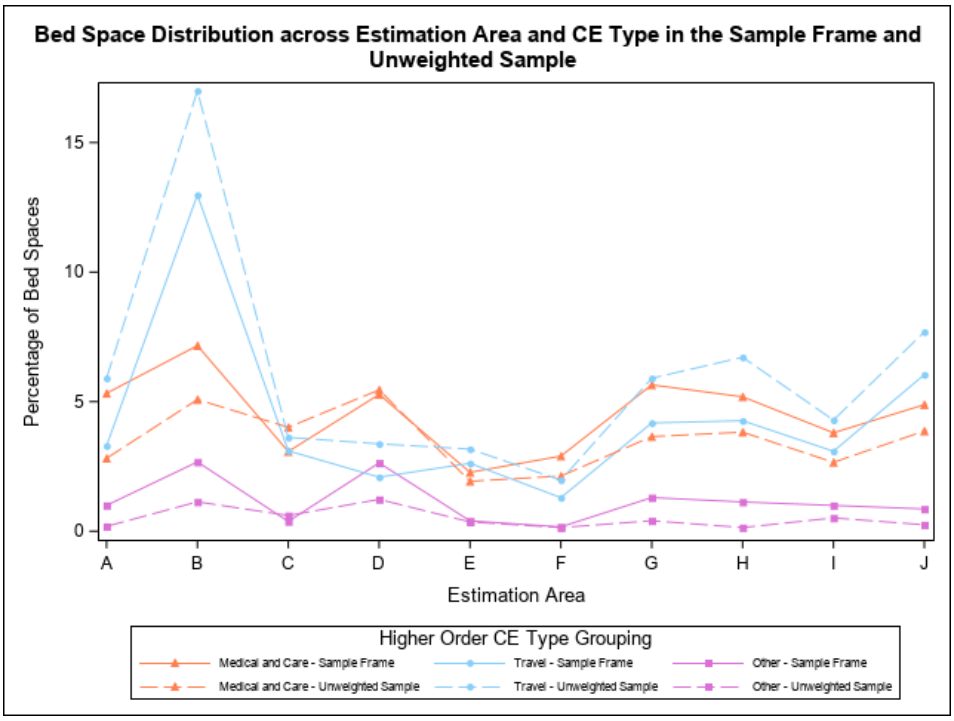

Figure 1: Scatter plot illustrating the distribution of bed spaces across Estimation Areas and CE Type in the Sample Frame (solid line) compared to the Unweighted Sample (dashed line).

Weightings were then applied, and the distribution of bed spaces across the Sample Frame (solid line) were compared to that of the newly weighted sample (dashed line), as seen in Figure 2.

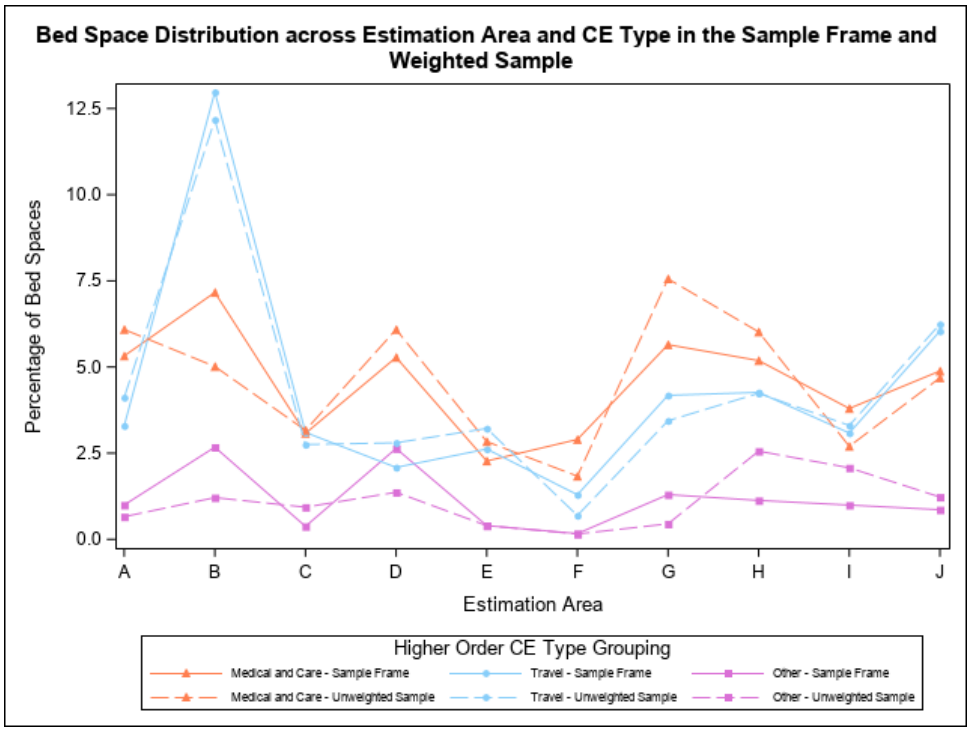

Figure 2: Scatter plot illustrating the distribution of bed spaces across Estimation Areas and CE Type in the Sample Frame (solid line) compared to the Weighted Sample (dashed line).

When these weightings are applied to account for the average size of CE across strata, a more proportionate distribution can be seen when comparing the Sample Frame and the Sample – particularly evident in Travel Establishments in Estimation Area A and J.

Given the beneficial impact on the representativeness of the bed space distribution of the sample when compared to the sample frame, the proposal is to use weightings in the creation of allocation proportions of the CE Boost Sample.

6.2 Boost Sample Stratification

6.2.1 Geographical Stratification

The option of no stratification was considered, however given that this approach would not yield meaningful strata and would subsequently result in unrepresentative

samples, a lack of stratification was ruled out and other stratification options were considered.

The additional Boost CE sample could be stratified geographically – which would ensure an even coverage across the country. If stratifying by Local Authority (LA, see Annex 2), this would result in 32 strata. The total number of CEs in each LA was investigated, illustrated by Figure 3.

