Article:
Mills, O, Shackleton, N, Colbert, J et al. (3 more authors) (2022) Inter-relationships between geographical scale, socio-economic data suppression and population homogeneity. Applied Spatial Analysis and Policy, 15 (4). pp. 1075-1091. ISSN 1874-463X

https://doi.org/10.1007/s12061-021-09430-2

Inter-relationships between geographical scale, socio-economic data suppression and population homogeneity

Abstract

Over time, technology has greatly enhanced access to vast amounts of public data in government datasets. At the same time there has been an increase in 'neighbourhood' level research, in which researchers typically select an administrative unit for their analysis. As the demand for data driven insights and decision making continues to rise, researchers face a tradeoff between data suppression (to protect the privacy of citizens) and homogeneity (the similarity of individuals within an area unit for given characteristics). In this paper, we explore the extent that different scales of geography impact data suppression and spatial homogeneity using the intra-class correlation and the D-Statistic. We use age, sex, ethnicity, education and income data from the 2013 New Zealand Census to assess a) the extent to which data are suppressed, and b) the spatial homogeneity of these variables across 5 scales of 'small area' geography available to researchers in NZ. The data used for this paper was accessed via the Integrated Data Infrastructure (IDI), a large data repository of de-identified, linked microdata obtained from government agencies, and nationally representative surveys. The scales used in this study are the 2013 Meshblock, Statistical Area 1, Data Zone, Statistical Area 2 and Area Unit, each of which can be used to analyse patterns at the 'neighbourhood' scale. We found that Data Zones are a suitable choice for undertaking analyses of census data as they represent a 'medium' scale geography designed to reduce data suppression while maintaining reasonable levels of population homogeneity. The policy implications for this research relate to zone design and decisions relating to the definition of 'a small cell count' for data dissemination for different users of sociodemographic data.

Keywords

MAUP effects; Integrated data infrastructure (IDI); Data suppression; Data homogeneity; Intraclass correlation; Data Zones; New Zealand

Inter-relationships between geographical scale, socio-economic data suppression and population homogeneity

Introduction

Geographical scale and data suppression are no longer an afterthought in the social sciences. The effect of scale, also known as the modifiable areal unit problem (MAUP), can result in statistical bias for area level analyses (Gehlke & Biehl, 1934; Holt et al., 1996; Manley et al, 2006; Openshaw, 1984; Openshaw & Taylor, 1979). Geography also has an intricate inter-relationship with the makeup of communities and the characteristics of the people that live within them (Duncan et al., 1999; Macintyre et al., 2002; Manley et al., 2006; Tobler, 1970). In particular, Macintyre et al. (2002) argue that the role of geography is manifested by three main effects; compositional (the collective properties of individuals within a specific location), contextual (the collective properties of the physical environment) and collective (shared norms, traditions and values). In combination, these effects indicate that individuals with similar characteristics tend to cluster together spatially, and this is more likely to be the case within smaller areas compared to larger areas (Manley et al., 2006). This homogeneity can be useful for targeting policies to groups of individuals (Duncan et al., 1999). Homogeneity also matters for area based deprivation measures that are derived from aggregating individual level information, such as the New Zealand Deprivation Index (NZDep) (Salmond & Crampton, 2012). If individuals were distributed randomly across spatial units, indicating complete heterogeneity, a deprivation index would not be useful for predicting individual outcomes or measuring the overall well-being of a geographical zone.

The Integrated Data Infrastructure (IDI)

Another concern for social science is the impact of data suppression as a result of protecting individual privacy when using microdata. In a NZ setting this includes datasets contained in the IDI. The IDI is a large database of datasets linked at the individual level administered by Stats NZ Tatauranga Aotearoa and contains data sets from various sources such as government agencies, non-government agencies and the 2013 Census data (Milne et al., 2019). Consistent with the United Nations (2007) principles regarding microdata access, Stats NZ has implemented the 'five safes' framework similar to that used by the UK Data Service (2021) to ensure that the data held within the IDI is not misused while ensuring that the public good of New Zealand is not compromised (Stats NZ Tatauranga Aotearoa, 2020). The framework consists of the following criteria which must be met before a researcher is granted access:

1. Safe people - researchers undergo reference checks by Stats NZ.
2. Safe projects - project proposals are vetted to ensure they align with public interest.
3. Safe settings - the data can only be accessed in secure Data Labs.
4. Safe data - the data is de-identified so that individual identifiers are removed.

5. Safe output - output data is checked by Stats NZ to ensure that potentially identifying data is suppressed or confidentialised.

The IDI is important within New Zealand's data landscape because it allows researchers to investigate a myriad of social issues using multiple data sources. Research to-date using the IDI includes: developing activity-based population cohorts (Zhao et al., 2018); socioeconomic and ethnic inequities in childhood obesity (Exeter et al., 2019; Shackleton et al., 2018), and immunization (Charania et al., 2018), the impact injuries or chronic disease have on work and income (Davie & Lilley, 2018); and the impact the 2011/2012 Christchurch earthquakes had on residential mobility and cardiovascular disease outcomes (Teng et al., 2018).

In order to meet the 'safe output' criterion, Stats New Zealand publishes a set of suppression rules to which users must adhere, in order to avoid 'disclosure risk' (Stats NZ, 2016b). Examples of disclosure risk would be when an individual entity (e.g. a person, household, school, business, etc.) is able to be identified, or if key information is learned about an entity within the data. Various suppression rules apply for different data sets, but the following examples are pertinent to the present study:

- Random rounding to base 3 ensures that the results are still representative of the raw values but also introduces an element of random variability.
- Suppressing final counts that have an underlying count less than 6 for sensitive information such as Census, Ministry of Justice and Ministry of Health data.

Arsenault et al. (2013) provide a framework for choosing and assessing a suitable scale of geography. In this paper, we use intra-area homogeneity and sufficiently large population size (n), to assess different scales of geography.

Geographical Scales in the New Zealand Context

Table 1 summarises the geographical scales used in this paper. A Meshblock is the smallest statistical area defined by Stats NZ (2015). SA1s are relatively new areas, built by aggregating adjacent Meshblocks (Stats NZ, 2017). Data Zones were designed as an intermediary geography between Meshblock and census area units, large enough in terms of population size to avoid too much data suppression, but remaining small enough to represent the notion of a neighbourhood and to enable detailed geographical analysis (Exeter et al., 2017). The SA2s were also released in 2018 to replace Census Area Units (AUs) and were constructed by merging whole Meshblocks into zones of similar size to Census Area Units, with a more equal population distribution than AUs (Stats NZ, 2018). Census area units are comprised of aggregated groupings of Meshblocks. Aggregated units define regional council boundaries as well as territorial authorities (Stats NZ, 2016a).

Geography Type	*Number of Units*	*Population Size*
Meshblock 2013 (MB)	45,989*	Approx. 50-150 people
Statistical Area 1 (SA1)	29,778	Approx. 100-200 people (500 max.)
Data Zone (DZ)	5,958	Approx. 500-1,000 people (mean=712)
Statistical Area 2 (SA2)	2,171	Approx. 2,000-4,000 people
Census Area Unit (AU)	1,911	Approx. 3,000-5,000 people

*Table 1: Geographical scales used in analysis. *NB. Excluding water bodies and the coastal 12mile economic exclusion zone*

These geographies might be used by a researcher as the basis for 'area level' analyses. At the Meshblock level we expect higher suppression for some areas, due to small numbers of individuals with a given characteristic, but better ability to detect spatial clustering and higher levels of within area homogeneity. At the other end of the scale, we expect the census area units to exhibit minimal suppression, as they contain a larger number of people, but increased within area heterogeneity due to land area that these units cover.