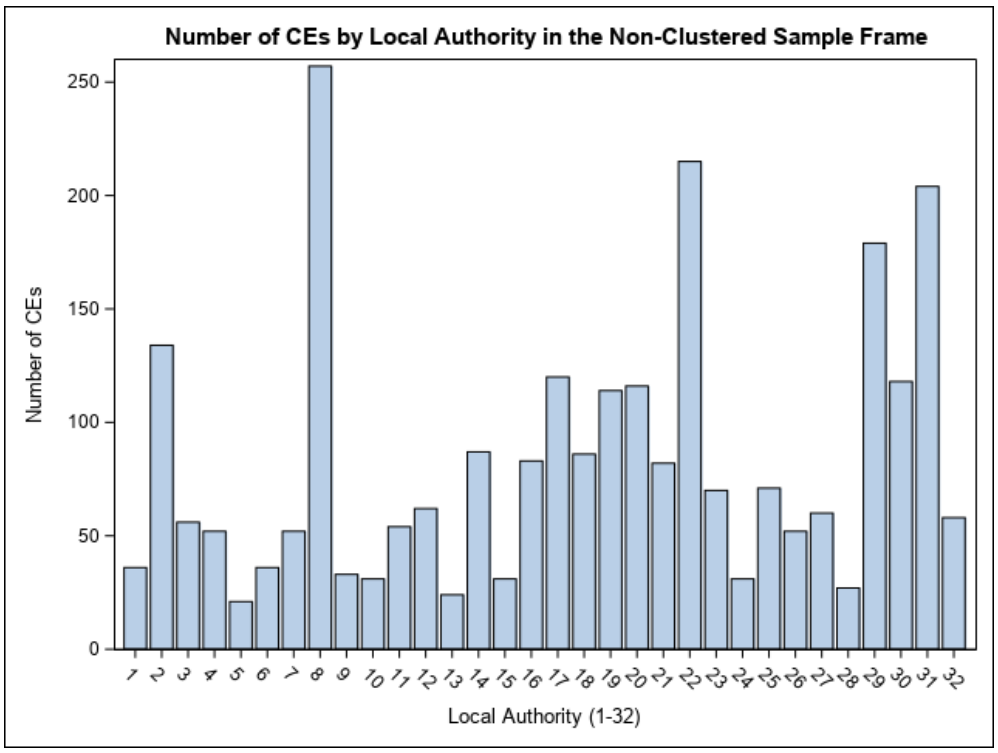

Figure 3: Bar Chart illustrating number of CEs within each Local Authority.

The purpose of this was to assess the potential level of collapsing that may be required if stratifying by LA. If some LAs contained few or no CEs, then these strata would require collapsing. However, even the LA with the lowest number of CEs within it (LA 5 with 21 CEs within it) has a sufficient size that significant collapsing should not be required. Therefore, LA appears an appropriate stratification variable if used individually. However, if stratifying the sample by another variable in combination with a geographical component, the use of LAs create too many strata – which can result in over-stratification and the subsequent risk of almost flat allocation, whereby the sample frames are insufficient to allocate proportionally.

An alternative in this instance would be considering 2011 Estimation Areas (EAs) as opposed to Local Authorities, with EAs being larger geographical areas which would decrease the number of strata. The ten Estimation Areas correspond to the ten Processing Units (PUs) used in 2011, with the Estimation & Adjustment Identifiers (E&A Identifiers, see Annex 2) corresponding to the EA grouping (e.g. PU B = EA B).

There is work being undertaken currently to update the Estimation Areas for 2022[4], however the areas used in 2011 are geographically contiguous and therefore form conventional geographic areas. Conversely, the proposed 2022 areas are guided by response characteristics for the household population, which is not relevant within the context of CEs. Therefore, although not the current EA groupings, the 2011 groupings appear more relevant for the purposes of CE stratification.

The number of CEs in each EA was investigated, and Figure 4 illustrates the more consistent spread of CEs across the country when using this form of geographical division.

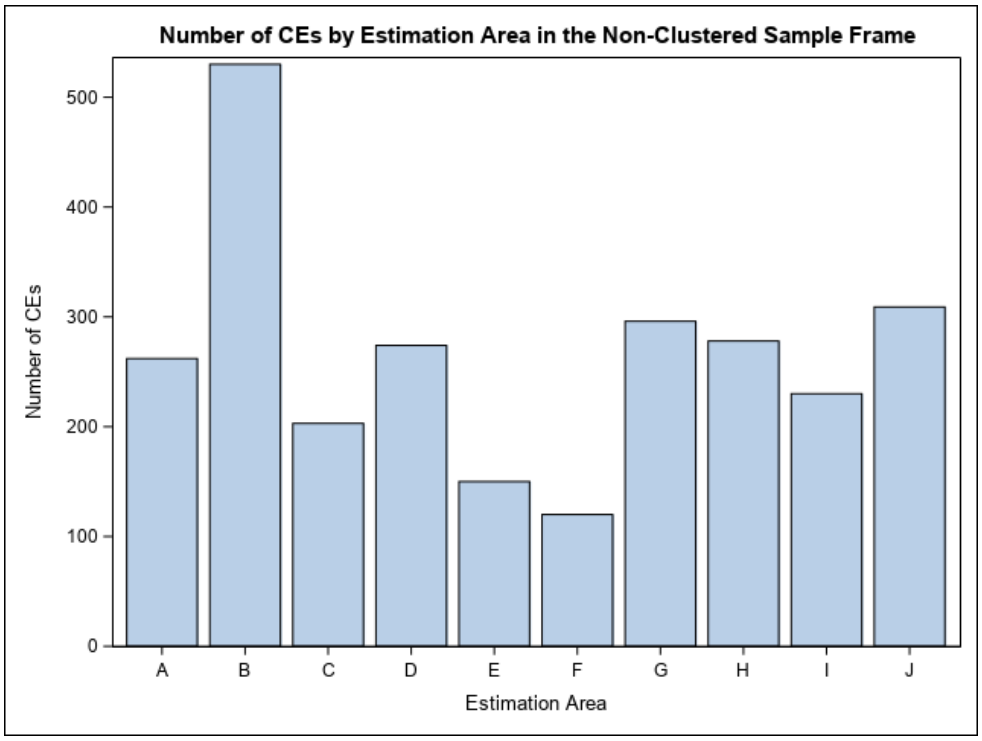

Figure 4: Bar Chart illustrating number of CEs within each Estimation Area (EA)

[4] See paper PMP009: Estimation Areas - Geographical grouping for the stratification of population estimates for more information.
https://www.scotlandscensus.gov.uk/external-methodology-assurance-panels-emaps-0