Historically in New Zealand the effect of area has been investigated solely using Meshblocks (MB) (e.g. Beere & Brabyn, 2006; Darlington-Pollock et al., 2017) and/or the Census Area Unit (AU) (e.g. Crampton et al., 2004; Marshall et al., 1991; Moon & Barnett, 2003). In this study, we compare within area homogeneity and suppression rates for the 2013 Meshblock, Statistical Area 1, Data Zone, Statistical Area 2 and Area Unit.

Data & Methods

Statistical Concepts

The intra-class correlation coefficient (ICC) can be used to assess how strongly individuals within groups resemble each other (Shackleton et al., 2016). The ICC for continuous variables is calculated using equation 1:

$$ICC = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2} \quad (1)$$

where σ_α^2 represents the variance of the level 2 residuals (geographical area) and σ_ϵ^2 represents the variance of the level 1 residuals (individual). However, for logistic models with binary outcomes equation 2 is more often used:

$$ICC = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\frac{\pi^2}{3}} \quad (2)$$

Where σ_α^2 represents the variance of the level 2 residuals and $\frac{\pi^2}{3}$ is assumed to be the residual variance of the level 1 unit, and $\pi = 3.142$. In our study many of the variables are binary so this is the equation that will be primarily used when calculating the ICC value. If the ICC value is close to 0 this means that observations

are mostly independent of each other, and conversely, if the ICC value is close to 1, we would expect responses to be very similar, or spatially clustered within the geographical scale unit of interest (Shackleton et al., 2016). An extension on the ICC is the D-statistic which is used for categorical variables. This correlation statistic "provides a measure of the contribution of each category to the overall statistic" (Burden & Steel, 2015: p. 570). This is essentially a weighted average of the ICC's across nominal categorical variables. In this study, we have used the ICC and the D-statistic as our primary measures of statistical homogeneity.

Sample & Measures

We analyzed data from the 2013 New Zealand Census by accessing microdata of each individual's census return. This approach provides us the flexibility of recoding the structure of census responses for this analysis, along with linkage to a range of geographical units. Our base population was the count of the usual resident population at the 2013 Census (n=4,242,048). We restricted to the usually resident adult population (aged 15+) (n=3,376,417). We chose to look at suppression for variables that would likely be of interest to social science researchers: age, gender, income, education level, and ethnic background For each census variable, unidentifiable or non-stated responses were set as missing observations. We categorised age into four groups: 15-24, 25-44, 45-64 and 65+, and divided the income and education measures into 'No', 'Low', 'Medium' and 'High' groups to further explore the level of suppression for these variables. For personal income (i.e. for individuals), the cutoff points were chosen so that 'Low', 'Med' and 'High' accounted for the bottom 20%, middle 60% and upper 20% of observations. The cutoff bands for household income were adjusted slightly to ensure that the sample sizes were statistically robust. In this case we used the bottom 30%, middle 35% and upper 35% of observations for 'Low', 'Med' and 'High' variables. Similarly, the education cutoffs were chosen so that 'Low', 'Med' and 'High' accounted for the bottom 33%, middle 30% and upper 37% of observations respectively.

Suppression Analysis

The process of suppression analysis is illustrated in Figure 1. The first part of this process involved assessing if each variable of interest would be suppressed within a geographical unit. This process checked the suppression condition ('is count <6') assigning a binary Y/N flag to each unit (i.e. individual zone) in a data set. The result was a table comprising 0's and 1's for 16 variables, across each unit of the 5 geographical scales (Table 2). We also considered suppression across each category for these 16 variables as well as stratifying gender and ethnicity by age group (Supplementary Table 1).

The geography of suppression was determined by adding all the instances of suppression across each scale unit. These suppression scores were summed row-wise such that each MB unit, SA1 unit, DZ unit, SA2 unit and AU unit had a unique suppression score. A score of 0 meant that none of the variables were

suppressed whereas a score of 16 indicated that all the variables were suppressed for that unit of analysis. We would expect more suppression of variables for the smallest units (MBs) and incrementally less suppression as the units become larger.

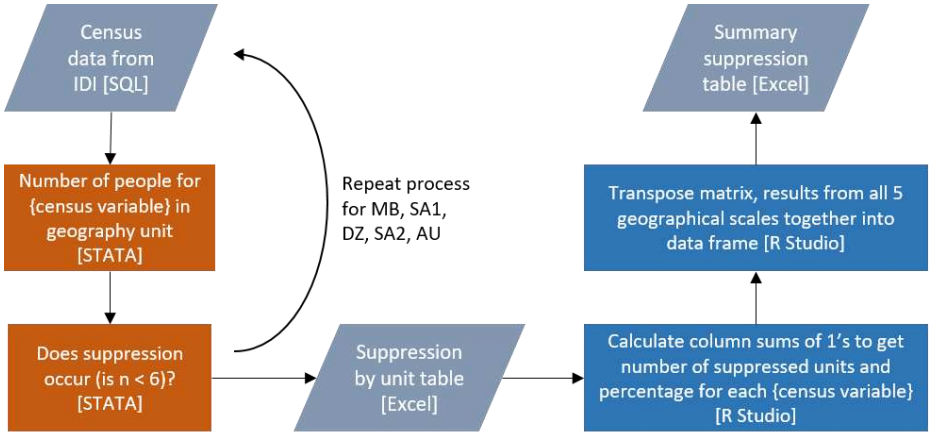

Figure 1: Conceptual suppression process.

Homogeneity Analysis

Calculating the intra-class correlation coefficient provides a simple measure of variation between the Level 1 units (individuals) and the Level 2 units (MB, SA1, DZ, SA2, AU). For example, if we consider the outcomes of two random individuals from a given Meshblock, the ICC would represent the correlation between the individuals. As discussed previously, higher ICC values indicate similarity between individuals who live in the same spatial unit whereas small ICC values indicate high variability between individuals. The ICC values were calculated in STATA using the 'estat icc command'. For binary outcomes, a multi-level mixed effects logistic model was fitted. For ordinal variables, to estimate the ICCs, a multi-level mixed effects linear regression model was fitted. For each variable we have included the ICC estimate that was output from STATA using the '*estat icc'* command, as well as the relevant upper and lower limits of the 95% confidence interval. In some cases, the confidence intervals were inestimable using STATA, so they have been stated as 'NA'.

For nominal variables the D-statistic was instead calculated as the measure of similarity, which required a series of dummy variables to be calculated for each nominal response. For example, the census question relating to smoking classified individuals into three groups (i.e. regular smoker, ex-smoker, never smoked), so three variables smoke1, smoke2, smoke3 were created. The ICC value was calculated for each dummy category and then the D-statistic was calculated using a proportionally weighted sum of the dummy ICC's.

The number of unique geographical units which had at least one resident per zone at the 2013 Census was: 44,196 (MB), 28,997 (SA1), 5,958 (DZ), 2,181 (SA2) and 1,918 (AU). Note that in the suppression analyses, only Data Zones captured information for every geographic unit, while the four other geographies comprised geographic areas that had no population, according to the 2013 Census, excluding the water

bodies and the 12 mile economic exclusion zone around New Zealand's Coast. Therefore, there were 1,877 (4.1%) of the 2013 Meshblocks in the NZ Landmass which had no population, compared with 796 (2.1%) of SA1s, 20 (0.9%) SA2s and 32 (1.7%) of Area Units. As these areas with no recorded population would be 'suppressed' we removed them from the analysis to avoid over-stating the extent of suppression for Meshblocks.

Results

Data Suppression

Table 2 shows the final counts of suppression across the 5 geographical scales for the 16 unique variables. Table 2 can be used to assess the proportion of information on smokers, for instance, that would be suppressed across the different spatial scales, and the corresponding heat map relates to the suppression percentage to illustrate the variables and scales that were highly suppressed. In general, the suppression values decrease as we move from small to larger geographical scales.