6.2.2 Estimation Area (EA) and Communal Establishment Type

One possibility is stratifying by CE type which would produce a sample that was representative of the range in establishment types. The major categories of CE are described below:

- Medical and Care
  - General Hospital
  - Mental Health hospital (including inpatient units)
  - Other hospital
  - Care home without nursing
  - Care home with nursing
  - Children's home
  - Other medical and care establishment
- Education
  - School
  - Halls of residence / student accommodation
  - Other educational establishment
- Travel
  - Hotel, guest house, B&B, youth hostel
  - Leisure / holiday establishment
  - Other travel establishment
- Hostel or shelter
  - Hostel or shelter for the homeless
  - Other hostel or shelter establishment
- Other
  - Religious establishment
  - Staff / worker accommodation only
  - Other establishment

It is important to note that Armed Forces and Detention CE types have been deemed out of scope for the purposes of the CCS so are not included. If major CE type is

used as a stratification variable, this would result in five strata, which may lack the required granularity to ensure adequate and representative coverage of Scotland. However if CE type was considered alongside geographical stratification to more efficiently capture response rate variation, five CE types would increase the risk of over-stratification and difficulty allocating proportionally.

Therefore, the number of CE types can be reduced through the use of higher order groupings of CE types. When considering categories that could be collapsed together into higher order groupings, incomplete responses to the 2011 CCS were analysed. It was found that elderly care and travel/temporary accommodation were the two types of CE of under 100 bed spaces that were most likely to have incomplete CE responses, i.e. where the number of completed individual forms was incongruous with the number of residents recorded by the CE manager. The latest version of the Communal Establishment Register (CER, v11) was analysed and Figure 5 illustrates that Medical and Care (which encompasses elderly care) and Travel are the CE types with the highest frequencies.

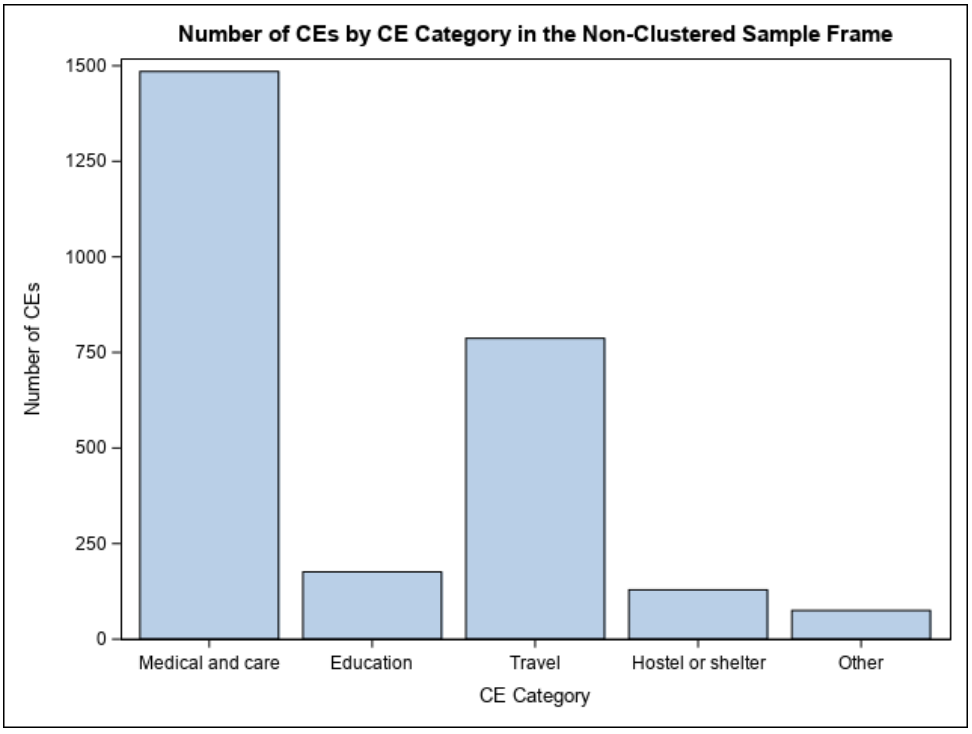

Figure 5: Bar chart illustrating number of small Communal Establishments by Major Category

The higher prevalence of incomplete responses from CEs in these categories is therefore likely to be proportional to the frequency of CEs themselves. Further, both

establishment types can conceivably be thought to have high resident turnovers and a lower proportion of long term residents given their natures, compared to other CE types.

Therefore, it may be prudent to maintain higher level categories of CEs that encompass care and travel establishments separately. Although high in frequency, the perceived low volumes of usual resident data available to enumerate in these establishments highlights the need for increased consideration for these CE types. By categorising these separately, this will ensure that these establishment types are represented within the sample, and increased coverage may offset any impact of high resident turnover and low volumes of data. Therefore, one proposed higher order categorisation is as follows:

- Medical and Care
    - General Hospital
    - Mental Health hospital (including inpatient units)
    - Other hospital
    - Care home without nursing
    - Care home with nursing
    - Children's home
    - Other medical and care establishment
- Travel
    - Hotel, guest house, B&B, youth hostel
    - Leisure / holiday establishment
    - Other travel establishment
- Other
    - Education
        - School
        - Halls of residence / student accommodation
        - Other educational establishment
    - Hostel or shelter
        - Hostel or shelter for the homeless
        - Other hostel or shelter establishment

- Other
  - Religious establishment
  - Staff / worker accommodation only
  - Other establishment

Figure 6 illustrates the maintained trend of Medical and Care and Travel being the most frequent, with these groupings combining the less frequent CE types that would otherwise be likely to require collapsing.

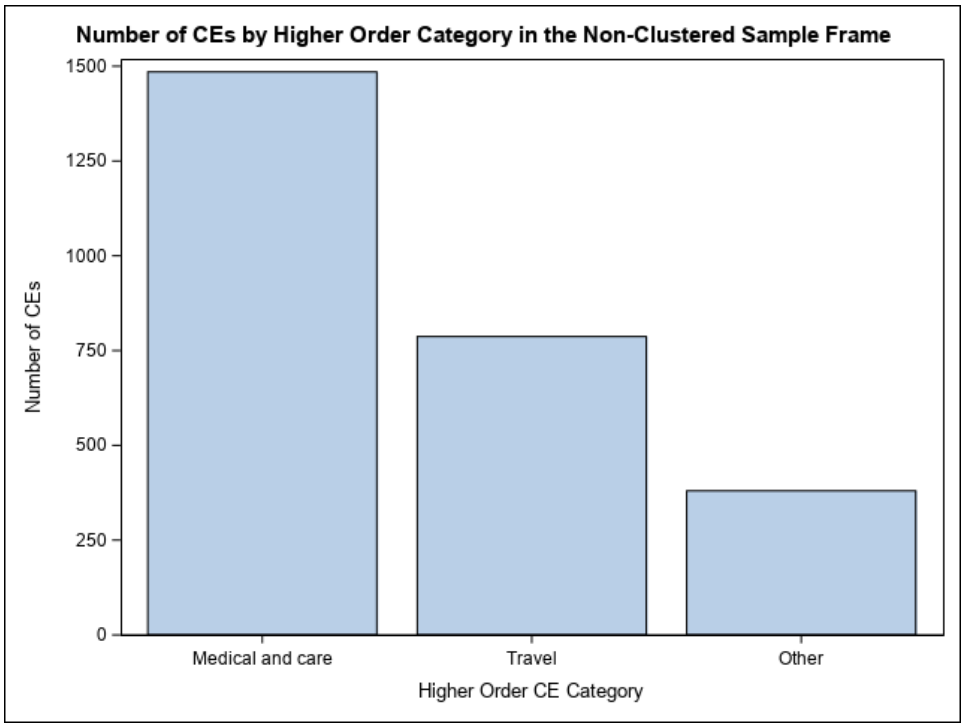

Figure 6: Bar chart illustrating number of small Communal Establishments by Higher Order Category

Given the perceived difference in response characteristics of the collapsed categories, i.e. hostel and shelter establishments compared to education establishments, ideally these would remain in discrete categories. However, the small numbers of CEs in these categories make it difficult to consider any feasible alternatives, therefore this grouping is the only practical collapsing approach.

Figure 7 illustrates the distribution of CEs by higher order type across Estimation Areas (EAs). This would result in 30 strata (3 CE types by 10 EAs), which would appear to avoid the risk of over-stratification. Further, using these stratification variables will maintain a geographical distribution of the sample, whilst ensuring that

each of the higher order categories of CE type are represented in the sample where possible.

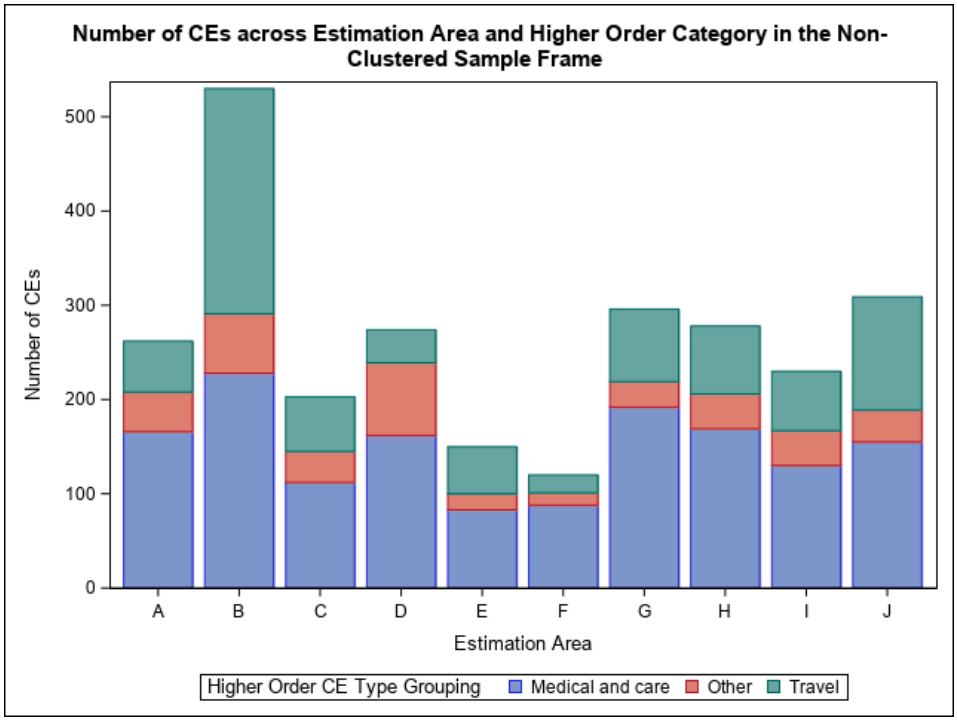

Figure 7: Bar chart illustrating the number of small Communal Establishments in each Processing Unit (PU), split into the three higher order categories described in Section 4.2.2

6.2.3 Bed Space Grouping and CE Type

A further option is using an indicator of the size of each CE as a stratification variable. The two main alternatives are either the potential maximum occupancy of an establishment, e.g. bed spaces, or the number of usual residents living there. Currently bed space information is being used to determine if CEs are within scope, and usual resident information is planned to aid the operational approach of CE enumeration – with only usual residents being enumerated within CEs. The address check phase has not yet been undertaken by the Enumeration team, with usual resident information being available after the completion of this. As mentioned previously, bed space information will be used as an alternative indicator of CE size in the absence of usual resident information.

Figure 8 illustrates the mean number of bed spaces in each higher order CE type. Travel establishments have the highest average number of bed spaces, and

amalgamating other categories results in the mean number of bed spaces of the higher order Other category being only marginally lower than that of Medical and Care establishments.