Approximately 8% of male and female observations are suppressed for MB and SA1s. Table 2 also highlights a high degree of suppression for Māori, Pacific and Asian ethnic groups, and the variability of suppression across the scales. For example, for counts of the Pacific population, 83% of Meshblocks and 75% of SA1s are suppressed, while a quarter of Data Zones (26%) and Area Units (14%) are suppressed: 10% more than observed for SA2s (16%).

Census Variable	MB (%)	MB (n)	SA1 (%)	SA1 (n)	DZ (%)	DZ (n)	SA2 (%)	SA2 (n)	AU (%)	AU (n)
Male	7.85	3,469	8.8	2,551	0	0	1.15	25	2.08	40
Female	8.24	3,641	8.89	2,578	0	0	1.28	28	2.24	43
Nominal Smoking	4.81	2,124	8.66	2,510	0	0	1.05	23	1.93	37
Ordinal Income (ind.)	4.86	2,148	8.66	2,512	0	0	1.01	22	1.88	36
Ordinal Income (hhld.)	5.99	2,646	8.84	2,563	0	0	1.15	25	2.03	39
Ordinal Education	4.95	2,189	8.66	2,510	0	0	1.01	22	1.88	36
European	6.68	2,951	9.08	2,634	0	0	1.15	25	1.98	38
Māori	51.45	22,738	32.62	9,459	0.25	15	2.8	61	4.69	90
Pacific	82.77	36,581	74.45	21,589	26.26	1,565	16.27	355	23.81	457
Asian	67.58	29,867	53.6	15,542	10.1	602	8.62	188	17.51	336
MELAA	96.66	42,721	94.43	27,384	63.79	3,801	42.35	924	45.08	865
Other	97.27	42,991	93.76	27,189	25.07	1,494	9.17	200	15.74	302
Age 15-24	31.48	13,915	12.91	3,744	0.02	1	2.06	45	3.7	71
Age 25-44	15.79	6,980	9.77	2,832	0.03	2	1.65	36	2.76	53
Age 45-64	12.33	5,450	9.49	2,753	0.05	3	1.37	30	2.19	42
Age 65+	31.74	14,028	15.44	4,476	0.27	16	2.43	53	3.7	71
Total unique spatial units	-	44,196	-	28,997	-	5,958	-	2,181	-	1,918
Total units no population	4.1	1,877	2.1	796	0	0	0.9	20	1.7	32

Table 2: The impact of geographic scale on suppression of 16 variables from the 2013 Census, the percentage and number of units with no population for each geography

Counterintuitively, in some cases, the proportion of suppressed SA1s was higher than the proportion of suppressed Meshblocks. For example, for the European ethnic group, approximately 7% of MBs were suppressed compared to 9% of SA1s. This likely results from the aggregation processes used to construct SA1s, and would potentially occur when a meshblock with small counts (and therefore suppressed in this case) was combined with adjacent (suppressed count) meshblocks, which would reduce the absolute number and a higher proportion of SA1s being suppressed.

Geography of Suppression

Whereas Table 2 shows the overall proportion of zones that are suppressed for each variable, we were also interested to determine whether there was some spatial distribution of data suppression. If particular localities (i.e. geographic units) are continuously omitted from analyses due to data suppression, there is a risk that communities in need of support and resources will be ignored due to the under representation of the issue. Figure 2 shows the geography of suppression for each of the 5 scales of interest. In this analysis, we calculated the proportion of our 16 census variables that would be suppressed, for each geographic unit at each scale of analysis. At the Meshblock level (Figure 2a), for example, only 29 of the 44,196 zones (<0.1%) would have populations large enough to not have any data suppressed. Nearly half (48.1%) would have fewer than 25% of the variables suppressed, and approximately 5% of Meshblocks would have 75% or more of the variables suppressed.

There is a scattering of units (shaded white) through the spine of the South Island and the central plateau of the North Island in Figure 2a, representing those locations with no reported residents on census day (and therefore had no data to be suppressed), 2013. Figure 2b shows the distribution of suppression for SA1's across New Zealand. Interestingly, while 75% of SA1s had fewer than 25% of variables suppressed, there were 8.4% of SA1s – predominantly in rural areas at the bottom of the South Island and near the Tongariro National Park in the central North Island – which had between 75% and 100% of its data suppressed. The Data Zones (Figure 2c) showed a very consistent pattern – with all but 1 of the 5,958 Data Zones having 25% or fewer variables suppressed. At the SA2 scale (Figure 2d), more than half of the units had no data suppressed (53.6%), although there were a few areas located in the central North Island which required between 50% and 100% of our 16 variables of interest to be suppressed. The pattern for the Area Units (Figure 2e) was almost identical to SA2s, however, the location of heavily suppressed zones – in the South Island in particular – was more similar to the distribution of SA1s.

We have also developed an interactive web app version of Figure 2 to enable readers to explore these patterns for specific parts of New Zealand.