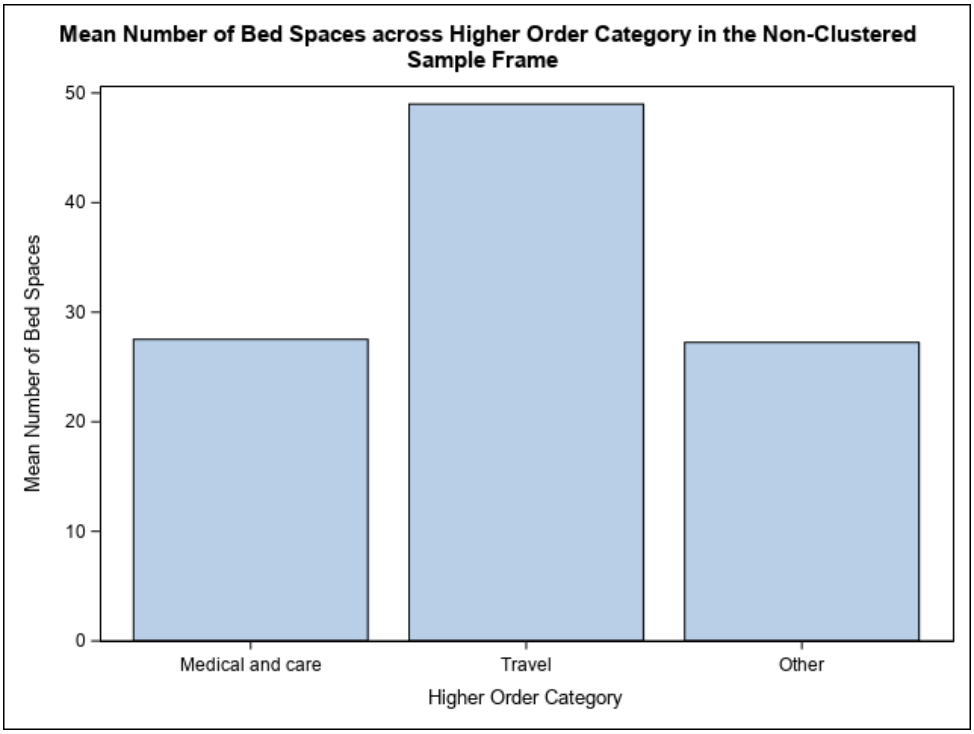

Figure 8: Bar chart illustrating the mean number of bed spaces of small CEs based on higher order CE category

One possible division of bed spaces could be into 4 categories:

- <30 beds
- 30-49 beds
- 50-75 beds
- >75 beds

The higher level categorisation of CE type combined with the four divisions of CEs based on bed spaces would result in 12 strata, illustrated by Figure 9. This lower number of strata could create more representative samples while avoiding the risk of flat allocation resulting from over-stratification discussed previously. There is a lack of geographical distribution of the sample, however both bed space information and CE type provide valuable insight into the nature of establishments being sampled, which may be more meaningful than geographical location.

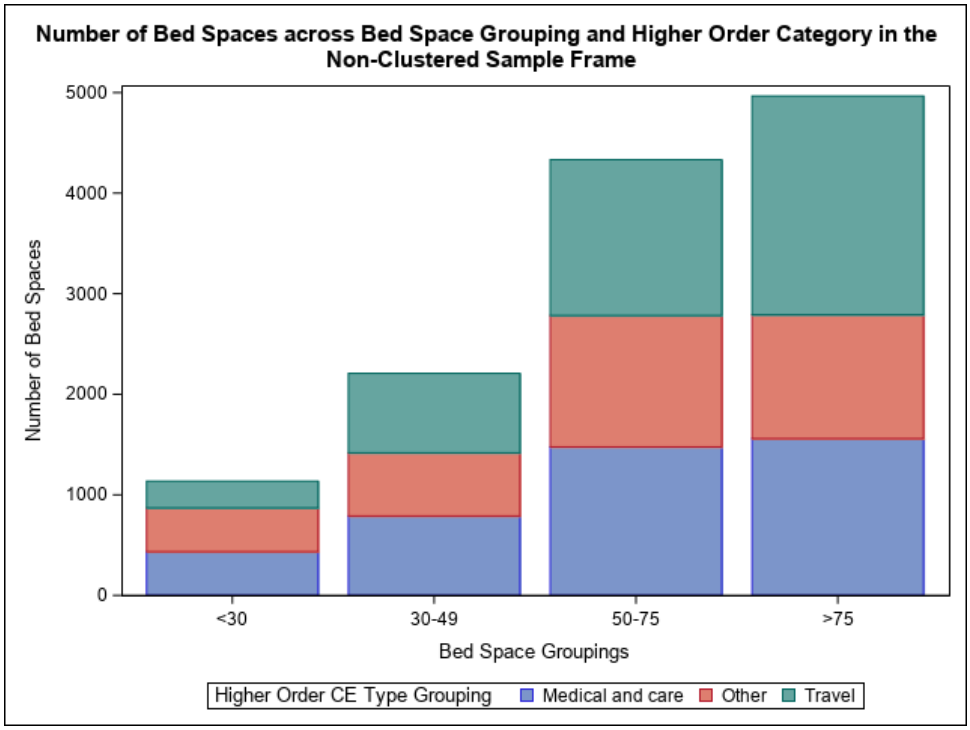

Figure 9: Bar Chart illustrating Total Number of Bed Spaces of CE Categories based on Bed Space Divisions

6.2.4 Proposed Approach

The main stratification options discussed are as follows:

1. Geographical Stratification – stratifying by Estimation Area alone (10 strata).
2. Estimation Area and Higher Order CE type – combining the 10 Estimation Areas with three higher level categories of CE type (30 strata).
3. Bed Space Grouping and Higher Order CE type – combining three higher level categories of CE type with four divisions of bed space groupings (12 strata).

Stratification by Higher Level CE type and Bed Space Groupings (Option 3, discussed in Section 6.2.3) was considered with interest given the question around whether geographical stratification is as relevant to response patterns within the context of CEs as other variables that may be more directly informative of establishment nature. However, the lack of geographical stratification was cause for concern, and therefore the proposal is to implement stratification by Estimation Area and Higher Order CE Type (Option 2, discussed in Section 6.2.2).

6.3 Boost Sample Clustering

There was concern that the CE Boost sample frame contains all within-scope CEs out-with CCS areas across Scotland, and stratified random sampling of this could result in some CEs falling significant distances from existing CCS areas. This could cause operational difficulties, by increasing field force travel times to enumerate such CEs.

A geographically clustered approach was considered as an alternative. As the CCS Planning Areas (PAs, see Annex 1) are already clustered as part of the main CCS sample design[5], creating a sample frame of only the CEs in closest geographical proximity to existing CCS PAs would emulate this approach for the CE Boost sample design. Unlike a standard stratification approach where stratum selection precedes clustering and allocation of the sample, this approach involves the clustered sample frame being created initially – with stratification and allocation occurring subsequently.