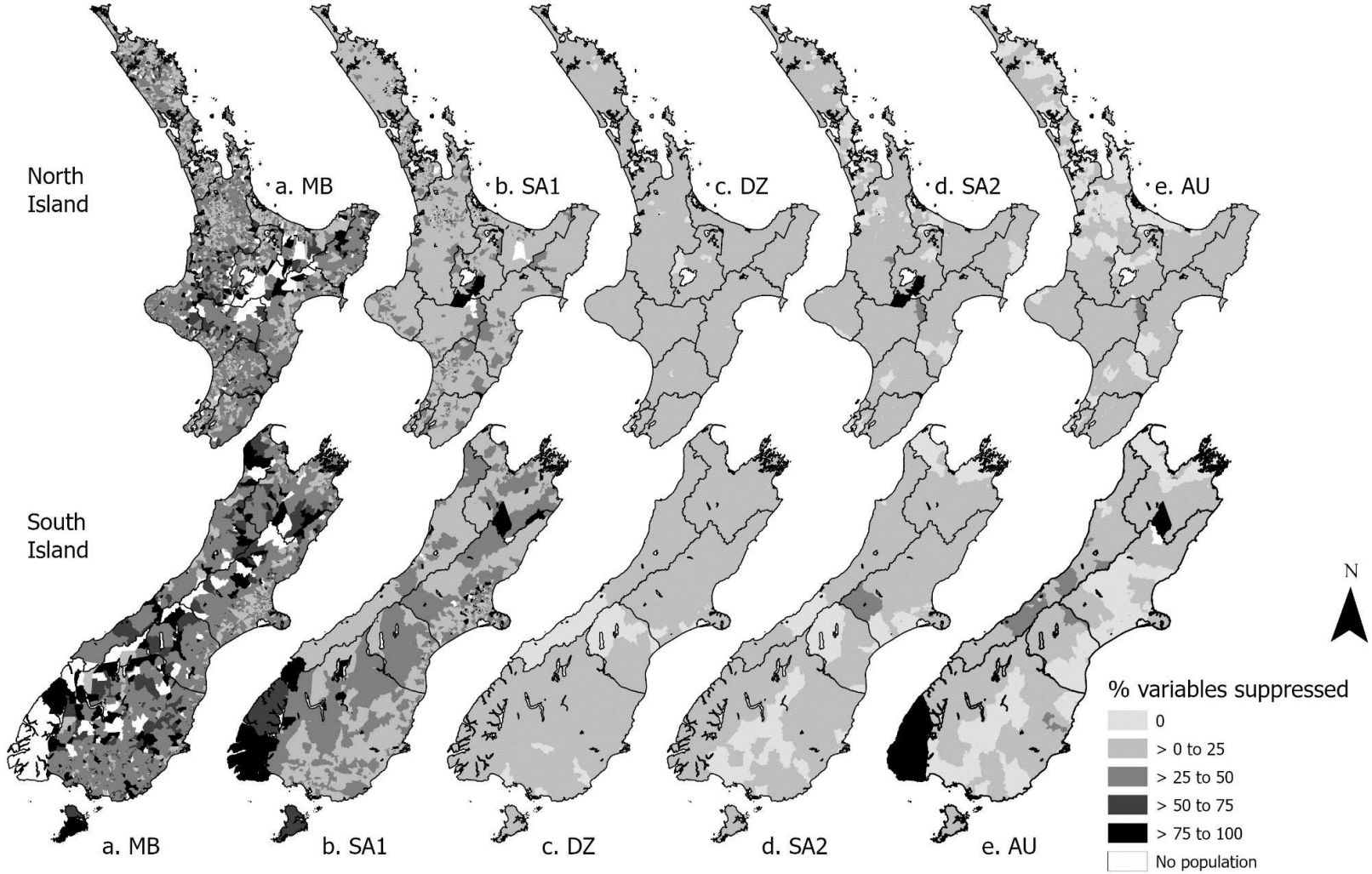

Figure 2: The geography of data suppression in New Zealand

Homogeneity

Higher Intra-Class Correlation (ICC) scores (Table 3) are associated with geographic units in which populations are most similar. Meshblocks are the smallest spatial unit included in this analysis, and unsurprisingly had the highest maximum ICC value (0.518) as well as the highest mean (0.185) and median (0.127) ICC values overall for the 16 variables in Table 3. As the geographical unit increases, the ICC values tend to decrease, with the maximum, median and mean ICC values for the Census Areas (0.349, 0.066, 0.122 respectively) considerably lower than for the Meshblock.

The 16 variables of interest represent four broad groups: sex, age, ethnicity, and socio-economic indicators. Table 3 shows that, as expected, the ICC scores were consistently low (0.002 or 0.003) for both Males and Females across the five geographic scales. Of the four age bands, the ICCs were consistently highest for the population aged 65+, followed by the population aged 15 to 24 years, then 25 to 44 and 45 to 64 years. There was, however, more variability in the size of the ICCs for each age band. For example, for the 45 to 64 years age band, the ICC for Meshblocks (0.05) was 1.5 times greater than that for Area Units (0.032), and for the 15 to 24 years age band the Meshblock ICC (0.127) was 2.5 times higher than for Area Units (0.05). The ICC values were consistently larger for the 5 main ethnic groups (Pacific, Asian, MELAA [Middle Eastern, Latin American and African] European and Māori respectively). For the Pacific ethnic group ICCs were similar for CAUs (0.348) and SA2s (0.346). The CAU ICCs (0.228) were also marginally higher than the ICCs for SA2s (0.208) for the Māori population

Consistent with the ICCs, the estimated D-Statistic for smoking and prioritized ethnicity demonstrates higher homogeneity among Meshblocks and least in Area Units (Table 4).

	Meshblock 2013			Statistical Area 1			Data zone			Statistical Area 2			Census Area Unit		
Census Variable	ICC	CI (L)	CI (U)	ICC	CI (L)	CI (U)	ICC	CI (L)	CI (U)	ICC	CI (L)	CI (U)	ICC	CI (L)	CI (U)
Male	0.002	0.002	0.003	0.003	0.002	0.003	0.003	0.003	0.003	0.003	0.002	0.003	0.003	0.002	0.003
Female	0.002	0.002	0.003	0.003	0.002	0.003	0.003	0.003	0.003	0.003	0.002	0.003	0.003	0.002	0.003
Smoking	0.066	NA	NA	0.061	NA	NA	0.053	NA	NA	0.046	NA	NA	0.046	NA	NA
Income (ind.)	0.068	0.067	0.069	0.066	0.065	0.067	0.057	NA	NA	0.049	NA	NA	0.047	NA	NA
Income (hhld.)	0.161	0.159	0.164	0.139	0.137	0.142	0.106	NA	NA	0.096	NA	NA	0.096	NA	NA
Education	0.112	0.110	0.113	0.107	0.106	0.109	0.097	NA	NA	0.087	NA	NA	0.087	NA	NA
European	0.364	0.360	0.368	0.343	0.339	0.347	0.303	0.296	0.311	0.277	0.265	0.289	0.273	0.260	0.286
Māori	0.301	0.298	0.305	0.265	0.261	0.269	0.217	0.211	0.223	0.208	0.199	0.219	0.228	0.217	0.240
Pacific	0.518	0.512	0.523	0.495	0.489	0.501	0.403	0.394	0.412	0.346	0.332	0.361	0.348	0.332	0.364
Asian	0.453	0.448	0.458	0.437	0.431	0.442	0.372	0.363	0.380	0.360	0.346	0.375	0.349	0.334	0.365
MELAA	0.377	0.369	0.385	0.351	0.343	0.359	0.259	0.249	0.269	0.230	0.217	0.244	0.219	0.205	0.234
Other	0.127	0.123	0.132	0.098	0.094	0.102	0.055	0.052	0.059	0.051	0.047	0.055	0.049	0.045	0.054
Age 15-24	0.127	0.123	0.132	0.079	0.077	0.081	0.063	0.061	0.065	0.057	0.053	0.060	0.050	0.046	0.053
Age 25-44	0.073	0.071	0.074	0.064	0.063	0.065	0.044	0.042	0.046	0.037	0.035	0.039	0.034	0.032	0.036
Age 45-64	0.050	0.049	0.051	0.046	0.045	0.047	0.039	0.037	0.040	0.037	0.035	0.040	0.032	0.030	0.034
Age 65+	0.166	0.164	0.169	0.149	0.146	0.151	0.110	0.106	0.114	0.090	0.085	0.095	0.081	0.076	0.086

Table 3: The influence of geographic scale on spatial homogeneity: Intra-Class Correlations of 16 variables from the 2013 Census

	Smoking			Prioritised Ethnicity		
	D-Statistic	**CI (L)**	**CI (U)**	**D-Statistic**	**CI (L)**	**CI (U)**
Meshblock	0.066	0.065	0.067	0.359	0.355	0.362
SA1	0.061	0.060	0.062	0.340	0.335	0.344
Data Zone	0.053	0.051	0.054	0.299	0.291	0.307
SA2	0.046	0.043	0.049	0.273	0.261	0.285
CAU	0.046	0.043	0.049	0.275	0.262	0.288

Table 4: Spatial homogeneity: D-statistic results for smoking and prioritised ethnicity

Discussion

To our knowledge this is the first paper that examines the effects geographic scale have on suppression and homogeneity of small areas in NZ. The results showed that scale choice does impact both suppression and homogeneity. In general, smaller geographies, such as the Meshblocks and SA1's, exhibited high suppression (less chance of data being output from the IDI) and higher ICC values (indicating greater population homogeneity – a clustering of more similar people). Conversely, larger geographies had lower suppression (more chance of data being output from the IDI) and lower ICC values (indicating more population heterogeneity). Medium scale geographies such as the Data Zones displayed relatively low

levels of suppression across the variables under investigation whilst also having favorable ICC values for area level analyses.

Another consideration is related to geographical scale design and implementation. For example, the Data Zones had the lowest median suppression, and this is likely due to the minimum population constraint. In fact, many variables had 0 counts of suppression for the Data Zones, whereas there are a high number of Meshblocks and SA1's that have fewer than 6 people meaning that they will be suppressed regardless of what variables you are investigating. These insights could be used by Stats NZ if they are designing or implementing new scales or areas for analysis.

Manley et al. (2006) investigated if scale effects existed for Census data in the United Kingdom. The authors used two variables to perform their analysis - the percentage of the population who are local authority renters (RLA) and the percentage of females. These two variables were chosen because they contrast in spatial distribution. Those who rent from local authorities tend to do so in clusters (high ICC) whilst the distribution of females to males is relatively even (low ICC). For scale choice, the authors used census enumeration districts (ED) to measure spatial clustering across Glasgow (for the two variables). Although not a direct comparison, we can think of the enumeration districts as similar scales to our Meshblocks (high suppression, high ICC values). The Female intra-area correlation coefficient (similar to the ICC) for Glasgow was 0.007 and 0.627 for the RLA variable. This is comparable to our results that showed low ICC values for the gender variables (0.002 - little spatial clustering) and high ICC values for variables that typically live in communities together (0.518 - Pacific ethnicity). Manley et al. (2006, p. 159) provide an eloquent summary on these results that fit in well with our work: "Through necessity, the census geography that is provided must be a compromise". They also state that the method described above can be used as a general methodology to explore scale effects - as we have done so in this paper. The major difference is that we looked at a larger list of variables and scales and in addition, investigated how suppression is affected by scale choice.

A further study by Jones et al. (2018) addressed scale effects, in particular the MAUP, and zonation effects by developing a multilevel model analysing segregation of the Indian population living in Leicester UK, at multiple scales simultaneously, rather than focussing specifically on data suppression as we do in this paper. Their analysis comprised three models: a hierarchical model of census Output Areas (OAs) units (similar to SA1s) and author created higher geography 'Zones' (similar to SA2s or CAUs), and two multiple membership cross-classified models which used a moving window ('patch') with varying numbers of contiguous OAs, and three different zonations. Using a complexity-penalised badness-of-fit statistic (DIC), Jones et al. (2018) found that for the hierarchical models, the more complex models (individuals within OAs, and OAs further nested in Zones) reduce the badness-of-fit greatly, while for the multiple membership cross-classified models, larger neighourhood models based on distance were preferred when a single spatial scale is considered. When two spatial scales are considered, a nested spatial scale model is found to represent segregation the best, indicating clustering of Indian ethnicities in larger spatial units of Leicester, and within these larger areas clustering in specific smaller spatial units. A strength of the work

by Jones et al. (2018) is that their approach is extensible to other variables such as multiple groups (e.g. total response ethnicity output).

Despite the complexity in their analyses, the modelling approach used by Jones et al. (2018) revealed similar results to ours, where clustering was evident for a specific ethnic group who may live in communities (or away from other ethnic groups, hence segregation). Jones et al.'s (2018) paper demonstrated the strengths of the multi-scalar approach to population segregation. For example, they found that the Indian population was clustered within larger spatial units of Leicester. Moreover, Jones et al. identified smaller zones within the larger spatial units in which the Indian population were further spatially segregated. Although our study only considered the scale effect, Jones et al.'s (2018) study may be seen as a possible extension to our work. Nevertheless, while the zonation effect is an important aspect of the Modifiable Area Unit Problem (MAUP) to consider, for researchers using areal units with fixed boundaries (such as census units), the scale effect is the predominant consideration in the selection of spatial units. Therefore, the results presented in our paper are immediately of use to researchers without the need to develop complex models.

The major strength of this study is that it is the first of its kind to be conducted in New Zealand. We found considerable variation in the levels of data suppression and homogeneity by geographic scale, and across a range of core demographic variables. The data suppression rule applied to the Census data (that cell counts fewer than 6 should be suppressed) is one of many rules that IDI users must use, depending on the type of data being output for public dissemination. For example, tables containing education attainment data must be suppressed if the underlying count of entities (i.e. schools) is fewer than 2 – regardless of the cell count itself (Stats NZ, 2016b. Therefore, the methods presented in this paper provide a framework for determining the optimal geographic scale of data output from other data sources within the IDI.

Another strength of this study is the insight for New Zealand researchers working with socio-demographic data in the IDI. Researchers can use tables 3 and 5 to assess which scale is most appropriate for a given variable of interest. For example, if we were interested in household income, the suppression percentages are 5.99% (MB), 8.84% (SA1), 0% (DZ), 1.15% (SA2) & 2.03% (AU). In relation to suppression only, it would be wise to perform the analysis using the Data Zones to avoid suppressing up to almost 9% of all observations for that variable.

In conjunction with the suppression counts, we also calculated the ICC values. The ICC's can be used to ensure that the scale choice also allows us to identify homogenous groups of individuals. This is useful when policy makers target groups of individuals based on residence in area units – for example to more efficiently provision health services to groups of similar people rather than at the individual level. ICCs are also informative for policy evaluations and for power calculations for future surveys, because adjustments need to be made to statistical analyses and sample sizes depending on the extent of 'clustering' Shackleton et al, 2016). The ICC values for 'Ordinal Income (hhld.)' are 0.161 (MB), 0.139 (SA1), 0.106 (DZ), 0.096 (SA2) and 0.096 (AU). In this case, the ICC values are higher at the Meshblock and SA1 level, so if the aim

is to identify groups of individuals with similar levels of household income, these lower level geographies will be more useful. However, given the suppression levels described above, and the fact that suppression patterns are different across the income distribution (supplementary appendix) potentially introducing bias, it may be more suitable to use Data Zones.

Limitations of this study were that the 2018 Census data were not available at the time of writing, and the lack of geographic concordance between 2013 Meshblocks and the contemporary SA1s and SA2s. This issue arose because the Meshblocks used in this analysis were created for the 2013 Census, however a more recent Meshblock pattern was used to design the SA1s and SA2s units, which were released in 2018. This created inconsistencies in the concordance tables, for example, where previous Meshblocks have been subdivided in 2018 it was impossible to determine which SA1 to allocate a 2013 Meshblock to. As there is generally a many-to-one relationship between smaller and higher geographies, we would expect a handful of Meshblocks aggregate into (most, if not all of) an SA1 unit. However, on rare occurrences, we found one-to-many instances in which a single 2013 Meshblock was split among more than one SA1 or SA2. To some extent, such issues could be resolved if there were published 'forward mapping' and 'backward mapping' concordance files (Norman et al., 2003), outlining the effect of boundary re-configurations over time. Indeed, one potential workaround we attempted was to use the 2018 census address data set recently made available in the IDI, which provided concordance between 2013 Meshblocks and 2018 Meshblocks (and subsequently to SA1s and SA2s). However, this data set had around 10% missing address data, so this approach brought about its own challenges. There is no doubt that these limitations have contributed to the perhaps counter-intuitive observation that SA1s have a higher proportion of data suppression than Meshblocks for some variables (see Table 2), which would likely be mitigated when making comparisons using geographies built on a common geographic base.

For the variables used in our study, we defined various 'cut-off' values for age, income, education level, and ethnic background. To a degree the decisions made for the cut-offs are arbitrary and we should recognise that different groupings could generate different results. Analogous to the 'modifiable areal unit problem' there can be a 'modifiable categorical unit problem' whereby differently specified cut-offs of continuous/categorical data may lead to different conclusions being drawn (Lomax et al., 2021).

The policy implications for this research relates to zone design and decisions relating to the definition of 'a small cell count' for data release. Many national statistical offices create local -level synthetic geographies to provide both geographic and socio-demographic detail. Inevitably, these represent a compromise in which the smallest zones (usually determined by threshold counts of persons or households) will offer users fewer variables for analyses (Exeter et al 2014). By contrast, larger zones offer more detailed variable combinations. In subsequent releases of 'custom' data extractions, as we have detailed in this paper, cell counts below a threshold level will lead to data suppression.

Paramount to data dissemination and the integrity of research is that the confidentiality of individuals is preserved. There is a need for datasets provided to be safe and non-disclosive but geographical variables

within a dataset create a challenge in this regard. Whilst combinations of personal attributes (age, sex, tenure, ethnicity, health status, etc.) may be observations which are unique in the population or sample, the risk of disclosure is heightened when geographical location is added. A tension therefore exists between making data available which both protects the confidentiality of individuals while containing sufficiently detailed information to underpin the utility of research. Currently, it is likely that neither data suppliers nor data users know the degree to which the data specification decisions made affect the utility of research results.