To implement this approach, a simulated CCS sample used in the development of the Field Force Model was used to identify an example set of CCS Planning Areas (PAs). The CER (v11) was then used to identify all within scope CEs that did not fall within existing CCS postcodes (as these would be opportunistically sampled as part of the main CCS sample). A geographic information system called ArcGIS was used to compare grid references of CEs to those of PAs to identify CEs located within existing CCS PAs (with physical distance being recorded as zero); as well as geographic PA centroids, to identify the closest CEs out-with CCS PAs, and the distance between the CE and the PA centroid.

6.3.1 Bias Analysis

There were, however, concerns around the possibility of intrinsic bias associated with this approach. By selecting only CEs that are in close geographical proximity to

[5] More information on the CCS sample methodology can be found in the CCS Sample Methodology and the CCS Sample Allocation and Reserve Sample papers
https://www.scotlandscensus.gov.uk/external-methodology-assurance-panels-emaps-0

CCS Planning Areas (PAs), this could bias the sample frame towards certain establishment types that by nature are generally situated near heavily populated areas, and create bias against the selection of more remotely situated CEs.

Analysis was conducted to compare the clustered and non-clustered sample frames, specifically investigating the distribution of number of CEs and number of bed spaces across both sample frames. Figures 10 shows the distribution of CEs across the non-clustered and clustered sample frames, respectively.

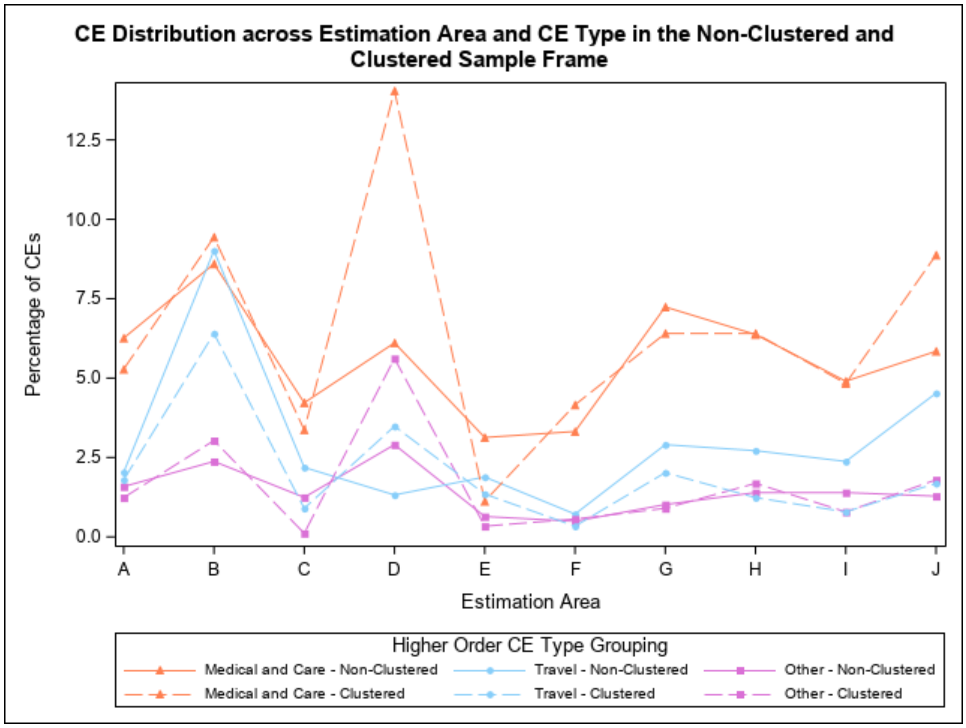

Figure 10: Scatter plot illustrating the distribution of small CEs across Estimation Areas and CE Type in the Non-Clustered Sample Frame (solid line) and Clustered Sample Frame (dashed line).

Estimation Area (EA) D is particularly of note – when considering Medical and Care establishments, the non-clustered sample frame contains nearly double the percentage of CEs as the clustered sample frame. This demonstrates the lack of consistent distribution patterns in the number of CEs within the non-clustered and clustered sample frames.

Similarly, Figure 11 illustrates the bed space distribution in the non-clustered (solid line) and clustered sample frames (dashed line), respectively.

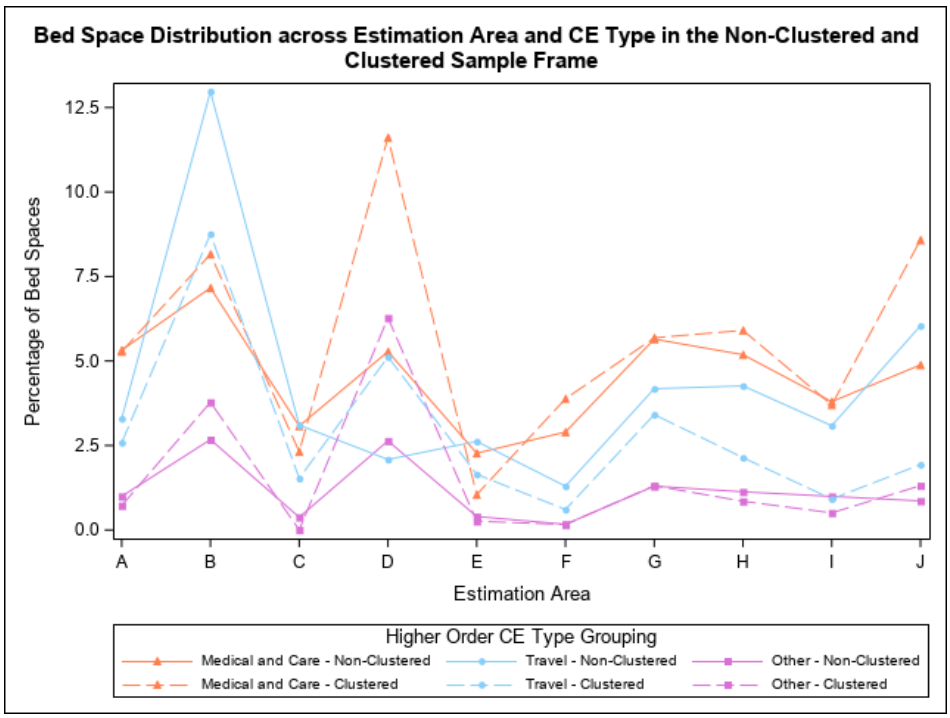

Figure 13: Scatter plot illustrating the bed space distribution across Estimation Area and Higher Order CE Type in the Non-Clustered Sample Frame (solid line) and the Clustered Sample Frame (dashed line).