Conclusion

An important consequence to consider is the impact that suppression has on service provision and distribution of central and local government services, particularly when we acknowledge that similar people live closer together due to the collective effects of geography and neighborhood effects (among other effects). The most heavily suppressed groups of individuals when using lower levels of geography (those aged 65+, 15-24, and of non-European ethnicity) are also more likely to engage with core government services. In general, future research ought to consider the implications of where the suppression occurs, as this may have implications for targeted funding according to area circumstances.

On the other hand, understanding the impact of geographical scale on the similarity of individuals within area units can be used to efficiently allocate public services or interventions to a group. This is especially relevant for communities that need public services the most, such as neighbourhoods or groups of similar people with limited access to health care, education and welfare. At the Meshblock level, the following groups had high ICC values in Table 3: Māori, Pacific, Asian and MELAA groups. The risk that we run into when changing scale to avoid suppression is also creating inefficiencies with targeting public services, due to the homogeneity value decreasing.

The results showed that scale choice does impact both suppression and homogeneity. In general, smaller geographies, such as the Meshblocks, exhibited high suppression and higher ICC values. Conversely, larger geographies had lower suppression and lower ICC values. The Data Zones appeared to be a good candidate for a 'medium' scale geography to mitigate the negative consequences of the effects listed above. Overall, the most important takeaway message is that these effects do exist and thus should be treated with caution alongside the myriad of other challenges that the spatial analyst faces when dealing with similar data. Bearing all of this in mind, this project should be informative rather than prescriptive. It is by no means a complete picture of suppression and homogeneity – the process undertaken in this paper (the code for which is available here) could easily be replicated for different scale units, variables and datasets in other jurisdictions.

Supplementary tables

Census Variable	MB (%)	MB (n)	SA1 (%)	SA1 (n)	DZ (%)	DZ (n)	SA2 (%)	SA2 (n)	AU (%)	AU (n)
Reg. Smoker	37.05	16,375	17.97	5,210	0	0	2.06	45	3.23	62
Ever Smoked	11.51	5,087	9.17	2,659	0	0	1.33	29	2.34	45
No Income (ind.)	61.04	26,978	33.59	9,740	0	0	3.07	67	5.84	112
Low Income (ind.)	27.88	12,323	10.87	3,151	0	0	1.88	41	3.49	67
Med Income (ind.)	6.93	3,062	8.83	2,560	0	0	1.24	27	2.14	41
High Income (ind.)	49.05	21,678	30.1	8,729	1.63	97	3.07	67	5.37	103
No Income (hhld.)	99.08	43,792	98.27	28,495	78.27	4,664	35.88	783	41.9	804
Low Income (hhld.)	21.14	9,342	11.31	3,280	0.05	3	1.7	37	2.66	51
Med Income (hhld.)	17.26	7,627	10.09	2,926	0.05	3	1.88	41	2.92	56
High Income (hhld.)	26.89	11,883	14.53	4,214	0.1	6	2.25	49	3.91	75
No Education	27.64	12,217	14.42	4,181	0.18	11	1.79	39	2.97	57
Low Education	14.94	6,604	10.11	2,931	0.1	6	1.51	33	2.5	48
Medium Education	12.14	5,367	9.06	2,626	0	0	1.33	29	2.34	45
High Education	29.51	13,042	13.2	3,829	0	0	2.25	49	3.65	70
Prioritised Ethnicity	4.6	2,032	8.64	2,506	0	0	0.96	21	1.88	36
Prioritised Māori	51.45	22,738	32.62	9,459	0.25	15	2.8	61	4.69	90
Prioritised Pacific	84.79	37,473	77.55	22,489	34	2,026	20.85	455	29.03	557
Prioritised Asian	68.51	30,279	54.96	15,937	12.2	727	9.9	216	18.81	361
Prioritised European	8.59	3,797	10.08	2,924	0.02	1	1.24	27	2.08	40
Male 15-24	54.72	24,186	27.23	7,897	0.02	1	2.57	56	4.9	94
Male 25-44	31.49	13,917	12.4	3,597	0.03	2	2.11	46	3.34	64
Male 45-64	26.72	11,808	11.42	3,313	0.17	10	1.97	43	2.5	48
Male 65+	57.46	25,396	33.24	9,639	0.6	36	3.3	72	4.59	88
Female 15-24	56.37	24,916	29.14	8,449	0.05	3	2.89	63	5.84	112
Female 25-44	28.4	12,550	11.43	3,314	0.05	3	2.06	45	3.7	71
Female 45-64	25.39	11,223	11.15	3,233	0.2	12	2.11	46	3.18	61
Female 65+	54.38	24,035	30.95	8,976	0.59	35	3.67	80	5.37	103
European 15-24	45.99	20,325	22.73	6,591	0.1	6	2.47	54	4.48	86
European 25-44	24.96	11,030	13.47	3,906	0.15	9	2.15	47	3.13	60
European 45-64	20.69	9,145	13.33	3,866	0.32	19	1.83	40	2.61	50
European 65+	40.52	17,910	23.68	6,868	1.33	79	2.98	65	4.17	80
Māori 15-24	87.09	38,492	77.37	22,435	16.06	957	8.11	177	12.92	248
Māori 25-44	81.16	35,872	67.98	19,714	8.06	480	5.96	130	10.42	200
Māori 45-64	86.8	38,361	77.6	22,501	13.93	830	6.14	134	10.16	195
Māori 65+	98.11	43,362	96.28	27,918	69.51	4,142	34.14	745	35.64	684
Pacific 15-24	93.99	41,541	90.59	26,270	64.05	3,817	42.35	924	46.38	890
Pacific 25-44	92.25	40,773	87.88	25,482	56.07	3,341	35.2	768	39.55	759
Pacific 45-64	95.09	42,029	92.16	26,726	70.8	4,219	48.76	1,064	50.18	963
Pacific 65+	99.03	43,770	98.32	28,510	88.59	5,279	79.38	1,732	78.27	1,502
Asian 15-24	90.51	40,003	84.71	24,565	49.3	2,938	37.08	809	41.01	787
Asian 25-44	81.42	35,985	71.48	20,728	26.01	1,550	18.84	411	26.68	512
Asian 45-64	88.3	39,024	81.6	23,662	41.94	2,499	27.54	601	34.39	660
Asian 65+	97.63	43,149	95.85	27,795	74.68	4,450	60.63	1,323	61.54	1,181

Supplementary table 1: The impact of geographic scale on suppression of additional variables from the 2013 Census

	Meshblock 2013			Statistical Area 1			Data zone			Statistical Area 2			Census Area Unit		
Census Variable	ICC	CI (L)	CI (U)	ICC	CI (L)	CI (U)	ICC	CI (L)	CI (U)	ICC	CI (L)	CI (U)	ICC	CI (L)	CI (U)
Reg. Smoker	0.129	0.126	0.131	0.116	0.114	0.119	0.094	0.091	0.098	0.080	0.076	0.085	0.082	0.077	0.087
Ever Smoked	0.062	0.061	0.063	0.057	0.056	0.059	0.050	0.048	0.052	0.044	0.042	0.047	0.045	0.042	0.048
No Income (ind.)	0.065	0.064	0.067	0.062	0.060	0.063	0.050	0.048	0.052	0.041	0.038	0.043	0.039	0.036	0.042
Low Income (ind.)	0.057	0.056	0.059	0.054	0.053	0.055	0.045	0.043	0.046	0.037	0.035	0.040	0.033	0.031	0.035
Med Income (ind.)	0.063	0.061	0.064	0.059	0.058	0.060	0.048	0.047	0.050	0.043	0.040	0.045	0.040	0.038	0.043
High Income (ind.)	0.178	0.175	0.180	0.176	0.173	0.179	0.161	0.156	0.167	0.137	0.130	0.144	0.135	0.128	0.144
No Income (hhld.)	0.435	0.425	0.445	0.367	0.357	0.377	0.192	0.183	0.202	0.125	0.116	0.135	0.111	0.102	0.120
Low Income (hhld.)	0.176	0.173	0.178	0.147	0.144	0.149	0.107	0.104	0.111	0.096	0.091	0.102	0.098	0.093	0.105
Med Income (hhld.)	0.070	0.069	0.072	0.053	0.052	0.054	0.029	0.028	0.031	0.023	0.021	0.024	0.023	0.021	0.025
High Income (hhld.)	0.243	0.240	0.247	0.207	0.204	0.210	0.153	0.148	0.158	0.136	0.129	0.144	0.140	0.132	0.149
No Education	0.143	0.140	0.145	0.137	0.135	0.140	0.127	0.123	0.131	0.112	0.390	0.441	0.109	0.103	0.116
Low Education	0.133	0.131	0.135	0.128	0.126	0.130	0.120	0.116	0.124	0.107	0.101	0.113	0.104	0.098	0.110
Medium Education	0.019	0.019	0.020	0.018	0.017	0.018	0.014	0.013	0.014	0.011	0.011	0.012	0.010	0.010	0.011
High Education	0.157	0.155	0.159	0.148	0.146	0.151	0.132	0.127	0.136	0.119	0.113	0.126	0.122	0.115	0.129
Prioritised Ethnicity	0.359	NA	NA	0.340	NA	NA	0.299	NA	NA	0.273	NA	NA	0.275	NA	NA
Prioritised Māori	0.301	0.297	0.305	0.265	0.261	0.269	0.217	0.211	0.223	0.208	0.199	0.219	0.228	0.217	0.240
Prioritised Pacific	0.538	0.532	0.544	0.517	0.511	0.523	0.423	0.413	0.433	0.365	0.350	0.380	0.368	0.352	0.385
Prioritised Asian	0.478	0.473	0.483	0.453	0.447	0.458	0.386	0.377	0.395	0.375	0.360	0.390	0.367	0.350	0.383
Prioritised European	0.335	0.331	0.338	0.320	0.316	0.324	0.290	0.282	0.297	0.260	0.249	0.272	0.260	0.248	0.273
Male 15-24	0.063	0.061	0.064	0.057	0.056	0.059	0.041	0.039	0.042	0.036	0.034	0.039	0.033	0.031	0.036
Male 25-44	0.042	0.041	0.043	0.040	0.039	0.041	0.031	0.030	0.033	0.027	0.025	0.029	0.025	0.023	0.027
Male 45-64	0.027	0.026	0.028	0.028	0.027	0.028	0.026	0.025	0.027	0.026	0.024	0.028	0.023	0.021	0.025
Male 65+	0.089	0.087	0.091	0.085	0.083	0.087	0.072	0.069	0.074	0.062	0.058	0.066	0.056	0.052	0.060
Female 15-24	0.069	0.067	0.071	0.065	0.063	0.066	0.050	0.048	0.052	0.044	0.042	0.047	0.040	0.038	0.043
Female 25-44	0.035	0.034	0.035	0.034	0.033	0.035	0.027	0.026	0.028	0.075	0.071	0.081	0.020	0.019	0.022
Female 45-64	0.023	0.022	0.024	0.079	0.077	0.082	0.024	0.023	0.025	0.023	0.022	0.025	0.019	0.018	0.020
Female 65+	0.135	0.133	0.137	0.126	0.124	0.129	0.103	0.100	0.107	0.087	0.082	0.092	0.077	0.072	0.082
European 15-24	0.096	0.094	0.098	0.088	0.086	0.090	0.071	0.068	0.074	0.061	0.057	0.065	0.056	0.053	0.060
European 25-44	0.107	0.105	0.109	0.101	0.099	0.103	0.087	0.084	0.090	0.071	0.067	0.076	0.068	0.064	0.073
European 45-64	0.129	0.127	0.131	0.129	0.127	0.132	0.126	0.122	0.131	0.112	0.106	0.119	0.104	0.098	0.110
European 65+	0.218	0.215	0.221	0.204	0.201	0.208	0.174	0.168	0.179	0.143	0.135	0.150	0.129	0.122	0.137
Māori 15-24	0.239	0.235	0.244	0.222	0.218	0.227	0.175	0.169	0.181	0.157	0.149	0.166	0.160	0.151	0.170
Māori 25-44	0.215	0.211	0.219	0.202	0.198	0.206	0.172	0.166	0.177	0.159	0.151	0.168	0.162	0.153	0.172
Māori 45-64	0.241	0.237	0.245	0.217	0.213	0.221	0.181	0.175	0.187	0.173	0.164	0.182	0.181	0.171	0.191
Māori 65+	0.334	0.327	0.342	0.304	0.297	0.312	0.247	0.238	0.257	0.240	0.227	0.253	0.250	0.236	0.265
Pacific 15-24	0.548	0.541	0.555	0.474	0.466	0.482	0.395	0.383	0.406	0.339	0.323	0.355	0.336	0.318	0.354
Pacific 25-44	0.441	0.434	0.447	0.434	0.427	0.441	0.368	0.358	0.378	0.315	0.300	0.330	0.311	0.295	0.328
Pacific 45-64	0.464	0.456	0.471	0.459	0.451	0.467	0.387	0.376	0.398	0.325	0.309	0.341	0.325	0.308	0.343
Pacific 65+	0.512	0.500	0.524	0.519	0.507	0.531	0.471	0.456	0.487	0.416	0.394	0.439	0.403	0.379	0.426
Asian 15-24	0.406	0.399	0.412	0.402	0.395	0.409	0.369	0.359	0.380	0.363	0.346	0.379	0.352	0.334	0.370
Asian 25-44	0.398	0.393	0.404	0.386	0.380	0.392	0.338	0.329	0.347	0.332	0.317	0.346	0.319	0.303	0.335
Asian 45-64	0.376	0.370	0.382	0.370	0.363	0.376	0.327	0.317	0.336	0.307	0.292	0.321	0.302	0.286	0.319
Asian 65+	0.417	0.407	0.426	0.415	0.406	0.425	0.383	0.370	0.396	0.372	0.353	0.391	0.360	0.340	0.382

Supplementary table 2: the influence of geographic scale on spatial homogeneity for additional variables from the 2013 Census

References

Arsenault, J., Michel, P., Berke, O., Ravel, A., & Gosselin, P. (2013). How to choose geographical units in ecological studies: Proposal and application to campylobacteriosis. *Spatial and Spatio-temporal Epidemiology, 7*, 11-24. https://doi.org/10.1016/j.sste.2013.04.004

Beere, P., & Brabyn, L. (2006). Providing the evidence: Geographic accessibility of maternity units in New Zealand. *New Zealand Geographer, 62*(2), 135-143. https://doi.org/10.1111/j.1745-7939.2006.00056.x

Burden, S., & Steel, D. (2015). Constraint choice for spatial microsimulation. *Population, Space and Place, 22*, 568-583. https://doi.org/10.1002/psp.1942

Charania, N. A., Paynter, J., Lee, A. C., Watson, D. G., & Turner, N. M. (2018). Exploring immunisation inequities among migrant and refugee children in New Zealand. *Human Vaccines & Immunotherapeutics, 14*(12), 3026-3033. https://doi.org/10.1080/21645515.2018.1496769

Crampton, P., Salmond, C. E., Kirkpatrick, R. (2004). Degrees of deprivation in New Zealand: An atlas of socioeconomic difference (2nd Revised Edition). David Bateman.