Taken together, these graphs illustrate the lack of consistent distribution of either number of CEs or number of bed spaces across the non-clustered and clustered sample frames. This supports the suggestion that the clustered approach may introduce bias around particular establishment types based on proximity to densely populated areas, therefore not creating a sample frame that is representative of the total distribution of CEs across Scotland.

6.3.2 Distance Analysis

Additionally, distance analysis was conducted to assess the magnitude of improvement that geographically clustering the sample frame would yield. Figure 15 illustrates the distances associated with both the non-clustered and clustered approaches, using distances created through ArcGIS analysis.

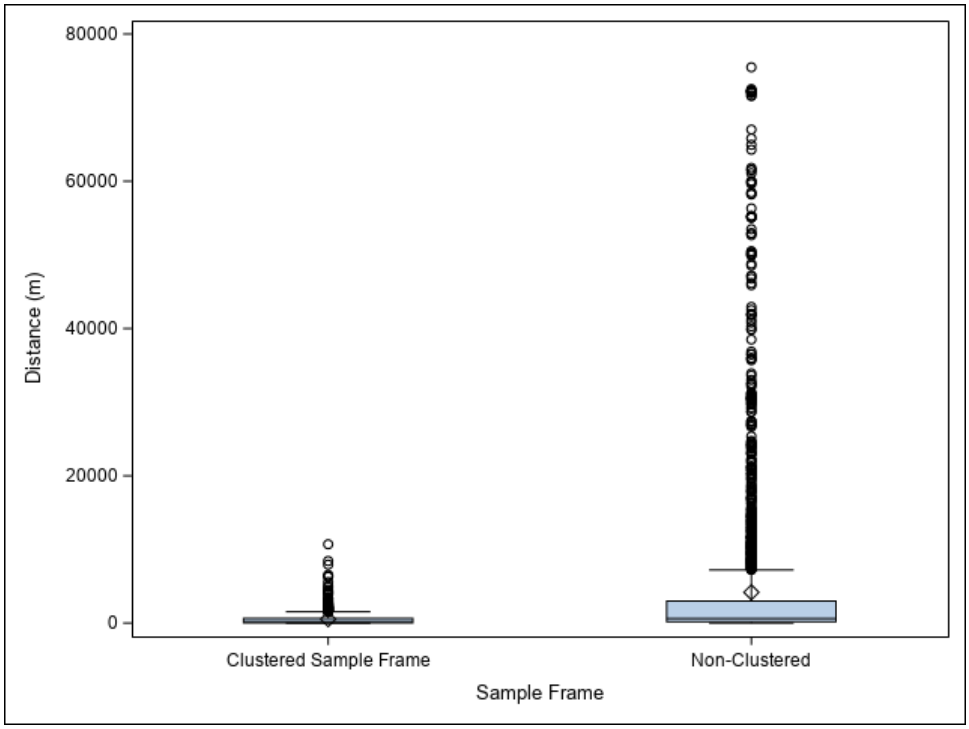

Figure 15: Box plot illustrating the distances associated with the clustered and non-clustered sample frames.

The mean distance from a CCS Planning Area to a CE in the non-clustered Boost sample is 4,175.03m, with a standard deviation of 10,113.01m. Over 80% of the CEs in the non-clustered sample were within 5km of CCS Planning Areas, and the maximum distance that a field worker may have to travel is 75,482.35m (75km). Conversely, the mean distance from a CCS PA to a CE in the clustered sample is 531.56m, with a standard deviation of 948.64m. Over 80% of CEs in the clustered sample were within 1km of CCS PAs, and the maximum distance was 10,698.04m (11km).

Although clustering has a demonstrably beneficial impact on both the average distance between CCS PAs and Boost CEs and the maximum distances required to be travelled, it may not be a significant enough improvement to warrant the risk of bias introduction as a result of implementing a clustered approach.

6.3.3 Operational Considerations

One further aspect of implementing the clustered approach was considered – the operational timelines. If a non-clustered approach was adopted, the preliminary CE

Boost sample frame could be calculated at any point after the Address Check has been conducted by the Enumeration team – expected in January 2022. This consolidates the information on the CER and creates the Address Enumeration extract Communal Establishment Register (CER) that will be used to create the sample frame, which includes additional information not present on the CER – including usual resident information. Once the content of the CER extract has been validated, this can be used as a basis for the sample frame – and sample allocation can occur after this.

Conversely, if a clustered approach is implemented, a sample frame cannot be created until the CCS areas are finalised. Two weeks before the CCS is carried out, a decision will be taken whether to activate the Flexible sample[6], based on preliminary Census response patterns. If the Flexible sample is activated, 20% of the overall CCS sample could be allocated to predicted final 2022 Census response rates, to mitigate the over-allocation based on outdated response patterns. Therefore, CCS areas will not be finalised until after this decision, and so the CE Boost sample frame cannot be geographically clustered around these areas until this point – only two weeks before the start of live CCS operations.

Operationally, this leaves a very short window to create a geographically clustered sample frame, allocate the sample, and subsequently allocate the sampled CEs to CCS fieldworkers. If the sample frame was not geographically clustered, this information would be available from January 2022 – providing an additional three months to fine tune the operational approach. Therefore, this raises the question of whether the operational gains in reduced distances to travel warrant the risk of introducing bias, and the problems that may arise as a result of a short timeframe to implement this approach operationally.

[6] More information on the flexible sample can be found in the CCS Sample Allocation and Reserve Sample Methodology paper
https://www.scotlandscensus.gov.uk/external-methodology-assurance-panels-emaps-0

6.3.4 Proposed Approach

Geographically clustering the CE Boost sample was considered as a way of increasing operational efficiency of the enumeration of Boost CEs, by reducing the distances field workers would need to travel to enumerate additional CEs.
There was an improvement in distance when implementing a clustered approach, reducing the average distance between CCS Planning Areas and CEs from around 4km to around 0.5km, and the maximum distance from around 75km to just under 11km. However, these gains are not significant enough to warrant the risks involved with this approach, including the potential of intrinsic bias associated with implementing this approach (described in Section 6.3.1) and the risk of a short timeframe to operationally implement this approach (described in Section 6.3.3). Therefore, the proposal is to implement a non-clustered design.