Darlington-Pollock, F., Norman, P., Lee, A. C., Grey, C., Mehta, S., & Exeter, D. J. (2016). To move or not to move? Exploring the relationship between residential mobility, risk of cardiovascular disease and ethnicity in New Zealand. *Social Science & Medicine, 165*, 128-140. https://doi.org/10.1016/j.socscimed.2016.07.041

Davie, G., & Lilley, R. (2018). Financial impact of injury in older workers: Use of a national retrospective e-cohort to compare income patterns over 3 years in a universal injury compensation scheme. *BMJ Open, 8*(4), e018995. http://dx.doi.org/10.1136/bmjopen-2017-018995

Duncan, C., Jones, K., & Moon, G. (1999). Smoking and deprivation: Are there neighbourhood effects? *Social Science & Medicine, 48*(4), 497-505. https://doi.org/10.1016/S0277-9536(98)00360-8

Exeter, D. J., Rodgers, S., & Sabel, C. E. (2014). "Whose data is it anyway?" The implications of putting small area-level health and social data online. *Health Policy, 114*(1), 88-96. https://doi.org/10.1016/j.healthpol.2013.07.012

Exeter, D. J., Zhao, J., Crengle, S., Lee, A., & Browne, M. (2017). The New Zealand Indices of Multiple Deprivation (IMD): A new suite of indicators for social and health research in Aotearoa, New Zealand. *PloS ONE, 12*(8), e0181260. https://doi.org/10.1371/journal.pone.0181260

Exeter, D. J., Shackleton, N., Browne, M., Zhao, J., Lee, A., & Crengle, S. (2019). Different domains of deprivation and their relationship with obesity in New Zealand 4-year-old children. *Pediatric Obesity, 14*(8), e12520. https://doi.org/10.1111/ijpo.12520

Gehlke, C. E., & Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association, 29*(185A), 169-170. https://doi.org/10.1080/01621459.1934.10506247

Holt, D., Steel, D. G., Tranmer, M., & Wrigley, N. (1996). Aggregation and ecological effects in geographically based data. *Geographical Analysis, 28*(3), 244-261. https://doi.org/10.1111/j.1538-4632.1996.tb00933.x

Jones, K., Manley, D., Johnston, R., & Owen, D. (2018). Modelling residential segregation as unevenness and clustering: A multilevel modelling approach incorportating spatial dependence and tackling the MAUP. *Environment and Planning B: Urban Analytics and City Science, 45*(6), 1122-1141. https://doi.org/10.1177/2399808318782703

Lomax, N., Norman, P., & Darlington-Pollock, P. (2021). Defining distance thresholds for migration research. *Population, Space and Place, 27*(4), e2440. https://doi.org/10.1002/psp.2440

Macintyre, S., Ellaway, A., & Cummins, S. (2002). Place effects on health: How can we conceptualise, operationalise and measure them? *Social Science & Medicine, 55*(1), 125-139. https://doi.org/10.1016/S0277-9536(01)00214-3

Manley, D., Flowerdew, R., & Steel, D. (2006). Scales, levels and processes: Studying spatial patterns of British census variables. *Computers, Environment and Urban Systems, 30*(2), 143-160. https://doi.org/10.1016/j.compenvurbsys.2005.08.005

Marshall, R. J. (1991). Mapping disease and mortality rates using Empirical Bayes Estimators. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 40*(2), 283-294. https://doi.org/10.2307/2347593

Milne, B. J., Atkinson, J., Blakely, T., Day, H., Douwes, J., Gibb, S., Nicolson, M., Shackleton, N., Sporle, A., & Teng, A. (2019). Data Resource Profile: The New Zealand Integrated Data Infrastructure (IDI). *International Journal of Epidemiology*, *48*(3), 677-677e. https://doi.org/10.1093/ije/dyz014

Moon, G., & Barnett, R. (2003). Spatial scale and the geography of tobacco smoking in New Zealand: A multilevel perspective. *New Zealand Geographer, 59*(2), 6-15. https://doi.org/10.1111/j.1745-7939.2003.tb01662.x

Norman, P., Rees, P., & Boyle, P. (2003). Achieving data compatibility over space and time: Creating consistent geographical zones. *Population, Space and Place, 9*(5), 365-386. https://doi.org/10.1002/ijpg.294

Openshaw, S. (1984). The modifiable areal unit problem. *Concepts and Techniques in Modern Geography, 38*. Geo Abstracts, Norwich.

Openshaw, S., & Taylor, P. J. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In N. Wrigley (Ed.), *Statistical Applications in the Spatial Sciences* (pp. 127-144). Pion.

Salmond, C. E., & Crampton, P. (2012). Development of New Zealand's Deprivation Index (NZDep) and its uptake as a national policy tool. *Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique*, *103*(2), S7-S11. https://www.jstor.org/stable/41995682

Shackleton, N. (2018). Hierarchical Linear Modeling. In B.B.Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 774-778). Sage.

Shackleton, N., Hale, D., Bonell, C., & Viner, R. M. (2016). Intraclass correlation values for adolescent health outcomes in secondary schools in 21 European countries. *SSM - Population Health, 2*, 217-225. https://doi.org/10.1016/j.ssmph.2016.03.005

Shackleton, N., Milne, B. J., Audas, R., Derraik, J. G. B., Zhu, T., Taylor, R. W., Morton, S. M. B., Glover, M., Cutfield, W. S., & Taylor, B. (2018). Improving rates of overweight, obesity and extreme obesity in New Zealand 4-year-old children in 2010-2016. *Pediatric Obesity, 13*(12), 766-777. https://doi.org/10.1111/ijpo.12260

Stats NZ Tatauranga Aotearoa. (2015). *Meshblock 2013* [Data set]. https://datafinder.stats.govt.nz/layer/8347-Meshblock-2013/

Stats NZ Tatauranga Aotearoa. (2016a). *Area Unit 2013* [Data set]. https://datafinder.stats.govt.nz/layer/25743-area-unit-2013/

Stats NZ Tatauranga Aotearoa. (2016b). *Microdata output guide (Fourth edition)*. https://www.stats.govt.nz/assets/Uploads/Integrated-data-infrastructure/microdata-output-guide-fourth-edition.pdf

Stats NZ Tatauranga Aotearoa. (2017). *Statistical Area 1 2018 (generalised)* [Data set]. https://datafinder.stats.govt.nz/layer/92210-statistical-area-1-2018-generalised/

Stats NZ Tatauranga Aotearoa. (2018). *Statistical Area 2 2018 (generalised)* [Data set]. https://datafinder.stats.govt.nz/layer/92212-statistical-area-2-2018-generalised/

Stats NZ Tatauranga Aotearoa. (2020). *Integrated Data Infrastructure*. https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/

Teng, A. M., Blakely, T., Ivory, V., Kingham, S., & Cameron, V. (2018). Living in areas with different levels of earthquake damage and association with risk of cardiovascular disease: A cohort-linkage study. *The Lancet Planetary Health, 1*(6), e242-e253. https://doi.org/10.1016/S2542-5196(17)30101-8

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 46*(sup1), 234-240. https://doi.org/10.2307/143141

United Nations. (2007). *Managing statistical confidentiality & microdata access – Principles and guidelines of good practice*.

https://unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

UK Data Service (2021). *What is the Five Safes framework?* https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/ last accessed 1 September 2021

Zhao, J., Gibb, S., Jackson, R., Mehta, S., & Exeter, D. J. (2018). Constructing whole of population cohorts for health and social research using the New Zealand Integrated Data Infrastructure. *Australian and New Zealand Journal of Public Health, 42*(4), 382-388. https://doi.org/10.1111/1753-6405.12781

Statistics New Zealand Disclaimer:

The results in this article are not official statistics They have been created for re- search purposes from the Integrated Data Infrastructure (IDI), managed by Statistics New Zealand. The opinions, findings, recommendations, and conclusions expressed in this article are those of the author(s), not Statistics NZ, or The University of Auckland. Access to the anonymised data used in this study was provided by Statistics NZ under the security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular per- son, household, business, or organisation, and the results in this article have been confidentialised to protect these groups from identification and to keep their data safe. Careful consideration has been given to the privacy, security, and confidentiality issues associated with using administrative and survey data in the IDI. Further detail can be found in the Privacy impact assessment for the Integrated Data Infrastructure available from www.stats.govt.nz.