7. Strengths and Limitations of Methodology

7.1 Allocation and Weighting

The proposal is to allocate based on the size of CEs included within the strata – using bed space information as an alternative measure of CE size until usual resident information is available – as opposed to the number of CEs themselves. This will result in a sample that is allocated based on the potential volumes of data to be gathered, creating a more meaningful sample distribution and a stronger relationship between statistical design and the associated operational approach. Further, by including weightings in the allocation proportions, this provides a representative sample frame that accounts for and reflects varying CE size across the strata, again producing a more meaningful sample distribution.

7.2 Stratification

The proposal is to stratify by Estimation Area and Higher Order CE Type. Using two stratification variables, related to both geographic distribution and establishment nature, will produce appropriate coverage across the country while suitably representing the variation in establishment types – creating meaningful strata.

7.3 Geographically Clustered Approach

The proposal is to adopt a non-clustered approach, given the risk of bias introduction and the short timeframe to operationally implement the approach. This will allow far more time to create the sample frame, allocate the sample and then operationally allocate these CEs to existing CCS fieldworkers – having the information to create the sample frame in January 2022, compared to the end of April 2022.

The benefits of adopting a non-clustered approach, in terms of increasing the timeframe in which this must be operationally implemented, outweighed the benefits of geographically clustering the sample frame. However, by not implementing a clustered approach, this could mean that CCS fieldworkers may need to travel more significant distances in order to enumerate some CEs within the Boost sample. If the enumeration of CEs that are significant distances away from CCS areas is not planned carefully, this could have a detrimental impact on the efficiency of the field force. However, the additional time to allocate Boost CEs to CCS fieldworkers – as a result of implementing a non-clustered approach – may allow more consideration around the allocation of these CEs to reduce the impact on field force efficiency.

8. Conclusion and Recommendations

The proposal is to define stratum size as relative to the total number of usual residents as opposed to the number of CEs themselves within a stratum. For the purpose of the paper, number of bed spaces were used as a measure, however when usual resident information is available in January 2022 this is planned to be used instead. Weightings were discussed and the impact of this was demonstrated for one stratification option (Figures 1-3), appearing to increase proportionality of bed space distribution across strata of the sample to the sample frame. Therefore, the proposal is to apply weightings to the creation of allocation proportions in the CE Boost Sample.

Three main stratification options are discussed:

1. Stratifying by Estimation Area (EA) alone

2. Stratifying by Estimation Area (EA) and Higher Order CE Type
3. Stratifying by Bed Space Grouping and Higher Order CE Type

Concerns around the lack of geographical stratification in Option 3 resulted in the proposal to implement Option 2 – stratification by Estimation Area and higher level CE type.

A clustered approach was discussed, whereby the sample frame would be created based on the proximity of CEs to existing CCS Planning Areas. This would aid operational ease by reducing travel times, however given the concern over the introduction of bias and a short timeframe to operationally implement this approach, the proposal is to adopt a non-clustered approach.

9. References

Estimation and Adjustment Methodology
https://www.scotlandscensus.gov.uk/documents/Scotlands%20Census%202021%20-%20SMDP%20-%20Estimation%20and%20Adjustment%20Methodology%20paper%20(pdf).pdf

Developing a Hard-to-count Index
https://www.scotlandscensus.gov.uk/external-methodology-assurance-panels-emaps-0

CCS Sample Methodology
https://www.scotlandscensus.gov.uk/documents/Scotlands%20Census%202021%20-%20SMDP-%20Census%20Coverage%20Survey%20(CCS)%20-%20CCS%20Sample%20Methodology%20paper%20(pdf).pdf

CCS Sample Allocation and Reserve Sample Methodology
https://www.scotlandscensus.gov.uk/documents/Scotlands_Census_2022_-_PMP003_-_Census_Coverage_Survey_(CCS)_-_Sample_Allocation_and_Reserve_Sample_Methodology_paper(1).pdf

Estimation Areas - Geographical grouping for the stratification of population estimates
https://www.scotlandscensus.gov.uk/external-methodology-assurance-panels-emaps-0

Annex 1: Glossary of Acronyms

Term	Definition
Communal Establishment (CE)	Classified in Scotland's Census as managed residential accommodation.
Dual System Estimation (DSE)	The process used to account for people who are missed or counted more than once in the population totals.
Hard to Count Index (HtC)	A categorisation (1-5) developed to identify geographical areas which were expected to be difficult to enumerate (i.e. to contain a high proportion of non-responding households).
Relative Standard Error (RSE)	A measure of the extent to which estimates are likely to deviate from the true population.
Communal Establishment Register (CER)	A list of all Communal Establishments in Scotland recorded in the Census.
Estimation Area (EA)	Updated terminology for Processing Unit
Processing Unit (PU)	A method of grouping data together during all stages of data processing, made up of Local Authorities. In 2011 each PU contained at least one Local Authority,and where possible, each PU contained geographically adjacent LAs (one exception was Shetland being grouped with Aberdeen).
Local Authority (LA)	Local Authorities are the 32 council areas within Scotland (see Annex 2).
Planning Area (PA)	These are ideally contiguous areas covering relatively small populations (averaging between 200-400 residential addresses), built from groups of postcodes nestled within Local Authorities.

ArcGIS	Geographic Information System used for creating and using maps, compiling geographic data, and analysing mapped information.

Annex 2: 2011 Census Processing Units

Pre-E&A Identifier	E&A Identifier	LA	LA Code
PU1	PUA	Borders	05
		East Lothian	12
		South Lanarkshire	29
PU2	PUB	Dumfries & Galloway	08
		East Ayrshire	10
		North Ayrshire	22
		South Ayrshire	28
PU3	PUC	Edinburgh	14
		Midlothian	20
PU4	PUD	North Lanarkshire	23
		West Lothian	31
PU5		CCS	
PU6	PUE	Clackmannanshire	06
		Falkirk	15
		Fife	16
PU7	PUF	Glasgow	17
PU8	PUG	West Dunbartonshire	07
		East Dunbartonshire	11
		East Renfrewshire	13
		Inverclyde	19
		Renfrewshire	26
PU9	PUH	Angus	03
		Dundee	09
		Perth and Kinross	25
		Stirling	30
PU10	PUI	Aberdeen	01

		Aberdeenshire	02
		Shetland	27
PU11	PUJ	Argyll & Bute	04
		Highland	18
		Moray	21
		Orkney	24
		Na h-Eileanan Siar	32
PU12		Late Returns